北京交通大学 CWMT2013 评测技术报告

吴培昊　　徐金安　　苏晨　　张玉洁
北京交通大学计算机与信息技术学院　　北京　　100044
E-mail: {12120465,jaxu,12120447,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 CWMT2013 评测的情况。本次评测，本研究组一共参加了英汉新闻、汉英新闻以及汉英科技三个项目的机器翻译评测任务。本文主要介绍了本研究组参加各个评测任务的系统框架、模型以及评测结果。

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 CWMT2013 机器翻译评测情况进行描述。本次评测本研究组参与了包括英汉新闻领域、汉英新闻领域以及英汉科技领域在内的三个子项目。在三个领域中，本研究组均使用了基于层次短语的统计机器翻译模型。

本文第 2 章简单介绍了参评系统，第 3 章介绍了实验流程及评测结果，第 4 章对本研究组此次评测进行总结。

2　参评系统描述

BJTU-NLP 在本次评测中利用开源工具 NiuTrans 搭建基于短语的统计机器翻译系统以及基于层次短语的统计机器翻译系统。

在参与的三个评测任务中，本研究组在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中，使用了基于短语的统计翻译模型和基于层次短语的统计机器翻译模型。同时，针对汉英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。

在英汉新闻领域机器翻译评测任务中，除了已有的开源汉语分词工具外，本研究组尝试使用了一种领域自适应的汉语分词方法。

本章将对各个系统及方法进行简要介绍。

2.1　基于短语的统计

本次评测中，本研究组使用开源工具 NiuTrans 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 为目标语言句子，f 为源语言句子，$h_m(e,f)$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。

系统采用的节本特征参数为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、语言模型等特征。

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

2.2　基于层次短语的统计机器翻译模型

层次短语模型可以认为是对基于短语的统计机器翻译模型的扩展，它可以抽取源语言句子中非连续部分，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机器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，其规则行事如公式(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做是一个不断使用过程推导的过程.。翻译模型同样采用对数线性模型。使用了包括翻译概率$P(\alpha|\gamma)$和$P(\gamma|\alpha)$，词汇化权重$P_w(\gamma|\alpha)$和$P_w(\alpha|\gamma)$,n-gram 语言模型，规则个数以及目标单词数等。翻译系统最终选择分数最大的推导生成翻译结果。

本次评测中我们使用 NiuTrans 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。

2.3 基于领域自适应的汉语分词

科技领域的汉英机器翻译系统开发中，需要对大规模的汉英平行语料处理以获取翻译知识，同时作为翻译对象会有大规模的该领域的汉语生语料。我们使用融合汉语生语料中的 n-gram 统计特征和汉英语料上的分词引导特征的分词方法。它基于以下想法：利用科技领域的汉语生语料的统计特征实现汉语分词向科技领域的自适应，而利用汉英语料上的英语单词边界和双语对齐特征引导汉语分词；为了融合性质不同的特征，分别实现分词系统，再对各自系统的分词结果进行融合。融合系统的总体框架如图 1 所示。

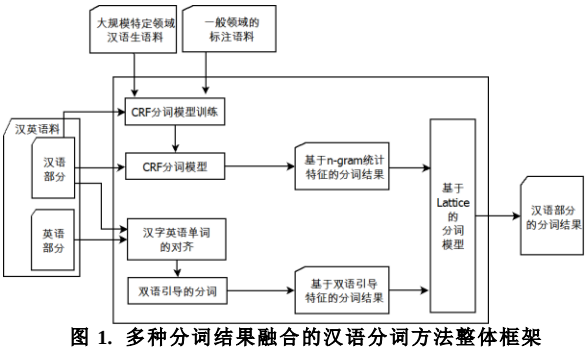

图 1. 多种分词结果融合的汉语分词方法整体框架

为了利用汉语生语料的统计特征，我们实现了基于 n-gram 统计特征的汉语分词系统[Guo Z et al., 2013];同时为了利用汉英语料的分词引导特征，我们实现了基于单词对齐的分词[奚宁等, 2012]，最后我们使用线性融合的技术将两类不同的分词结果进行融合，可描述为公式 3。

$$F(i,j) = \sum_{i=1}^{S} \lambda_i \cdot Conf_{CRF1} \cdot seg_1(i,j) \quad (3)$$

公式 3 中，F(i, j)表示汉语句子中第 i 个汉字到第 j 个汉字构成一个单词的支持度；$Conf_{CRFl}(1\le l\le S)$是对应单词的概率得分；$seg_l(i, j)$是一个二值函数，当第 i 种分词结果中第 i 个汉字到第 j 个汉字为一个单词时，$seg_l(i, j)$值为 1，否则为 0；λ_l $(1\le l\le S)$表示分词结果的特征权重。

我们使用格状结构(Lattice)表示 F(i, j)，Lattice 的解码是一个动态规划的过程，寻找一个支持度乘积最大的分词结果。

本文采用基于网格的线性优化算法[William H et al., 2002]训练参数 λ。首先在多维参数空间中初始化一个点；然后迭代优化参数，每步迭代在固定其他维度参数条件下，优化一个维度的参数使得相应的分词结果 F 值最高；当分词结果的 F 值收敛到了某种期望的程度，结束迭代。为了避免训练参数局部最优，我们选择多个不同的初始点进行参数训练。

3 实验

3.1 系统硬件配置

在本次评测中使用计算机配置与操作系统如表 1 所示。

CPU	内存	操作系统
Intel Xeon CPU E5620 2.4GHz 16 核	48G	Ubuntu desktop 10.04

表 1. 机器硬件配置与操作系统

下面简述三个评测方向的训练及测试过程。

3.2 数据处理方法及工具

对中午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 转移字符转化为对应的特殊字符，全角转半角，中文标点转化为对应的英文标点，标点符号的分离，大写转小写。

对于用于构建语言模型的单语语料，除上述所描述的预处理，还进行去噪处理。

由于 NiuTrans 的分词系统中，包含了时间词、数词等的预处理。因此，本研究组先使用 NiuTrans 自带的分词系统，获取句子中的命名实体。将句子中的命名实体泛化后，对于中文句子使用 NLPIR2013 对句子进行重新切分，以得到最终使用的中文分词结果；英文端则直接使用 NiuTrans 的分词结果。

英文的小写、标点合并等使用了 NiuTrans 自带的 Niurans-training-recase-model.pl 以及 Niutrans-detokenizer.pl 等。

3.3 数据使用

本实验室使用的评测语料，均为评测方所提供的训练数据。在英汉新闻领域机器翻译中，hanukkah 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两版的合并数据。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放提供的开发集。预处理后的训练数据如表 2 所示。

评测项目	翻译模型	语言模型	开发集
英汉新闻	4,541,387 句对	16,070,930 句	1000 句，四个参考译文
汉英新闻	4,541,387 句对	46,037,784 句	1006 句，四个参考译文
英汉科技	897,962 句对	911,527 句	1116 句，四个参考译文

表 2. 评测系统使用的预处理后的数据

3.4 语言模型

语言模型使用的数据与训练模型使用的数据的处理方式相同。预处理后使用 NiuTrans 自带的语言模型训练工具进行训练，最终获得英汉科技的汉语端 5 元语言模型，具体训练命令为：

```
perl NiuTrans-training-ngram-LM.pl \
    -corpus train.txt \
```

-ngram 5 \
-vocab LM/lm.vocab \
-lmbin LM/lm.trie.data

3.5 翻译模型

本次评测中三个评测项目均使用了基于层次短语的统计机器翻译模型，其中英汉新闻领域以及汉英新闻领域还使用了基于短语的统计机器翻译模型。

1) 基于短语模型的统计机器翻译训练
在本次评测中，对于预处理后的语料，使用 mGIZA 进行词对齐。完成双向对齐后，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。
使用 NiuTrans 的短语抽取工具进行短语抽取。短语抽取时，短语长度限制为10，并使用词汇化重排序模型。
规则抽取命令如下所示：

```
perl NiuTrans-phrase-train-model.pl \
        -tmdir model
        -s train.source \
        -t train.target \
        -a alignment.txt
```

2) 基于层次短语模型统计机器翻译训练
使用 mGIZA 进行词对齐，采用启发式规则“grow-diag-final-and”合并双向词对齐结果。
层次短语规则抽取使用 NiuTrans 中的层次短语规则抽取脚本进行抽取，具体命令如下所示：

```
perl NiuTrans-hierarchy-train-model.pl \
        -src train.source \
        -tgt train.target \
        -aln alignment.txt \
        -out hierarchy.rule.table
```

3) 英文大小写转换模型训练
汉英新闻领域机器翻译任务中，需要对译文进行大小写还原，因此训练了英文大小写转化模型，训练命令为：

```
perl NiuTrans-training-recasing-model.pl
        -corpus lm.en \
        -modelDir model.recasing
```

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

在使用开发集进行最小错误率训练时，本研究组使用了两种目标端的语言模型进行对比：1) 以全部英汉新闻训练语料的汉语部分作为训练集获得的语言模型 base.5lm; 2) 以全部英汉新闻语料的汉语部分以及搜狗全网新闻语料合并后作为训练集得到语言模型 all.5lm。测试时，使用基于层次短语的模型作为翻译模型。开发集上的实验结果如表 3 所示：

语言模型	BLEU-SBP5
base.5lm	0.3207
all.5lm	0.3314

表 3. 不同语言模型在开发集上 BLEU 值对比(英汉新闻)

如表 3 结果所示，使用大规模的语言模型进行训练后，开发集上 BLEU 值提升值超过 1 个百分点。因此，在最终解码测试集时，选择语言模型 all.5lm。

本实验组提交了基于短语模型以及基于层次短语模型的翻译结果。评测结果表明，在该领域，本实验组的基于层次短语模型的翻译结果性能高于基于短语模型的翻译结果性能。最终的评测结果如表 4 所示：

测试集	BLEU5-SBP	BLEU5	BLEU6	NITS6	NIST7	GTM
2009-ec-news-phrase	0.2897	0.3006	0.2413	8.9375	8.9529	0.7397
2011-ec-news-phrase	0.2803	0.2947	0.2348	8.4988	8.5012	0.7305
2009-ec-news-hierarchy	0.2961	0.3068	0.2450	9.0287	9.0372	0.7557
2011-ec-news-hierarchy	0.2913	0.3015	0.2422	8.8747	8.8835	0.7484
	mWER	mPER	ICT	METEOR	TER	
2009-ec-news-phrase	0.6493	0.3956	0.3587	0.3913	0.5219	
2011-ec-news-phrase	0.6482	0.3891	0.3546	0.3876	0.5146	
2009-ec-news-hierarchy	0.6714	0.4140	0.3861	0.4388	0.5457	
2011-ec-news-hierarchy	0.6555	0.4042	0.3664	0.4015	0.5322	

表 4. 英汉新闻领域评测结果

3.6.2 汉英新闻领域机器翻译评测记过与分析

与英汉新闻领域相似，在对开发集进行最小错误率训练时，汉英新闻领域也使用两种目标语言端的语言模型进行对比：1) 以全部英汉新闻语料的英语部分作为训练集得到的语言模型 base.5lm; 2) 以全部汉英新闻语料的英语部分和路透社 RCV1 新闻语料合并后作为训练集得到的语言模型 all.5lm。测试时，使用基于层次短语的模型作为翻译模型。

开发集上的实验结果如表 5 所示。从表 5 中可以看出，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 all.5lm。

语言模型	BLEU-SBP4(大小写不敏感)
base.5lm	0.2589
all.5lm	0.2693

表 5. 不同语言模型在开发集上 BLEU 值对比(英汉新闻)

本实验组提交了基于短语模型以及基于层次短语模型的翻译结果。最终评测结果如表 6 所示：

测试集	BLEU4-SBP	BLEU4	NIST5	GTM	mWER
2009-ce-news-phrase	0.1436	0.149	5.7326	0.6152	0.7793
2009-ce-news-hierarchy	0.2256	0.2428	7.6903	0.7104	0.7095
	mPER	ICT	METEOR	TER	
2009-ce-news-phrase	0.5742	0.2781	0.126	0.742	
2009-ce-news-hierarchy	0.5065	0.3157	0.145	0.7051	

表 6. 英汉新闻领域评测结果

3.6.3 汉英科技文献领域机器翻译评测结果与分析

在汉英科技领域中，本实验组对于两种不同的汉语分词方法进行比较。除了 NLPIR2013 外，还使用了章节 2.3 节所述的方法简称为 BJTUSEG。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，使用层次短语模型作为翻译模型。开发集上的测试结果如表 7 所示

汉语分词方法	BLEU_SBP5
NLPIR2013	0.2589
BJTUSEG	0.2693

表 7. 不同汉语分词结果在开发集上 BLEU 值对比(英汉科技)

从表 7 中看出，与 NLPIR2013 比较，本实验组额外使用的分词方法在该训练集与开发集中，没有得到较好的翻译结果。

测试集	BLEU5-SBP	BLEU5	BLEU6	NITS6	NIST7
NLPIR	0.3576	0.3704	0.304	10.0475	10.067
BJTUSEG	0.3543	0.3614	0.2947	9.9371	9.9562
	mWER	mPER	ICT	METEOR	TER
NLPIR	0.6468	0.3281	0.3541	0.5175	0.4897
BJTUSEG	0.6238	0.3114	0.3536	0.5287	0.4993

表 8. 英汉新闻领域评测结果

在最终解码测试集时，本实验组同时使用了两个分词方法，最终的评测结果如表 8 所示。

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 CWMT2013 评测的情况。这是 BJTU-NLP 小组第二次参加全国机器翻译研讨会组织的评测。本实验组较为年轻，在机器翻译的各个方面均存在不足，需要不断学习和实践。本研究组使用开源翻译工具 NiuTrans 实现了基于短语、基于层次短语的翻译系统。在训练过程中，加强了训练语料的预处理，降低了噪声干酪，提高了训练语料的质量，实验结果表明，本实验组的系统性能在汉英新闻领域以及英汉领域相对于基线系统，有一定的提升。

今后，本实验组将在统计机器翻译的预处理、后处理以及基于句法的翻译模型等方面进行深入的研究和探讨。

致谢

本论文受国家自然科学基金项目(批准号: 61370130)和北京交通大学人才基金(2011RC034)资助。

参考译文

Chiang D. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 263-270.

Chiang D. Hierarchical phrase-based translation[J]. computational linguistics, 2007, 33(2): 201-228.

Och F J, Ney H. A systematic comparison of various statistical alignment models[J]. Computational linguistics, 2003, 29(1): 19-51.

Och F J. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 160-167.

Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 48-54.

Koehn P. Pharaoh: a beam search decoder for phrase-based statistical machine translation models[M]//Machine translation: From real users to research. Springer Berlin Heidelberg, 2004: 115-124.

Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation[C]//Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2007: 177-180.

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. 2012. NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation. In *Proc. of ACL*, demonstration session.

Guo Z, Zhang Y, Su C, et al. Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation[M]//Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2012: 121-131.

奚宁, 李博渊, 黄书剑, 等. 一种适用于机器翻译的汉语分词方法[J]. 中文信息学报, 2012, 26(3): 54-58.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. Numerical Recipes in C++. Cambridge University Press, Cambridge, UK.