北京交通大学 CWMT-2015 评测技术报告

明芳，徐金安，陈钰枫，张玉洁，李少童，郑晓康，王楠

（北京交通大学计算机与信息技术学院，北京,100044）

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 CWMT2015 评测的情况。本研究组参加了评测项目中的两个子项目评测：英汉新闻以及汉英新闻的机器翻译评测任务。本文主要介绍了研究组参加各个评测任务的实现框架、模型以及它们在评测数据上的性能表现，同时，以优化分词效果和改善命名实体音译质量为重点，结合翻译结果进行比较和分析。

关键词：层次短语模型；错误驱动学习方法；命名实体音译法

BJTU-NLP LAB Technical Report for the 2015 China Workshop on Machine Translation

Fang Ming, Jin'an Xu, Yufeng Chen, Yujie Zhang, Shaotong Li, Xiaokang Zheng，Nan Wang

（School of Computer and Information Technology, Beijing Jiaotong University，Beijing, 100044）

Abstract: This paper describes the details of machine translation and evaluation task for BJTU-NLP LAB's joining in the China workshop on Machine Translation in 2015(CWMT2015). In this task, we take part in 2 sub-tasks: Chinese-to-English and English-to-Chinese translation for news. The report mainly describes the implementation framework, model and the performance in the evaluation data set of the evaluated translation system. In addition, regarding optimizing segmentation effects and improving name entity transliterating results as key points, all translation results are compared and analyzed.

Key words: Hierarchy phrase-based model; Error-driven rule learning method; Name entity transliterating method

1 引言

北京交通大学自然语言处理研究组(BJTU-NLP)参加了 2015 年第十一届全国机器翻译研讨会（CWMT2015）组织的机器翻译评测活动。在所有评测项目中，本研究组参与了包括英汉新闻领域、汉英新闻领域两个子项目。

近年来，基于统计的机器翻译方法[Koehn et al., 2010]以其强泛化、无指导的学习能力、无需语言学家编写语言翻译规则等优势，克服了基于规则的机器翻译方法的缺点而占据主导地位。统计机器翻译方法[Huang, 2005]有很多，包括短语模型[Koehn et al., 2003]、层次短语模型 [Chiang, 2005; Chiang, 2007]、以及基于句法的统计翻译模型等。其中，基于句法的统计翻译方法还包括树到串模型[Liu et al., 2006]、森林到串模型[Liu et al., 2007]、树到树模型[Eisner et al., 2003；Zhang et al., 2008]等。层次短语模型是对基于短语的统计翻译模型的扩展，继承了短语翻译模型的优点，通过引入层次短语的概念，依靠变量规则直接进行依据语言规律的调序，排除了很多歧义性，更好地解决短语调序问题。但层次短语模型也存在如下主要缺点：抽取规则计算开销大；解码时间长；规则抽取过程约束不足，规则库存在巨大冗余；规则中只包含一个变量 X，规则的使用除了规则中的词汇信息外无其它约束。

为了解决以上问题，研究者提出了融合语义知识的层次短语模型[Xiao et al., 2014]，将源语言的语义信息引入到层次短语模型当中，另有研究者在此基础上提出了软依赖匹配的层次短语模型[Cao et al., 2014]，不仅能够获取规则内部各部分之间的依赖关系，还可以从全局的角度获取模型规则之间的依赖关系和上下文关系。鉴于以上研究，本文采用层次短语模型，重点研究优化分词结果和改善命名实体音译等内容，旨在优化层次短语模型。

本文的结构安排如下：第二章简要介绍评测系统架构及相关原理；第三章介绍实验流程及评测结果；第四章对评测结果进行总结。

2 参评系统描述

BJTU-NLP 在本次评测中利用开源工具 NiuTrans[Xiao et al., 2012]搭建基于层次短语的统计机器翻译系统。本研究组在本次的汉英、英汉新闻领域机器翻译评测任务中，除了使用已有的开源工具和方法外，还采用了错误驱动学习方法和命名实体音译的方法，提升翻译性能。

本章将对各个系统及方法进行简要介绍

2．1 基于层次短语的统计机器翻译模型

层次短语模型抽取元语言句子中非连续部分。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，其规则形式如公式(1)所示：

$$\chi \rightarrow <\gamma, \alpha, \sim> \quad (1)$$

其中 χ 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 SCFGs 文法，通过启发式规则从字对齐的平行语料中获取 SCFGs，首先抽取初始短语对，然后获取层次短语规则。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目标语言词串；同时为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。翻译模型采用对数线性模型。使用了包括翻译概率 $P(\gamma|\alpha)$，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特征函数。翻译系统最终选择分数最大的推导生成翻译结果。

2．2 语料预处理

本次评测中用到的所有评测语料（包括训练集、开发集、测试集以及测试集的参考集等）均采用如下预处理方法：

中文语料：

1） 常见 html 转义字符转化为对应的特殊字符，全半角转化；

2） 对于数字、 日期、时间、网址等不可枚举，无法通过词典简单查找的数据稀疏问题，可以采用正则表达式或者自动机进行自动识别，并给予特殊名字进行泛化，通过泛化的方法解决该问题；

3） 中文分词。

英文语料：

1） 全半角转化，常见 html 转移字符转化为对应的特殊字符；

2） 大写转小写；

3） 将数字、日期、时间、网址等不可枚举的类型进行识别，然后分别采用特殊名字进行泛化处理；

4） 将英文句尾结束符与句尾最后一个单词用空格分开。

对于用于构建语言模型的单语预料，除上述描述的预处理外，还需进行去噪处理，过滤不规范中英文句子。

2．3 错误驱动学习方法

支持向量机 SVM(Support Vector Machine）是一个有监督的学习模型，通常用来进行模式识别、分类、以及回归分析。中文分词的精度对于机器翻译的精度有很大的影响，本文使用 SVM 模型和错误驱动的学习方法相结合进行中文分词数据修正。

1）模型特征:

本文将汉字进行数字映射，取句子中 5 个字符和其对应的字符类别作为特征训练模型。

2）学习规则：

进行文本分词后，先取出部分已经分好词句子的 5 个字符进行人工判断正确错误，并同时修正的分词结果，融合的规则集合进行特征标注。同时，取出的每条规则去验证测试集，然后比对 PRF 值提升来确定，是否将此条规则并入规则集中。

3）重构字典和训练集：

每次进行错误驱动后，将修改后的新词并入字典中，同时，将修正后的句子并入训练集中，重新训练判别模型。

错误驱动学习方法，在初始小规模数据中学习规则，通过模型对数据进行判别和修正，同时在判别和修正的过程中扩大学习规则集，达到 SVM 判别模型的对数据精度不断提升，减少人工修正数据规模，最终进行修正数据，提升文本分词的精度，同时提升翻译性能。

2．4 命名实体音译模型

由于人名、地名、机构名等命名实体在新闻领域出现较为频繁，其翻译效果对整体翻译结果有较大影响。尤其是人名的音译尤为关键，因此我们对抽取出的人名进行了单独的音译处理。首先我们从训练语料中抽取汉英实体翻译对作为评测中人名音译所需的词典和训练语料。针对人名翻译，我们采用多特征融合方法[Wang et al., 2015]进行音译，音译过程可以被描述为，对于给定的源语言人名 f，从所有可能的音译结果 e 中，找到最优的音译结果 $\hat{e}$ 满足：

$$\hat{e} = \arg\max_{e} \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e,f)\right)}{\sum_{e'} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e',f)\right)} \quad (2)$$

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。

由于音译过程是从左到右顺序翻译，不存在句子翻译的调序问题，因此音译模型不需要置换模型特征，并且由于音译翻译中的“短语翻译对”的内部不存在词汇（英文字母和汉字）对应关系，在音译模型中我们不采用短语词汇化特征和反向短语词汇化特征。最终在本系统使用了如下特征：

（1）正向短语音译概率：$P(e|f)$

（2）反向短语音译概率：$P(f|e)$

（3）人名的长度特征

（4）长度惩罚

汉语短语 f 翻译为英语短语 e 的概率 $P(e|f)$，英语短语 e 翻译为汉语短语 f 的概率 $P(f|e)$ 与长度惩罚因子 $I(e|f)$ 与 $I(f|e)$ 分别按照公式(3)计算，其中 len(c) 代表汉语短语包含的汉字的个数，len(e)代表英语短语中包含的辅音个数。

$$P(e|f) = \frac{count(e|f)}{count_{-e}(e|f)}$$

$$P(c|e) = \frac{count(f,e)}{count_{-f}(e,f)}$$

$$I(e|f) = \frac{|len(f) - len(e)|}{len(f)}$$

$$I(f|e) = \frac{|len(e) - len(f)|}{len(e)}$$

（3）

（5）汉语及英语的语言模型 $lm(c)$。

基于短语的翻译模型在翻译各个源短语片段时，不需要考虑它前面的源语言短语片段是什么，各个源语言短语是独立翻译的，各个源语言短语的目标语言短语翻译之间仅仅靠目标语言的语言模型进行联系。

2．5 数据后处理

层次英汉翻译结果的处理：

1） 将系统输出的英文标点符号转化为中文标点；

2） 去掉中文分词之间的空格；

3） 进行 A3 区的全角到半角的转化。

汉英翻译结果的处理：

1） 英文恢复大写信息，该过程分为两步，首先需要训练大写信息恢复模型，然后对翻译结果进行大写信息恢复；

2） 去掉标点与最后一个单词之间的分词间隔；

3） 半角转全角。

3 实验

3.1 实验环境

本次评测中使用计算机配置和操作系统如表 1 所示：

表 1. 机器硬件配置和操作系统

CPU	内存	磁盘	操作系统
Intel Xeon CPU E5620 2.4GHz 16 核	64G	5T	Ubuntu 15.04 desktop 1004

3.2 实验设置

本实验基线系统：数据经预处理之后，使用NiuTrans系统自带的分词工具对数词、时间、日期等命名实体词汇做泛化处理；利用GIZA++[Och and Ney, 2003]训练词对齐模型；翻译模型使用对数线性模型，在扩展的开发集上过滤无效的翻译规则；采用NiuTrans 自带的语言模型训练工具训练 5 元语言模型；在开发集上使用最小错误率训练(MERT)方法[Och et al., 2003]对对数线性模型参数进行优化。

本实验主系统：在基线系统基础上，使用 NiuTrans 系统自带分词工具之后，抽取人名等命名实体，即抽取汉英实体翻译对作为评测中实体翻译所需的词典和训练语料。再利用 CRF++工具对分词结果做进一步优化时，采用错误驱动学习方法优化分词结果；GIZA++训练词对齐模型；翻译模型同样使用对数线性模型，并做进一步优化，在扩展的开发集上过滤无效的翻译规则；采用NiuTrans 自带的语言模型训练工具训练 5 元语言模型；在开发集上使用最小错误率训练(MERT)方法对对数线性模型参数进行优化。

3.3 实验数据

本实验使用的训练集、开发集、测试集均为评测方提供的训练数据。实验数据规模如表 2 所示：

表 2. 翻译模型训练集数据

评测项目	训练集对数	开发集	测试集
汉英新闻	6493765	1006 句，4 个参考翻译	1100 句
英汉新闻	6493765	1000 句，4 个参考翻译	1100 句

在选择语言模型数据规模时，本实验组首先做了 2 组对比实验，针对汉英新闻，仅以目标语言训练集作为语言模型训练数据，得到 5 元语言模型 Base.5lm;将目标语言训练集与路透社新闻语料的合并数据作为语言模型，得到 5 元语言模型 All.5lm。对比结果如表 3 所示：

表 3 不同语言模型在开发集上 BLEU 值对比（汉英新闻）

语言模型	BLEU-SBP4
Base.5lm	0.1468
All.5lm	0.1619

另一组实验是英汉新闻，同样仅以目标语言训练集作为语言模型训练数据，得到 5 元语言模型 Base.5lm;将目标语言训练集与搜狗全网新闻语料的合并数据作为语言模型训练数据，得到 5 元语言模型 All.5lm，对比结果如表 4 所示：

表 4. 不同规模大小的语言模型在开发集上 BLEU 值对比（英汉新闻）

语言模型	BLEU-SBP5
Base.5lm	0.2663
All.5lm	0.2855

经上述对比实验，做如下选择：在英汉新闻领域机器翻译中，语言模型的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新闻语料的合并数据。因此，经预处理后的语言模型数据规模如表 5 所示：

表 5. 语言模型训练集数据

评测项目	语料名称	语言模型
汉英新闻	目标语 + routers	16409735 句
英汉新闻	目标语 +sogou 新闻	27693591 句

3.4 实验结果

3.4.1 汉英新闻

在汉英新闻领域，本实验组首先对汉语源语言分词结果采用2.3节所述错误驱动学习方法对分词结果进行优化处理，为对比处理前后分词效果，本实验组在小规模语料上进行了实验测评，预测整体分词效果，PRF值作为评价指标，结果显示较初始分词效果更佳,因此最终选用优化后的分词结果。

本实验组提交了基于层次短语模型的 2 组翻译结果，分别是基于层次短语的基线系统的翻译结果ce-contrast和优化基线系统所得到的翻译结果 ce-primary,最终评测结果如表 6 所示。

表 6. 汉英新闻领域机器翻译结果

system	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
ce-primary	0.1659	0.1819	5.9812	0.6525	0.7689	0.5693	0.2705	0.1555	0.7392
ce-contrast	0.1619	0.1772	5.8119	0.6472	0.7695	0.5760	0.2698	0.1519	0.7417

表 7. 英汉新闻领域机器翻译结果

system	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
ec-primary	0.2884	0.3087	0.2468	09.2912	9.3019	0.7850	0.7333	0.4104	0.3335	0.4762	0.5853
ec-contrast	0.2855	0.3103	0.2482	9.1263	9.1369	0.7783	0.7309	0.4183	0.3378	0.4684	0.5853

对于汉英翻译，源语言句子：“*李克强在政府工作报告中指出，要把“大众创业、万众创新”打造成推动中国经济继续前行的“双引擎”之一*”。 基线系统翻译结果如下：“*in the government work report that the peoples of innovation , ″ mass business , ″ to make china ’s economy continues to move on , ″ one of the twin - engine ″ ”*.结果显示对人名翻译缺失，通过优化基线系统，翻译结果为：*“Li Keqiang in the government work report that the peoples of innovation , ″ mass business , ″ to make China ’s economy continues to move on , ″ one of the twin - engine ″ ”*.对比翻译结果，可知对人名翻译有较大提升。评测结果表明，在汉英新闻领域，本实验组的方法相对基线系统性能有一定提升，对翻译性能提高能起到一定的作用。

3.4.2 英汉新闻

本实验组最终提交了基于层次短语模型的两组翻译结果，分别是基于层次短语的基线系统的翻译结果 ec-contrast，和优化基线系统所得到的翻译结果 ec-primary。最终评测结果如表 7 所示。

例如：对于英汉翻译源语言句子：“*Zhao Hongzhu, deputy head of the Central Commission for Discipline Inspection, said recently that more than half of government offices and public institutions are still to be visited by disciplinary investigators.*”，基线系统翻译如下：“*中央纪律检查委员会，一 所近表示，半数以上的政府办公室和公共机构还将参观由管理人员*”。结果显示对人名翻译效果较差，通过优化基线系统，翻译结果为：“*赵,中央纪律检查委员会副处长,最近表示,超过半数的政府部门和公共机构仍然是纪律调查人员访问*”。虽然不能完整翻译出人名，但已经有所提高，同时翻译效果也相对流畅许多。评测结果表明，在英汉新闻领域，本实验组的方法相对基线系统性能有一定提升，对翻译性能提高能起到一定的作用。

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 CWMT-2015 评测的情况。这是 BJTU-NLP 小组第三次参加全国机器翻译研讨会组织的评测。本研究组使用了开源翻译工具 NiuTrans 实现基于层次短语的翻译系统，采用了错误驱动学习方法和命名实体音译的方法改进翻译系统。在训练过程中，加强了训练语料的预处理，降低了噪声干扰，提高了训练语料的质量，实验结果表明，通过优化层次短语模型的翻译规则库和英汉、汉英音译处理，在汉英新闻领域以及英汉新闻领域的翻译结果，相对于基线系统翻译精度得到了一定的改善。

今后，我们将深入研究语法知识和层次短语模型的深度融合、以及基于句法的翻译模型的改良等。

致谢

本论文受国家自然科学基金(61370130 和 61473294)；中央高校基本科研业务费专项资金 (2015JBM033)；国家国际科技合作专项资助(2014DFA11350)资助。

参考文献

1. 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.
2. Chiang D. Hierarchical phrase-based translation[J]. computational linguistics, 2007, 33(2): 201-228.
3. Cao H, Zhang D, Zhou M, et al. Soft Dependency Matching for Hierarchical Phrase-based Machine Translation[J].
4. Eisner J. Learning non-isomorphic tree mappings for machine translation[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2. Association for Computational Linguistics, 2003: 205-208.
5. Huang L. Ten Original Ideas in the Twenty Years of Statistical Machine Translation[J].
6. 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.
7. Koehn P. Statistical machine translation[M]. Cambridge University Press, 2009.
8. 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.
9. Liu Y, Liu Q, Lin S. Tree-to-string alignment template for statistical machine translation[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 609-616.
10. Liu Y, Huang Y, Liu Q, et al. Forest-to-string statistical translation rules[C]//Annual Meeting-Association For Computational Linguistics. 2007, 45(1): 704.
11. Och F J, Ney H. A systematic comparison of various statistical alignment models[J]. Computational linguistics, 2003, 29(1): 19-51.
12. 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.
13. Wang D, Yang X, Xu J, et al. A Hybrid Transliteration Model for Chinese/English Named Entities—BJTU-NLP Report for the 5th Named Entities Workshop[C]//Proceedings of NEWS 2015 The Fifth Named Entities Workshop. 2015: 67.
14. Xiao T, de Gispert A, Zhu J, et al. Effective Incorporation of Source Syntax into Hierarchical Phrase-based Translation[C]//COLING. 2014: 2064-2074.
15. Xiao T, Zhu J, Zhang H, et al. NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation[C]//Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012: 19-24.
16. Zhang M, Jiang H, Aw A, et al. A tree sequence alignment-based tree-to-tree translation model[J]. ACL-08: HLT, 2008: 559.