CWMT2013 哈工大机器智能与翻译研究室技术报告

陈科海[1],史华兴[1],张捷鑫[1],张瑞鹏[1],孙海鹏[1],刘乐茂[1],张春越[1],赵铁军[1]
(1.哈尔滨工业大学,机器智能与翻译研究室 黑龙江 哈尔滨市,150001)

摘要：本文详细介绍了哈尔滨工业大学机器智能与翻译研究室(HIT-MITLAB)参加 2013 年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，我们在开源翻译引擎 Moses 的基础上构建了 2 个翻译系统,它们分别是基于短语的翻译模型和基于层次短语的翻译模型，共参与 CWMT2013 评测中的 6 个子项目：汉英、英汉新闻领域，英汉科技领域，蒙汉日常用语，藏汉政府文献和维汉新闻领域的翻译。在这 6 个子项目中，我们共提交了 21 个翻译结果。
关键字：机器翻译，短语翻译，层次短语翻译

HIT-MITLAB Technical Report for the 2013 China Workshop on Machine Translation

Kehai Chen[1],Huaxing Shi[1],Jiexin Zhang[1],Ruipeng Zhang[1],Haipeng Sun[1],
Lemao Liu[1],Chunyue Zhang[1], Tiejun Zhao[1]

(1. Harbin Institute of Technology,Machine Intelligence and translation Laboratory, Harbin 150001)

Abstract:*This paper describes the details of machine translation and evaluation task for HIT-MITLAB's joining in the China workshop on Machine Translation in 2013(CWMT2013).In this task, we make use of open source translation engine Moses to submit 2 translation systems, which are Phrase-Based Translation Model and Hierarchical phrase-based model; and take part in 6 translation sub-tasks: Chinese-to-English and English-to-Chinese News, English-to-Chinese Science and Technology, Mongolian-to-Chinese Daily Expression, Tibetan-to-Chinese Government Document and Uighur-to-Chinese News. In these sub-tasks, 21 translation results are submitted. Chinese News. In these sub-tasks, 21 translation results are submitted.*
Keywords:*Machine Translation,Phrase-based translation,Hierarchical phrase-based translation*

1 引言

哈工大机器智能与翻译研究室参加了 2013 年全国机器翻译评测(CWMT2013) 中的 6 个子项目：英汉新闻领域翻译评测；汉英新闻领域翻译评测；英汉科技领域翻译评测；蒙汉日常用语新闻领域翻译评测；藏汉政府文献和维汉新闻翻译评测。本文主要对哈工大的 2 个参评系统及其配置，数据的使用和处理进行了全面的描，同时对各个翻译实验进行了说明和分析。

2 系统描述

本次翻译测评中，我们在开源翻译引擎Moses的短语翻译模型和层次短语翻译模型的基础上来构建我们的翻译系统。

基于短语的翻译模型(Koehn et al.,2003)，最小翻译单元为连续的词序列所组成的形式化短语。该系统将给定的源语言句子切分为形式化短语，然后对每个形式化短语进行翻译，最后进行调序并输出翻译结果。该系统的翻译模型特征包含：正反向翻译概率、正反向词汇化概率、词汇化调序特征、词和短语惩罚特征、扭曲特征和语言模型。采用beam-search算法进行翻译解码，利用beam-size控制栈中翻译假设的数目，解码时beam-size设置为200，采用lazy-pruning剪枝策略。

在基于层次短语的翻译模型中(Chiang 2005)，它基于同步上下文无关文法。这些层次规则可以从双语对齐中进行抽取，将规则在句子中所覆盖的单词数限制为7个，切抽取的规则中含有终结符和非终结符的个数不超过5个。所抽取出来的规则为二元规则，即每条规则包含不多于两个非终结符。本系统所包含的特征有：翻译概率、词汇化翻译概率、短语惩罚、glue粘合翻译规则惩罚、单词惩罚和语言模型。系统解码采用cube-pruning的剪枝策略。

3 实验

3.1 数据使用

在本次CWMT2013评测中，将目标语言为汉语评测方向上的训练数据分为A，B两部分，A部分数据用于训练翻译模型也即抽取翻译规则，B部分数据用于训练语言模型。数据A为CWMT2013组织者提供的双语平行语料；数据B为双语平行语料中的目标语言汉语部分，同时还包含sogou2008中文语料（未追加后续sogou2012单语数据）。对于汉英新闻方向，训练数据为

翻译评测项目	训练集句对数		开发集句对数
	CWMT2013组织者句对数	预处理后句对数	
英汉新闻	4543807	3289497	1000
汉英新闻	4543807	3289497	1006
英汉科技	911525	861764	1116
藏汉政府文献	109380	109372	650
蒙汉日常用语	106469	104870	1000
维汉新闻	109922	109821	700

表1: 训练翻译规则的语料数据信息

翻译评测项目	语料名称	词数（过滤后）
英汉新闻	目标语+sogou2008 +sogou2012	85130045+369382933 +383500942
汉英新闻	目标语+routers	89856505+138161291
英汉科技	目标语+sogou2008	22221377+369382933
藏汉政府文献	目标语+sogou2008	1161244+369382933
蒙汉日常用语	目标语+sogou2008	1674124+369382933
维汉新闻	目标语+sogou2008	2094788+369382933

表2: 训练语言模型的语料数据信息

CWMT2013组织者提供的双语平行语料，语言模型为汉英训练语料中的英语部分和CWMT2013提供的路透社单语语料。语料数据的基本信息及其的路透社单语语料。语料数据的基本信息及其在6个评测子项目中的分布如表1，2所述。

我们此次评测任务完全是在CWMT2013组织者提供的训练集数据和开发集数据上完成的。此外，中文语言模型训练用到了CWMT2013组织者提供的搜狗(sogou2008)单语语料,英汉新闻子项目使用了搜狗(sogou2012)的单语语料；英文的语言模型的训练用到了CWMT2013组织者提供的路透社(routers)单语语料。

3.2 关键步骤

英汉科技、汉英英汉新闻方向的语料预处理：首先对中文使用stanford分词工具进行分词，然后采用实验室自主开发的语料过滤工具Jessica和基于困惑度的训练语料分析工具mozzie对这三个方向的训练语料进行过滤分析。后续对中文的预处理操作包括全角转半角、特殊标点符号的转义；后续对英文的预处理操作包括大写转小写和标点符号的分词处理。

维汉、藏汉和蒙汉语料预处理：首先对目标语言利用stanford分词工具进行分词，然后对蒙语和维语分别进行了转拉丁化处理，并尝试了初步切分；对于藏汉语料我们直接使用组织者提供的已切分的藏语语料。后续又对三个方向的语料进行了全角转半角、特殊符号转义以及标点符号分离操作。

语料对齐获取：对齐工具我们采用两种开源词对齐工具：mgiza++和bekerley对齐工具，来获取所有方向上的语料对齐文件。我们在评测中分别从两个对齐文件中提取翻译规则，然后融合在一起构成我们系统的翻译规则。

语言模型构建：由于sogou语言模型训练数据中存在着一些噪音，直接使用对翻译效果的提升不明显，因为调参后该语言模型对应的权重很小甚至为负数。我们适应一种基于困惑度的方法来改变sogou2008语言模型数据的分布。其主要思想就是增大和训练数据目标语言相近的句子的分布，同时降低和目标语言端相差较远的句子的分布。语言模型构建采用SRILM工具，使用训练语料和sogou2008语料创建ngram语言模型，并采用Kneser-Ney进行平滑。各个项目使用的语言模型的情况为：汉英新闻和英汉新闻包含两个9元语言模型，藏汉、蒙汉和维汉皆包含两个7元语言模型。

重排序规则训练：我们在评测中使用了双调序模型，一个是hier层次调序规则，另一个是msd调序规则。使用Moses训练脚本从对应方向的训练语料中获取，用语指导翻译解码。

训练解码：主要训练过程和解码过程使用开源机器翻译引擎Moses。

后处理：在英汉新闻翻译中，我们编写了一个人名、地名的音字转换程序，较好的转换了译文中的未识别的人名和地名。目标语言是中文的译文，去掉词与词之间的空格，目标语言是英文的译文，然后对所有译文中的相关英文缩写等专有名词进行了大小写转换和标点符号的合并处理。最后清除掉所有译文中的未识别词。

3.3 实验环境

硬件：

CPU 品牌：Intel
CPU 型号：Xeon E5-2670
CPU 主频：2.60Ghz
CPU 数量：8
内存容量：62GB

软件：

操作系统类型：Linux
操作系统版本：Centos Release6.0

3.4 实验结果及分析

6个子项目的开发集皆采用本次评测组织者提供的数据：汉英新闻的开发集共1006句；英汉新闻的开发集共1000句；英汉科技的开发集共1116句；蒙汉日常用语的开发集共1000句；藏汉政府文献的开发集共650句；维汉新闻的开发集共700句。英汉新闻和汉英新闻采用Bach-MIRA进行参数调优并采用Moses工具包中的实现，英汉科技、蒙汉日常用语、藏汉政府文献和维汉新闻皆采用Moses工具包中的MERT进行参数调优。

本次汉英、英汉新闻评测项目上，我们按照CWMT2013提出的受限领域评测规则进行训练，在训练过程中使用最大短语长度为10，并对短语分数计算进行了Kneser-Ney平滑，汉英评测项目上使用了训练集目标语言数据和路透社单语数据分别训练了两个7元语言模型，而在英汉评测项目上使用训练集目标语言数据和sogou2008单语数据分别训练了两个7元语言模型。参数调优过程中将kbest设置为300。评测结果如表3所示。

在英汉科技评测项目上，我们使用的都是基于短语的翻译系统，主系统和对比系统的区别在于，主系统使用了7元语言模型，而对比系统使用了9元语言模型。最大短语长度设置为15，并对短语打分进行了Kneser-Ney平滑，参数调优过程中将kbest设置为500。评测结果如表4、5所示。

在蒙汉日常用语评测项目上，我们使用了两种翻译模型，主系统采用基于短语的翻译模型，而对比系统使用拉丁化并随机切分后的蒙语以及目标语言汉语来构建翻译模型。语言模型利用训练数据目标语言汉语和sogou2008单语分别构建7元语言模型，最大短语长度设置为15，并对短语打分进行了Kneser-Ney平滑，参数调优过程中将kbest设置为300。在对比系统中，由于我们对蒙语进行了拉丁化处理以及随机切分操作，在某种程度上破坏了蒙语词固有的语言结构，所以在两个测试集上的分数较低，但这是我们为缓解蒙语形态学变化而进行的初步尝试。行的初步尝试。如表6所示。

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
2009-primary	0.2347	0.2487	7.9325	0.7195	0.7087	0.4942	0.3188	0.2113	0.6363

表3: 汉英新闻项目评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
2009-primary	0.3505	0.3674	0.3011	10.03	10.0413	0.8007	0.6418	0.3684	0.4
2011-primary	0.337	0.3544	0.2922	9.7092	9.7217	0.7834	0.6306	0.3714	0.3812

表4: 英汉新闻项目评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
2011-primary	0.3998	0.4154	0.3488	10.6054	10.6306	0.8386	0.596	0.3008	0.4015
2011-contrast	0.3979	0.4162	0.3497	10.5805	10.6048	0.8339	0.5972	0.3021	0.404
2013-primary	0.3701	0.385	0.3219	9.7469	9.764	0.8043	0.5846	0.3068	0.434
2013-contrast	0.3676	0.3805	0.3174	9.6951	9.7126	0.8014	0.582	0.3065	0.4387

表5: 英汉科技项目评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
2011-primary	0.3821	0.4215	0.3633	7.759	7.775	0.7646	0.3847	0.3334	0.6576
2011-contrast	0.2685	0.2837	0.2311	6.4316	6.4386	0.6908	0.4937	0.4219	0.4969
2013-primary	0.1185	0.1338	0.0938	5.249	5.2501	0.5822	0.6661	0.5458	0.4229
2013-contrast	0.0892	0.0959	0.0638	4.6256	4.626	0.5359	0.7116	0.5984	0.3487

表6: 蒙汉日常用语项目评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
2011-primary	0.5457	0.5755	0.537	9.4895	9.5402	0.8296	0.3697	0.2468	0.5844
2011-contrast	0.5356	0.5609	0.5223	9.4015	9.4505	0.8283	0.3833	0.253	0.5658
2013-primary	0.2362	0.2507	0.2014	7.4183	7.4271	0.7119	0.595	0.3962	0.3097
2013-contrast	0.2315	0.2437	0.1955	7.277	7.2854	0.7135	0.6083	0.4069	0.2927

表7: 藏汉政府文献项目评测结果

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
2011-primary	0.4483	0.4841	0.4249	10.3388	10.3689	0.8135	0.5051	0.3435	0.4864
2011-contrast-a	0.4472	0.4679	0.408	10.6478	10.678	0.8283	0.5098	0.3331	0.4637
2011-contrast-b	0.4362	0.4653	0.4039	10.545	10.5722	0.821	0.5028	0.3356	0.4879
2013-primary	0.4611	0.4866	0.4346	10.1905	10.2281	0.8059	0.4429	0.3081	0.5176
2013-contrast-a	0.4551	0.4767	0.4214	10.3014	10.3345	0.8142	0.4338	0.2894	0.523
2013-contrast-b	0.4499	0.4768	0.422	10.2273	10.2616	0.8084	0.4554	0.3012	0.4936

表8: 维汉新闻项目评测结果

在藏汉政府文献评测项目上，我们使用了两种翻译模型，主系统采用基于短语的翻译模型，而对比系统则使用基于层次短语的翻译模型。语言模型利用训练数据目标语言汉语和sogou2008单语分别构建7元语言模型。主系统中最大短语长度设置为15，对比系统中最大短语长度设置为10，两个系统中对短语打分都进行了Kneser-Ney平滑，参数调优过程中将kbest设置为300。在藏汉政府文献测评中，主系统和对比系统翻译性能比较接近，我们分析是由于藏语和汉语属于同一个语系，具有着相似的语法结构和语言学特点。如表7所示。

在维汉新闻评测项目上，我们使用了一个主系统和两个对比系统，主系统采用基于短语的翻译模型，而对比系统则使用基于层次短语的翻译模型和拉丁化并随机切分后的维吾尔语以及目标语言汉语来构建的翻译模型。语言模型利用训练数据目标语言汉语和sogou2008单语分别构建7元语言模型。主系统中最大短语长度设置为10，层次翻译模型中最大短语长度设置为7，另一个对比系统中最大短语长度设置为15。对短语打分都进行了Kneser-Ney平滑，参数调优过程中将kbest设置为300。如表8所示。

4 总结

在本次评测中我们使用了基于短语的翻译模型和基于层次短语的翻译模型。在 6 个评测项目中，我们的系统性能表现稳定，尤其是我们发现在大多子任务评测项目上，基于短语的翻译模型要比基于层次短语的翻译模型具有较好的翻译性能。在语料预处理中，我们自主开发的 Jessica 和 mozzie 较好地提高了训练语料的质量，为后续模型的建立奠定了良好的基础。同时我们对于小语种的翻译也进行了一些初步的探索，比如拉丁化处理等。此外，通过优化翻译模型，对未识别词的进行了部分处理，提高了翻译性能。本次评测与前次评测相比，我们的翻译结果有了较好的提升。最后通过此次 CWMT2013 评测，我们与国内外同行的交流，为我们今后更好的开展统计机器翻译的研究工作极大的开拓了思路。

References

P.Koehn,F.Och,and D.Marcu. Statistical Phrase-based Translation. In Proc. of NAACL, 2003.

David Chiang. A hierarchical Phrase-based Model for Statistical Machine Translation. In Proc. of ACL, 2005.

Franz Josef Och and Hermann Ney. Improve Statistical Alignment Models. In Pro. of ACL, 2000.

Franz Josef Och. Minimum Error Rate Training In Statistical Machine Translation. In Pro. of ACL, 2003.

Colin Cherry and George Foster. Batch Tuning Strategies for Statistical Machine Translation. In Proc. of NAACL, 2012.

A. Stolcke. SRILM - An extensible language modeling toolkit. In Proc. of ICSLP, 2002.

P.Koehn, H.Hoang, A.Birch, et al. Moses:Open Source Toolkit for Statistical Machine Translation. In Pro. of ACL, 2007.