2013 全国机器翻译评测新疆师范大学技术报告

杨勇，任鸽， 赛买提，阿迪力

新疆师范大学计算机科学技术学院， 新疆 乌鲁木齐 830054

摘要：本文介绍了新疆师范大学参加2013年全国机器翻译研讨会机器翻译评测的情况，本课题组参加的项目是面向新闻领域的维汉机器翻译。文章详细的介绍了系统的主要流程和细节，并对结果进行了分析。

关键字： 机器翻译； 维汉翻译； 短语翻译

中图分类号：TP391 **文献标识码：**A

XJNU Technical Report for the CWMT2013

Yong Yang,Ge Ren,Saimti,Adili

Xinjiang Normal University College of Computer Science and Technology ,Urumqi,XinJiang 830054

Abstract: In this paper, we describe the details of machine translation and evaluation task for XJNU's joining in the China workshop on Machine Translation in 2013(CWMT2013). We take part in Uighur-to-Chinese News MT. In this paper, describe the process of our system in details and anaylyze.the experimental results over the evaluation data.

Keywords: Machine translation, Uighur-Chinese Translation, Phrase-Based Translation

1 引言

新疆师范大学参加了2013年全国机器翻译评测(CWMT2013) 中的维汉新闻领域项目的翻译评测。本文主要介绍新疆师范大学的翻译系统和相关技术以及在维汉新闻领域项目中的性能表现。

2.系统描述

维吾尔语是典型的黏着语，具有丰富的形态学变化，同一个词干链接不同的词缀之后，从而表现出不同的形势。这种形态学变化往往会导致严重的数据稀疏等众多问题。因此使得维汉机器翻译面临着极大的挑战。目前本实验室对于维汉机器翻译的研究还处在起步阶段，本次翻译测评中，我们使用了基于短语的开源机器翻译引擎Moses搭建了一个维汉机器翻译系统。基于短语的翻译模型，将连续的词序列所组成的短语当做最小的翻译单元。短语翻译系统将给定的源语言句子切分为单词序列，然后对每个序列以及所组成的短语进行翻译，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：翻译概率、词汇化概率、词汇化调序特征、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 100。语言模型为两个5元语言模型，一个使用训练语料的目标语言构建，一个使用sogou单语语料构建。主系统中最大短语长度设置为10。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200。使用MERT进行参数调优。

3.数据处理及使用

在维汉新闻领域项目的测评中，我们使用的是CWMT2013官方所提供的109921句对训练语料。维吾尔语的处理包括老维文到拉丁维文的转码、Tokenization、去空格、处理不合理符号等。语言模型采用维汉训练集目标语言和sogou单语语料数据，利用srilm工具训练了两个5元语言模型。开发集使用CWMT2013提供的700句开发集进行模型调优。本次测评所使用的训练集、开发集和语言模型数据信息如表1。

翻译评测项目	训练集句对数	开发集句对数	语言模型(目标语+sogou)词数
维汉新闻	109722	700	85051021+354364529

表1：训练翻译规则的语料数据信息

4.实验

4.1 实验环境

本系统的运行环境：操作系统为 Ubuntu13.10 版的 Linux 平台；采用的 CPU 为 Intel 酷睿 i5,四核，主频 3.2G，内存 4G。

4.2 规则获取

维汉语料预处理：我们使用stanford分词工具对目标汉语进行分词，然后对维语分别转拉丁化处理，最后对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：我们利用开源词对齐工具giza++来获取维汉双语语料的对齐文件。

建立语言模型：语言模型构建采用SRILM工具，使用训练语料和sogou语料创建ngram语言模型，并采用Kneser-Ney进行平滑。

调序模型建立：在评测中使用Moses训练脚本从对应方向的训练语料中获取，用于指导翻译解码。

后处理：此次我们未对结果进行任何处理，只是剔除了空格之后直接提交结果。

4.3 实验结果及分析

实验结果如表2所示。

System	BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT	METEOR	TER
2011-primary	0.2966	0.3156	0.264	7.0202	7.0342	0.7392	0.6253	0.4591	0.4301
2013-primary	0.2339	0.2443	0.2036	5.4249	5.4349	0.6711	0.6377	0.4919	0.4019

表2 在维汉新闻项目上的评测结果

相比前三名的成绩我们的翻译结果得分较低。对实验结果进行详尽的分析后，发现影响

实验成绩主要有两个方面：一、对语料的处理不够统一，如在训练语料和测试集处理过程未能统一。二、对译文未进行后处理，如未剔除未识别词。

5 总结

在本次评测中我们使用了基于短语的翻译模型。在维汉新闻项目评测中，我们的系统性能表现不够理想，经过分析主要是我们对语料的预处理做的还不够细致，尤其是我们未对测试集译文进行后处理操作，直接导致测试成绩较低。这次我们本着想国内同行学习的态度，以便我们能够更好的对维汉机器翻译展开更加细致的研究。

参考文献

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

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

[3]A. Stolcke. 2002. SRILM - An extensible language modeling toolkit. In Proc. of ICSLP.

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

[5]http://code.google.com/p/giza-pp/downloads/list