新疆大学 CWMT2015 评测技术报告*

买合木提•买买提，卡哈尔江•阿比的热西提，吐尔根•依布拉音，艾山•吾买尔，艾山•毛力尼亚孜，麦热哈巴•艾力，

（1. 新疆大学信息科学与工程学院，乌鲁木齐 830046；2. 新疆多语种信息技术重点实验室，乌鲁木齐 830046）

摘要：本文介绍了新疆多语种信息技术重点实验室参加 2015 年全国机器翻译研讨会机器翻译评测的情况。今年我们参加了面向新闻领域的维汉机器翻译的评测任务。使用了基于短语的统计翻译模型的单系统。文章主要介绍了我们参加的评测任务的系统框架、模型及评测结果。

关键词：基于短语翻译模型；Moses；CWMT2015；维吾尔语

XJU Evaluation Technical Report for CWMT'2015

Maihemuti Maimaiti, Kahaerjiang Abiderexiti, Tuergen Yibulayin, Aishan Wumaier ,Aishan Maoliniyazi, Mairehaba Aili,

(1. School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, 830046, China; 2. Xinjiang Laboratory of Multi-language Information Technology, Urumqi, Xinjiang, 830046, China)

Abstract: This paper describes the XJU systems involved in the CWMT 2015 evaluation campaign. This year, we participated in one evaluation task: Uyghur-to-Chinese News. Phrase-based statistical machine translation system was used. The paper briefly describes the primary modules, the framework of implementation, and the evaluation results.

Key words: Phrase-based model; Moses; CWMT2015; Uyghur

1 引言

2015 年中国机器翻译研讨会（CWMT2015）机器翻译评测共含6个评测项目，分别为：汉英新闻领域机器翻译（CE）、英汉新闻领域机器翻译（EC）、蒙汉日常用语机器翻译（MC）、藏汉政府文献机器翻译（TC）、维汉新闻领域机器翻译（UC）、双盲评测机器翻译（DB）。本实验室参加了其中的维吾尔语—汉语新闻领域机器翻译的测评。在测评过程中我们只分别使用了测评组织方提供的语料和我们自己建立的语料，系统为基于短语的Moses翻译系统。我们提交了两个评测结果，主评测结果为评测组织方提供的语料，而对比评测结果为新疆大学维汉平行语料库加上评测组织方提供的语料库。下面我们介绍主要采取的方法和实验过程。

2 参评翻译系统概述

这次机器翻译评测中我们使用了基于短语的Moses统计机器翻译系统。Moses 是基于短语和层次短语的翻译模型，为开源系统；我们用该开源系统分别搭建了两个系统，一个系统叫做主（Primary）系统，另一个系统叫做对比（Contrast）系统。

2.1 Moses

Moses[Koehn et al.,2007]是由英国爱丁堡大学、德国亚琛工业大学等8 家单位联合开发的一个基于短语的统计机器翻译系统。

它的主要特点包括：

（1）基于beam search 的解码；

（2）基于短语的翻译模型；

（3）基于要素（factor）。

基金项目：国家重点基础研究发展计划（973）资助项目（2014cb340506）国家自然科学基金（61331011, 61462083,61262060）；国家社科基金 (10AYY006) ；

3 数据处理方法及工具

3.1 语料的预处理

我们对于主系统对所有语料（双语训练集，开发集，测试集，语言模型训练集）进行了如下预处理，预处理的步骤如下：

维吾尔文：

1、控制符和乱码去除处理；

2、Tokenization；

3、词干提取[2-6]。

中文：

1、全角转半角；

2、中文分词。

对于对比系统的双语训练集，开发集，测试集，语言模型训练集进行了如下预处理，预处理的步骤如下：

维吾尔文：

1、控制符和乱码去除处理；

中文：

1、全角转半角；

2、中文分词。

3.2 分词及词干提取

对汉语端使用ICTCLASS 2015 工具进行了分词。对维吾尔语使用新疆大学研发的分词和词干提取工具进行了分词和词干提取[2-6]。即先用工具分词和词干提取然后用机器翻译工具进行训练。

3.3 语言模型

中文语言模型为基于词的5元语言模型，语言模型的训练语料为搜狗全网新闻语料库，加上参与评测项目的训练集中对应的目标语言部分。

3.4 训练集和测试集

Primary 系统中的训练、开发和测试语料均来自评测主办方发布的语料，下面列出本系统使用的语料情况：

表 1 主系统语料情况表

原始规模	处理后规模	开发集	测试集
139925	139508	700	1000

Contrast 系统中的训练为新疆大学平行语料库加上评测方提供的训练语料库、开发和测试语料均为评测方发布的语料，下面列出 Contrast 系统使用的语料情况：

表2 对比系统语料情况表

原始规模	处理后规模	开发集	测试集
337362	336998	700	1000

4 实验结果

我们主评测结果和对比评测结果以及详细结果分别如表3，表 4 所示：

表3 维汉新闻领域机器翻译结果

参评系统	BLEU5-SBP	BLEU5	NIST6
主系统	0.3733	0.3866	9.374
对比系统	0.2886	0.3034	8.6228

表 4 维汉新闻领域机器翻译详细结果

结果 / 指标	主系统	对比系统
BLEU5-SBP	0.3733	0.2886
BLEU5	0.3866	0.3034
BLEU6	0.3390	0.2581
NIST6	9.3740	8.6228
NIST7	9.3874	8.6309
GTM	0.7682	0.7126
mWER	0.5013	0.5848
mPER	0.3697	0.4109
ICT	0.4590	0.3712
METEOR	0.5507	0.4940
TER	0.4505	0.5167

从表 3，表 4 可以看出，主系统的各个评测结果都比对比系统高。我们初步分析得导致这种结果的主要原因是，在对比系统中虽然语料规模比主系统大，但对维吾尔语没有进行词干提取，因此翻译质量

不如主系统。今后我们进一步研究该问题出现原因以及探索该问题的解决方法，进一步提高翻译质量。

5 讨论

本文主要介绍了新疆大学参加CWMT2015 评测的机器翻译系统的试验情况和测试结果，本次活动中我们参加了面向新闻领域的维汉机器翻译的评测任务。我们以基于短语的开源机器翻译系统Moses 作为我们的翻译的受限主系统，并以自己收集的平行语料库为主的基于Moses 短语的统计机器翻译系统作为非受限对比系统。

由于我们这次以人员以及时间等多种原因没能充分发挥我们组的实际势力，对此我们表示遗憾。总之，我们希望通过参加这次的测评，进一步沟通其他国内以及国外的研究机构，交流经验，进一步改进和完善自己的系统。

6 致谢

本项目得到国家重点基础研究发展计划（973）资助项目（2014cb340506）国家自然科学基金（61331011, 61462083,61262060), 国家社科基金(10AYY006)。

参考文献

[1] P. Koehn *et al.*, “Moses: open source toolkit for statistical machine translation,” in Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 2007, pp. 177-180.

[2]早克热·卡德尔 等., “混合策略的维吾尔语名词词干提取系统,” 计算机工程与应用, no. 01, pp. 171-175, 2013.

[3]艾山·吾买尔, 吐尔根·依步拉音,早克热·卡德尔, “基于噪声信道的维吾尔语央音原音识别模型,” 计算机工程与应用, vol. 46, no. 15, pp. 118-120,192, 2010.

[4]早克热·卡德尔 等, “维吾尔语名词构形词缀有限状态自动机的构造,” 中文信息学报, vol. 23, no. 6, pp. 116-121, 2009.

[5] A. Wumaier *et al.*, "Conditional Random Fields Combined FSM Stemming Method for Uyghur," *2009 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2009.* pp. 295-299.

[6] A. Wumaier *et al.*, "Maximum Entropy Combined FSM Stemming Method for Uyghur," *2009 Orental COCOSDA International Conference on Speech Database and Assessments, ICSDA 2009.* pp. 51-55.