新疆大学 CWMT2013 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•毛力尼亚孜，麦热哈巴•艾力，艾山•吾买尔

（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 2013 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；CWMT2013；维吾尔语

XJU Evaluation Technical Report for CWMT'2013

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

(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 2013 evaluation campaign. This year, we participated in one evaluation task: Uyghur-to-Chinese News. Phrase-based statistical single machine translation system was used. The paper briefly describes the primary modules, the implement framework, and the evaluation results.

Key words: Phrase-based model; Moses; CWMT2013; Uyghur

1 引言

2013 年中国机器翻译研讨会（CWMT2013）机器翻译评测共含六个项目，分别为：英汉新闻领域、汉英新闻领域、英汉科技领域、维汉新闻领域、蒙汉日常用语领域、、藏汉政府领域。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，以 Moses 单系统作为主要参评系统，自己开发的基于短语的翻译系统作为对比系统进行翻译。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两种统计机器翻译单一系统：Moses 和自己开发的基于短语的统计机器翻译系统。其中，Moses 是基于短语和层次短语的翻译模型，为开源系统；自己开发的基于短语的统计机器翻译系统是根据丝路系统开发的自建系统，它集成了命名实体识别，时间词、量词、数词识别及翻译等模块。

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）。

2.2 基于短语模型自建翻译系统

在人名翻译中，先利用我们用C#语言开发的基于规则和词典的人名识别程序，

基金项目：国家自然科学基金（61063026, 61063043, 61262060, 60963018）；国家社科基金 (10AYY006)； 新疆维吾尔自治区青年基金 (2011211B07)；新疆多语种信息技术重点实验室开放课题 (049807)；新疆大学博士启动基金

对语料中的人名进行识别，并通过基于词典的人名翻译方法对人名进行翻译。人名识别过程中虽然我们都对汉族人名、维吾尔族人名以及部分国家的外国人名，但对其进行翻译时因汉族人人名和外国人人名翻译语料库的规模不大，且我们系统根据词对齐结果基本上翻译所以我们只对维吾尔族人人名进行翻译，且只对系统没有进行翻译的人名进行翻译，但在评测语料中维吾尔族人人名的出现频率很低，占语料库的0.05%左右（我们人名识别程序的识别结果），所以维吾尔人人名翻译对系统几乎没有贡献。对于地名、数词、量词和时间词的翻译我们现阶段只通过简单的词典匹配方法，词典的规模也很小。我们人名、地名、数词、量词翻译方面今后继续探索。

3 数据处理方法及工具

3.1 语料的预处理

我们对所有语料（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。

3.2 分词及词干提取

对汉语端使用 ICTCLASS 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。

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 语料情况表

原始规模	处理后规模	开发集	测试集
109895	109527	1000	1000（current），574（progress）

Contrast（自建系统）系统的训练集使用的是新疆大学建立的 12 万条句子对齐语料库，开发和测试语料均来自评测主办方发布的语料。

4 实验结果

我们主系统和对比系统分别在 Progress 和 Current 中的测试结果如表 2 所示：

表 2 维汉新闻领域机器翻译结果

	测试集 2011	测试集 2013
参评系统	BLEU5-SBP	BLEU5-SBP
Moses	0.4512	0.4564
自建系统	0.4868	0.2354

表 3 维汉新闻领域机器翻译详细结果

	测试集 2011		测试集 2013	
系统	Moses	自建系统	Moses	自建系统
BLEU5-SBP	0.4512	0.4868	0.4564	0.2354
BLEU5	0.4831	0.4995	0.4741	0.2529
BLEU6	0.422	0.4542	0.4223	0.2043
NIST6	10.4296	10.7834	9.9658	7.6012

NIST7	10.4591	10.8104	10.001	7.6091
GTM	0.826	0.8093	0.8081	0.6631
mWER	0.5034	0.4955	0.4441	0.6498
mPER	0.3306	0.3414	0.3021	0.4558
ICT	0.5024	0.4045	0.5325	0.2836
METEOR	0.4371	0.3906	0.5311	0.3925
TER	0.4215	0.4494	0.3723	0.5706

从表 3 可以看出，基于 Moses 的短语翻译模型（Primary 系统）分别在 2011 和 2013 年度测试集上的表现比较稳定，自建系统（Contrast 系统）的表现差异较大，即自建系统 2011 年测试集上的表现明显好于 2013 年度测试集上的表现。造成这个结果的主要原因是自建系统使用的训练语料领域、时代与测试集的吻合度较低。

表 4 CWMT2013 维汉新闻人工评测结果（Current 测试集）

忠实度平均值	2.7706668
流利度平均值	3.0966685
总体平均值	2.93366765
评价句子总数	500

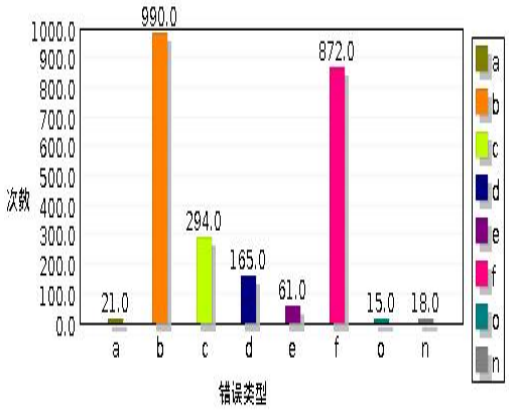

图 1 错误类型分析柱状图

错误类型说明：

a 类错误：译文和原文意思相反

b 类错误：实词译文缺失

c 类错误：词序不对

d 类错误：命名实体问题

e 类错误：数量词/时间词问题

f 类错误：译文选词错误

o 类错误：其它错误

n 类错误：没有错误

从图 1 可以看出，在人工评测当中我们主系统的实词译文缺失错误和译文选词错误占很大比例。我们初步分析发现，导致这种大量的错误的主要原因是维吾尔语和汉语的语序结构不同，容易出现词对齐结果不正确情况，从而导致短语表以及解码的错误。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 CWMT2013 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，并自己开发的基于短语的统计机器翻译系统作为非受限对比系统。该自建系统是根据丝路系统在.NET 上用 C#开发的，它集成了命名实体识别，时间词、量词、数词识别及翻译等模块。

由于我们这次又是像 CWMT2011 一样既是数据提供单位，是参评单位，且评测期间我们发现训练语料存在一些问题，因此我们评测期间对训练语料进行修改并重新发布，这对影响了我们的人力和时间。

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

本项目得到国家自然科学基金(61063026,61063043,61262060,60963018),国家社科基金 (10AYY006), 新疆维吾尔自治区青年基金 (2011211B07)新疆多语种信息技术重点实验室开放课题 (049807)和新疆大学博士启动基金的资助。

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.