新疆大学 CWMT2011 评测技术报告

麦热哈巴.艾力， 米日古.肉孜，撒依达，江阿古丽，吐尔根.伊布拉音
新疆大学 乌鲁木齐 830046
Email: marhaba@xju.edu.cn

摘要： 本文主要介绍了我们参加 CWMT2011 研讨会的参评系统和技术报告，我们参加的项目是面向新闻领域的维汉机器翻译、面向新闻领域的哈汉机器翻译以及面向新闻领域的柯汉机器翻译等三个项目。本文中详细的介绍了系统的主要流程以及细节。

关键词： 统计机器翻译， 语言模型， 词法分析

Technical Report of Xinjiang University for CWMT2011 Evaluation

Marhaba.eli, Mihrigul.rozi, Sayida, Janagul, Turgun.ibrayim
Xinjiang University Urumqi 830046
Email: marhaba@xju.edu.cn

Abstract: *In this paper ,we describe our SMT system for CWMT2011 and technical report. We take part in 3 subitems: Uyghur-Chinese MT , Kazak-Chinese MT and Kirgiz-Chinese MT and they all are based on News domain. In this paper, we describe the process of our system in detail .*

Keywords: *Sitatistical Machine Translation, Language Model , POS*

1 引言

随着机器翻译技术的发展，新疆大学多语种重点实验室也开始尝试少数民族语言与汉语之间的机器翻译研究。虽然我们研究工作起步不久，但为了尽快地了解国内外机器翻译研究的进展与评测方法，以便尽快地与国内外相关事项的研究接轨，本实验室参加了 2011 年中国机器翻译研讨会（CWMT2011）机器翻译评测。本次评测共设置了 8 种评测项目，本实验室参加了其中的维吾尔语—汉语、哈萨克语—汉语、柯尔克孜语—汉语的等三个语种的测评。在测评过程中我们只使用了测评组织方提供的语料资源，以 Moses 单系统作为主要参评系统，下面我们介绍主要采取的方法和实验过程。

2 系统描述

本次测评中，我们以基于短语的机器翻译开源系统——Moses 作为主要的参评系统，在不同的语种上，对它进行预处理以及改变了创建语言模型语料的规模。所有的系统中，对汉语端使用 ICTCLASS 工具进行了分词。下面详细介绍各个环节：

2.1 维汉机器翻译系统流程的描述

维吾尔语是典型的黏着语，具有丰富的形态变化，同一个词干在不同词缀的连接之下，由于维吾尔语的语音和谐规律现象，会出现某些字母的弱化、脱落或增音等现象，从而会显现出不同的形式，如：alma（苹果），almini（把苹果）。这现象往往会带来数据稀疏等众多问题，所以对词干进行还原是必要的。针对此问题，我们通过词法分析器对维语端进行了词法分析，把词干与各个词缀互相分离，同时将对词干进行了还原。

维吾尔语词缀在一个句子的生成中起到很重要的角色，大部分都带着“意思”，在汉语端会被翻译出来。但为了适应语音和谐规律，很多词缀都有变体。所谓的“变体”意思就是词缀的同义异形，如：表示复数的词缀“lar，ler”表示同一个意思，但是形式不同。对于维吾尔语词缀的这种现象，我们采取了将词缀抽象化表示的方法。

同时，维吾尔语词缀连接词干后起到的作用不同，有些词缀仅起到和谐维吾尔语词类的作用，不会被翻译出来；有些词缀具有实际意义，在汉语端会被翻译出来，如：herbiy rayoni 与 herbiy rayon 都会被翻译成“军区”等等。为了降低出现数据系数等现象的出现，我们对训练集使用 GIZA++进行对齐后，统计出各个词缀被翻译出来的概率并将经常不被翻译的词缀进行过滤操作，从而试图降低它们在词对齐以及短语抽取时的噪音。

最后，为了体现这些方法在机器翻译过程中起到的作用，我们分别构成了三个系统：对源语言进行 tokenization，而不进行任何其他分析，构成我们的主系统即：primary system a；对源语言进行词法分析，对词缀采取抽象化表示并将被翻译成空的概率高的词缀过滤掉后重新将剩下的词缀连接到词干上作为新的 token，构成我们的对比系统即：contrast system b；我们的第二个对比系统与第一个对比系统不同点是将词干与词缀不合并，而是分别看成是一个 token。

三个系统使用的语言模型都是一致的，即：我们从组评单位提供的搜狗语料中（2.08G）整理 139M 大小的语料在家训练语料 5 万个句对，用 SRILM 工具训练出了 5 元的语言模型。

2.2 哈汉机器翻译系统的描述

本实验室对哈汉机器翻译的研究还处在初步阶段，牵扯到这方面基础研究还处在试验阶段，这次测评中我们对哈语没有做任何其他处理，只是对哈语端进行了 tokenization 操作。我们通过更改语言模型的训练语料规模来构成主系统和对比系统，即：以训练集的 5 万句对训练出哈语的 5 元模型构成主系统即：primary system a；以搜狗语料（139M）加哈语训练集 5 万句对用 SRLIM 工具训练出 5 元模型，构成对比系统即：contrast system b；

2.3 柯汉机器翻译系统的描述

柯汉机器翻译的过程与哈汉机器翻译的过程是类似，不同点是：用搜狗语料（139M）与 5 万句对的训练语料训练出了 5 元的语言模型构成主系统即：primary system a；只用 5 万句对的训练语料构成了 5 元的语言模型构成了对比系统即：contrast system b。

3 实验

3.1 实验数据的准备

我们的训练集使用的是本次测评组织方提供的 5 万句对的训练语料，除此之外我们没加任何其他语料。对原始语料进行过滤、去重后的训练语料规模为以下表所示：

语种	领域	原始规模	处理后规模	说明
维汉	新闻领域	5 万句对	49，812 句对	UTF-8
哈汉	新闻领域	5 万句对	49，966 句对	UTF-8
柯汉	新闻领域	5 万句对	49，974 句对	UTF-8

3.2 维汉系统

本次评测（CWMT2011）中维汉系统的测试结果为：

系统	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
primary system a	0.4177	0.4428	0.3854	10.2294	0.78	0.4991	0.4991	0.3459	0.4272
contrast system b	0.3969	0.4182	0.3569	10.2761	10.2964	0.7994	0.5378	0.3434	0.432
contrast system c	0.2377	0.2445	0.2035	4.188	4.194	0.6438	0.6287	0.5263	0.4076

3.3 哈汉系统

本次评测（CWMT2011）中哈汉系统的测试结果为：

系统	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
primary system a	0.3702	0.409	0.3563	8.6584	8.6732	0.7295	0.5283	0.4027	0.4125
contrast system b	0.3725	0.4097	0.3563	8.6491	8.6637	0.733	0.5228	0.4024	0.4195

3.4 柯汉系统

本次评测（CWMT2011）中柯汉系统的测试结果为：

系统	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
primary system a	0.413	0.4407	0.3908	9.6019	9.6239	0.7591	0.4901	0.3377	0.3971
contrast system b	0.4072	0.4319	0.3806	9.5197	9.5397	0.7661	0.4772	0.3353	0.4116

4 讨论

本文主要介绍了新疆大学参加 CWMT2011 活动的机器翻译系统的试验情况和测试结果，本次活动中我们总共参加了维汉、哈汉以及柯汉这三个语种的翻译测评，我们以基于短语的开源机器翻译系统摩西（Moses）作为我们的翻译平台。虽然本实验室也在开发自己的翻译系统，但是由于还处在开发阶段而没来得及参加这次的测评。

整个参评过程中，我们采取的方法主要是对源语言端进行预处理以及语言模型训练语料规模的更改，而整个过程中系统参数主要采取了默认值。测评成绩出来后，才发现我们操作中的一些失误，比如：本打算对维汉翻译的主系统中采取将词缀抽象化的方法，而实际上未采取；对哈汉翻译系统的参数训练过程中发现当语言模型的训练语料规模为 5 万句对（组评单位提供的训练语料）时的 BLEU 值高于加搜狗语料后的值，而测评后得到的值恰恰相反。由于我们这次既是数据提供单位，又是参评单位，所以人力和时间都受到了一定的影响。

总之，我们希望通过参加这次的测评，能够与国内其他研究机构进一步沟通，努力学习，积累经验，取长补短，既要早点与国内外研究接轨，又要为进一步改进和完善自己的系统而努力。

参考文献

Necip Fazil Ayan and Bonnie J. Dorr. A maximum entropy approach to combining word alignments[C]. In

NAACL 2006(6): 96–103.

Franz Josef Och. Minimum error rate training in statistical machine translation [J]. In ACL '03 :160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation[C]. In ACL '02: 311–318.

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. Weighted alignment matrices for statistical machine translation[C]. In EMNLP '09, pages 1017–1026, August.

K.Oflazer, Ilknur Durgar El-Kahlout. Exploring Different Representational Units in Eghlish-to-Turkish Statistical Machine Translation[C]. in Proc.2nd workshop on Statistical Machine Translation. 2007(6):25-32