CWMT 2015 中国科学院新疆理化技术研究所评测报告

杨雅婷[1,2]，米成刚[1,2]， 董瑞[1,2,3]， 吐尔洪•吾斯曼[1,2,3]， 王磊[1,2]， 周喜[1,2]

（1.中国科学院新疆理化技术研究所，乌鲁木齐， 830011；
2. 新疆民族语音语言信息处理重点实验室， 乌鲁木齐， 830011；3. 中国科学院大学，北京，100049）

摘 要：本文详细介绍了中国科学院新疆理化技术研究所多语种信息技术研究室参加 2015 年全国机器翻译研讨会（CWMT 2015）翻译评测任务的情况。本次评测中，我们使用了 Moses_PR_Primary、Moses_PR_Contrast1 和 Moses_PR_Contrast2 三个机器翻译系统，参与了维汉新闻领域的受限和非受限两个评测项目，提交了测试集受限、非受限翻译结果。本文对评测所用系统的配置、数据的处理以及系统运行软硬件环境进行了详细的介绍，并对实验结果进行了分析。

关键字：机器翻译，维汉，短语翻译，层次短语翻译

XJIPC Technical Report for the CWMT 2015

YANG Yating[1, 2], MI Chenggang[1, 2], DONG Rui[1, 2, 3], TURGUN Osman[1, 2, 3],
WANG Lei[1, 2], ZHOU Xi[1, 2]

(1.Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; 2. Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China; 3. University of the Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper describes the details of machine translation and evaluation task for XJIPC's joining in the China workshop on Machine Translation in 2015 (CWMT 2015). In this task, we submit 3 translation systems, which are Moses_PR_Primary, Moses_PR_Contrast1 and Moses_PR_Contrast2, and take part in 1 translation sub-task, Uyghur-to-Chinese news. XJIPC submitted 3 translation results in the sub-task. This paper gives a detailed description of the translation systems, their configurations, the data and its processing methods, and gives the comparisons and analysis of the translation results.

Keywords: Machine translation, Uyghur-to-Chinese, phrase‐based translation, hierarchical phrase‐based translation

1 引言

2015 年全国机器翻译研讨会（CWMT2015）机器翻译评测共含 6 个项目，分别为：汉英新闻领域机器翻译、英汉新闻领域机器翻译、蒙汉日常用语机器翻译、藏汉政府文献机器翻译、维汉新闻领域机器翻译以及双盲评测机器翻译。

中国科学院新疆理化技术研究所多语种信息技术研究室参与了 CWMT 2015 评测项目中的维汉新闻领域翻译评测。CWMT 2015 机器翻译评测方案与上届评测（CWMT 2013）相比有较大变化：本次评测取消了上届评测中的灰箱测试和人工评测，增加了双盲评测（Double Blind Evaluation）项目。在双盲评测中，评测组织方将会选择一个语言对进行加密编码，呈现给各个参评单位的将是加密后的双语对照数据和句法树。这次的双盲测试将最大限度地抛开词法分析、句法分析、命名实体翻译等预处理和后处理对翻译的影响，将最大限度地关注翻译模型本身的问题。此外，评测组织方不再提供评测项目的“基线系统（Baseline System）”及相应的关键步骤中间结果文件。本文详细地介绍了中国科学院新疆理化技术研究的 Moses_PR_Primary、Moses_PR_Contrast1

和 Moses_PR_Contrast2 三个参评系统与实验数据的处理，并对实验结果进行了分析。

2 参评系统介绍

本次评测中，本单位共使用了三个统计机器翻译系统：Moses_PR_Primary、Moses_PR_Contrast1 和 Moses_PR_Contrast2。其中前者是主系统，后两者是对比系统。

2.1 Moses_PR_Primary

Moses_PR_Primary 是在对语料中维吾尔语词进行词干切分[1]等处理的基础上使用开源工具 Moses[2][3][4]搭建的维汉翻译系统。本系统使用 Moses 中的短语翻译模型[5]。在语言模型的训练过程中，元数设为 5 元；翻译模型训练时，最大短语长度设为 11，在进行解码时栈大小使用默认值。对开发集进行调参的过程中，抽取前 100 句（k-best）解码结果作为候选翻译。。

2.2 Moses_PR_Contrast(1&2)

Moses_PR_Contrast 使用开源工具 Moses 中的短语模型进行维汉机器翻译，以词干为翻译单位。在语言模型的训练过程中，元数设为 5 元，最大短语长度取 11。对开发集进行调参的过程中，抽取前 100 句解码结果作为候选翻译。对比系统中，我们使用了规模较大的训练语料获得翻译模型。另外，两个对比系统 Moses_PR_Contrast1 和 Moses_PR_Contrast2 在语料预处理阶段（汉语端分词等）采用了不同的方法，因此，获得了不同的翻译结果。

表 1：参评系统

翻译评测项目	系统编号	primary/ contrast
维汉新闻领域	Moses_PR_Primary	primary
	Moses_PR_Contrast1	contrast1
	Moses_PR_Contrast2	contrast2

3 实验

3.1 数据处理及其模型训练

（1）训练数据

本次评测我们参加了维汉新闻领域翻译的受限和非受限两个任务，所使用的语料可分为两类：受限语料和非受限语料，每一类语料又包括语言模型训练和翻译模型训练两个部分。受限语料的数量如下表所示。

表 2：翻译模型训练语料

翻译评测项目	句对数	源/目标语言词数
受限语料	139,792	2,762,094/2,755,436
非受限语料	300,000	7,359,296/7,818,017

受限语料，从评测组织方发布的 139,297 句对维汉平行语料中，删除过长、过短语料，保留 132,552 句对。

非受限语料，包含受限系统的所有语料，另外增加了本单位采集建设的维汉双语平行语料库资源，增加部分均为新闻语料。非受限语料共计 300,000 句对。

表 3：语言模型训练语料

翻译评测项目	语料名称+句子数	词数
维汉新闻领域	目标文件+139,792	2,755,436
维汉新闻领域（非受限）	目标文件+300,000	7,818,017

语言模型训练只采用了训练语料的汉语部分，未加入搜狗新闻语料。

（2）语料预处理

维吾尔语属于形态复杂语言，其每个词的变化形式最多都达到数百种之多。相比而言，汉语词几乎没有形态变化。对于形态变化相对比较简单的语言，可以基本上不考虑词形变化，把每个不同词形的词都当成独立的词语来

考虑即可。但对于复杂形态的语言，这种做法就会带来比较严重的问题。

维吾尔语通过词根和附加词缀构成多种词形，以表达不同的语法意义。据统计，维吾尔语有 3 万多个词根，100 多个后缀， 10 几个前缀，有的词根后面可以附加几个后缀，虽然并不是所有的词根和后缀的组合都是符合语法规则的，但这却暗示着维吾尔语的词汇量非常巨大，形态变化丰富。因为这类语言的词语变化非常灵活，形式多样，这样会导致翻译时出现大量未登录词（Out of Vocabulary，OOV），严重影响机器翻译的性能。

为解决数据稀疏和未登录词多的问题，对于参加维汉新闻领域翻译的双语语料进行了如下预处理：

维吾尔语端：

对维吾尔语端进行了 token，针对部分任务进行了维吾尔语词干切分。

首先，进行维吾尔语词 token：

1）语料里的非维吾尔文标点符号替换成相应的维吾尔文标点符号；

2）处理维吾尔语对偶词（对偶词是指有着相反或相近意义的字组成的词语，如：父母（ئاتا-ئانا））

其次，维吾尔语词词干切分 stemming：

维吾尔语是一种拼音式文字，基本特征表现在：存在元音和谐律、元音弱化现象；构词和构形附加成分很丰富；名词有数、从属人称、格等语法；词汇中除有突厥语族诸语言的共同词外，还有相当数量的汉语、阿拉伯语、波斯语和俄语的借词。针对维吾尔语自身特点，可以进行维吾尔语词干切分：

<1> 从词首取 N 个字母作为候选词干跟词干库匹配（N 从 1 开始）。

<2> 如匹配不上再检测元音弱化、脱落、增加等现象，并还原候选词干。

<3> 候选词干再跟词干库匹配。

<4> 所有跟词干库匹配的词干集合作为结果返回。

3）维吾尔语语料拉丁化

采用统计学习的思想，获取适应本次翻译任务的维吾尔语拉丁化规则。

汉语端：

首先，对汉语语料中的非汉字字符进行统一化处理；

其次，对语料进行全半角转换；

最后，使用分词工具对汉语语料分词。

（3）语言模型训练

语言模型是针对某种语言建立的概率模型，使得正确词序列的概率值大于错误词序列的概率值。语言模型主要应用在机器翻译、语音识别、拼写检查等自然语言处理领域。

对于词序列 $w_1 w_2 \cdots w_m$，其语言模型概率用 $P(w_1 w_2 \cdots w_m)$ 来表示：

$$P(w_1 w_2 \cdots w_m) = P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1 \cdots w_{i-1})$$

由于语言现象非常复杂，训练语料不能包含所有的语言现象，这就导致了数据稀疏问题。统计语言模型研究中提出了数据平滑技术来解决数据稀疏问题。主要的平滑技术包括：加 1 平滑、Good-Turing 平滑、插值平滑等。本次评测中语言模型训练采用插值平滑算法进行平滑。

插值平滑：将高阶模型和低阶模型做线性组合。参数 λ 参数 EM 算法进行估计

$$p_{inter}(w_i \mid w_{i-n+1}, \cdots, w_{i-1}) = \lambda \times p_{inter}(w_i \mid w_{i-n+2}, \cdots, w_{i-1}) + (1-\lambda) \times p_{inter}(w_i \mid w_{i-n+1}, \cdots, w_{i-1})$$

在本次评测中对参与语言模型训练的语料进行分词处理，使用开源工具 SRILM[6]进行语言模型的训练，经过前期的一系列实验，将语言模型元数设为 5 元，选用插值的方式进行平滑。

（4）翻译模型训练

本次评测主要用到了基于短语的翻译模型。模型主要包括以下几个模块：

1）词对齐模块

该模块首先利用词对齐工具进行源语言到目标语言和目标语言到源语言的双向词对齐。从双向词对齐的交集开始，采用启发式的方法向双向词语对齐的并集进行扩充。

2）词语评分模块

采用最大似然方法，根据上述词对齐结果，计算出词语之间翻译的最大概率。

3）短语抽取模块

短语抽取方法主要是对对齐矩阵中所有以对齐点为顶点的矩形区域进行提取。短语抽取原则为：与矩形所在行范围内的源语言词对齐的目标语言词也在这个矩形区域的列范围之内，反之亦然。

4）短语评分模块

计算五个概率：源语言到目标语言短语翻译概率，目标语言到源语言短语翻译概率，源语言到目标语言词典翻译概率，目标语言到源语言词典翻译概率以及短语惩罚概率。

本次评测中，使用 GIZA++[7]进行词对齐处理（包括词聚类部分及其对齐部分），根据前期实验的结果，针对维汉翻译的特定任务，词对齐选用 grow-diag-final-and 的方式，进行维吾尔语-汉语，汉语-维吾尔语的双向对齐。短语（规则）抽取使用 Moses 提供的脚本进行。

3.2 实验环境

（1）硬件:

CPU: Intel(R) Xeon(R) CPU E5-2690 0 @ 2.90GHz, 32 cores

Mem: 128G

HDD: 1.5T

（2）软件:

Operating System:Linux

Version:Ubuntu server 12.04.2

3.3 实验结果及其分析

维汉新闻领域翻译开发集采用评测组织方提供的 700 句（含 4 个参考译文）语料进行调参，参数调整采用 Moses 实现的 MERT[8]方法进行，训练过程使用 Moses 提供的脚本。调参过程中，k-best 大小取 100。实验结果如下表所示（primary 表示主系统（受限系统），contrast 表示对比系统（非受限系统））。

在本次评测中，组织方提供了 1,000 句测试语料。我们分别在 primary（主系统）和 contrast1&2（对比系统）上进行了实验。结果表明，在语料层面，因为两个 contrast 系统使用了较多的训练语料，涉及领域较广，因此翻译质量明显高于 primary；在模型层面，由于我们的机器翻译系统集成了自主研发的语料处理模块，primary 以及两个对比系统 contrast1 和 contrast2 给出的翻译结果均优于传统的翻译模型。对译文进行分析，由于我们的三个翻译系统都是基于词干，因此，与基于词的传统模型相比，翻译结果中未登录词所占比例较小；由于三个系统都基于 Moses 的短语模型，与同样条件下的层次短语模型相比，翻译速度较快。

本次评测的开发集中提供了一个源语言文件和四个参考译文，我们分别使用训练得到的 primary 和 contrast1 两个系统进行了参数调整实验，由于 contrast1 系统使用了较大规模语料，覆盖领域较广，因此调参结果明显好于 primary 和 contrast2。另外，实验发现使用四个参考译文产生的最终调参结果比使用单个参考译文产生的调参结果要好。

4 结论

在 CWMT 2015 的维汉新闻领域评测中，我们搭建了 Moses_PR_Primary, Moses_PR_Contrast1 和 Moses_PR_Contrast2 三个机器翻译系统。经过多组实验，在维汉新闻领域非受限评测中，Moses_PR_Contrast1 系统效果较好。但是，本研究组开展维汉机器翻译时间

较短，前期技术积累较为薄弱，在命名实体的识别和翻译、在句子调序、词对齐等方面没有进行过多的处理，Moses_PR_Primary 系统的效果不太理想。在后续的工作中，将针对评测中出现的问题，如未登录词过多、句子调序等进行深入的研究。

此次参与 CWMT 2015 机器翻译评测，是一次非常难得的机会，能够与国内外同行进行交流和探讨，及时发现我们自身存在的问题，为下一步开展工作带来诸多的帮助。在此向 CWMT 2015 的主办方表示诚挚的感谢，同时也对中国科学院新疆理化技术研究所所有参与本次评测的同事所付出的努力表示感谢！

表 4：测试集实验结果

Systems	Test Set	BLEU_SBP	BLEU	GTM	mWER	mPER	NIST	ICT
primary	1,000	0.4105	0.4262	0.7652	0.4723	0.3504	9.9230	0.4640
contrast1	1,000	0.4530	0.4693	0.7873	0.4396	0.3166	10.5995	0.4978
contrast2	1,000	0.4514	0.4658	0.7841	0.4421	0.3168	10.5193	0.5021

参考文献

[1] 古丽拉 阿东别克, 米吉提 阿布力米提.**维吾尔语词切分方法初探**[J].中文信息学报,2004,18(6):61-65.

[2] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, **Moses: Open Source Toolkit for Statistical Machine Translation**, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

[3] Richard Zens and Hermann Ney. **Efficient Phrase-table Representation for Machine Translation with Applications to Online MT and Speech Translation**, Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Rochester, NY, April 2007.

[4] Marcin Junczys-Dowmunt. **Phrasal Rank-Encoding: Exploiting Phrase Redundancy and Translational Relations for Phrase Table Compression**, Proceedings of the Machine Translation Marathon 2012, The Prague Bulletin of Mathematical Linguistics, vol. 98, pp. 63-74, 2012.

[5] Koehn P, Och F J, Marcu D. **Statistical phrase-based translation**[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 48-54.

[6] A. Stolcke, **SRILM - an extensible language modeling toolkit**, Proceedings of INTERSPEECH. 2002.

[7] Franz Josef Och, Hermann Ney. **A Systematic Comparison of Various Statistical Alignment Models** , *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

[8] Franz Josef Och. **Minimum Error Rate Training for Statistical Machine Translation**. Annual Meeting of the Association for Computational Linguistics (ACL), Japan, Sapporo, July 2003.