第九届全国机器翻译研讨会中科院智能所评测技术报告

杨振新，李淼，朱泽德，陈雷，张建
(中国科学院合肥智能机械研究所，安徽 合肥 230031)

摘要：本文是中科院智能所参加第九届全国机器翻译研讨会（CWMT2013）评测的技术报告。在本次评测中，我们参加了蒙汉日常用语机器翻译和维汉新闻领域机器翻译。本文详细介绍了我们参加的两个评测任务的各系统的理论模型、系统框架、实现方法。

关键词：机器翻译研讨会；评测；技术报告

中图分类号：TP391　　**文献标识码：**A

IIM Evaluation Technical Report for the 9th China Workshop on Machine Translation

YANG Zhenxin, LI Miao, ZHU Zede, CHEN Lei, ZHANG Jian
(Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China)

Abstract: This paper describes an overview of evaluation systems for the 9th China Workshop on Machine Translation (CWMT2013), submitted by the Institute of Intelligent Machines Chinese Academy of Sciences (IIM). We participated in two tasks: Mongolian-Chinese Translation for daily expressions, Uyghur-Chinese Translation for news. This paper introduces the model, the system framework and the implementation method of the two tasks.

Keywords: CWMT2013；evaluation；technical report

1　引言

中国科学院合肥智能机械研究所参加了第九届全国机器翻译研讨会（CWMT2013）评测中的两个项目：蒙汉日常用语机器翻译和维汉新闻领域机器翻译。本文从参评系统、数据处理与使用、实验设置、实验结果与分析等方面详细介绍了我们参加的两个评测任务的各系统的相关情况。

2　参评系统

本次评测中我们使用了两种不同类型的统计机器翻译单一系统，包括基于短语的统计机器翻译模型[1]和基于层次短语的统计机器翻译模型[2]。这里使用的是开源机器翻译平台 Moses[3]。除此之外，我们还对单系统解码出来的 nbest 译文进行词级系统融合，再通过 PageRank 算法，对所有可能的输出路径进行重排序。

2.1 Moses-BP

短语翻译模型是当前非常稳定的一种翻译模型，它的最小翻译单元为短语，即连续的词序列。基于短语的统计机器翻译通过对数线性框架将译文得分描述为若干特征的线性特征的组合。

基金项目：中国科学院信息化专项（XXH12504-1-10）；国家自然科学基金面上项目（61070099）
作者简介：杨振新（1990－），男，博士研究生，研究方向为自然语言处理、统计机器翻译；李淼（1955-），女，研究员，博士生导师，研究方向为自然语言处理与农业知识工程；张建（1954-），男，研究员，博士生导师，研究方向为人工智能与农业知识工程。

对于短语模型的实现，我们使用的是开源工具 Moses，所使用的主要特征包括：

- 正反向短语翻译概率
- 正反向词汇化翻译概率
- 双语言模型
- 词长度惩罚
- 双向 msd 调序模型

2.2 Moses-HP

层次短语模型是基于上下文无关文法的，不使用任何语言学知识的句法模型，该模型的规则定义为：

$$X \rightarrow < \gamma, \alpha, \sim >$$

其中，X 代表非终结符，γ、α 分别代表是由终结符和非终结符组成的源语言字符串和目标语言字符串，~代表的是 γ 和 α 中非终结符间的一一对应关系。

另外，层次短语包含了两条粘着规则：

$$S \rightarrow < S_1 X_2, S_1 X_2 >$$
$$S \rightarrow < X_1, X_1 >$$

规则抽取分为两步：抽取满足对齐一致性的初始短语；将初始短语中的子短语替换为非终结符得到层次短语。

对于层次短语模型的实现，我们使用的是开源工具 Moses，所使用的主要特征包括：

- 正反向规则翻译概率
- 正反向词汇化翻译概率
- 双语言模型
- 词长度惩罚
- 粘着规则惩罚
- 抽取规则惩罚

2.3 CN_PageRank

李文等[4]将 PageRank[5]重排序融入到混淆网络。在单一系统输出的 nbest 基础上，利用词级别的混淆网络技术生成新的翻译后选结果，通过 PageRank 算法，在新的候选翻译结果上进行排序，最终输出系统的翻译结果。系统整体流程如图 1 所示：

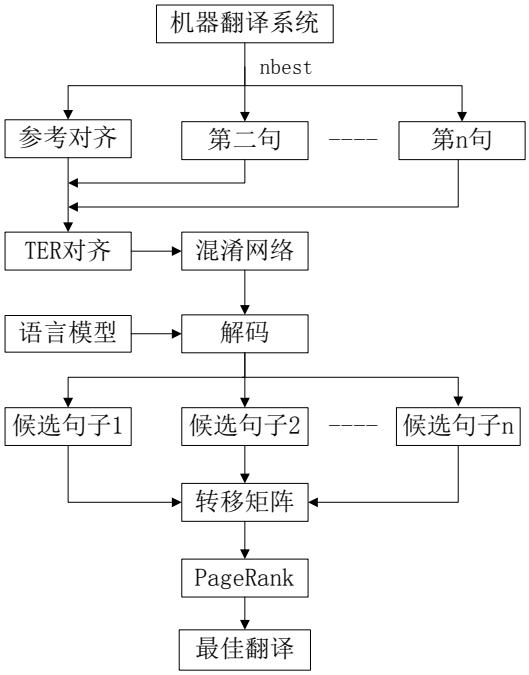

图 1 系统整体流程

本文将 PageRank 排序思想应用到机器翻译的 nbest 后处理技术中，所有的 nbest 视为一个图中的节点，句子 v_i 到句子 v_j 的相似度为对应节点 i 到节点 j 的一个链接或节点 i 给节点 j 的一个投票。句子 i 到句子 j 的投票权重由自身的重要性和相似度共同决定，因而 PageRank 也可以视为一个随机游走的过程。经过足够的时间游走后，节点 i 的 PageRank 值表示图中其它节点到达节点 i 的概率。根据最终的 PageRank 值的排序，选择相应的句子，值越大，句子的重要性也越大。

给定一个图和起始点，游动子以一定的

概率选择移动到图中任意节点。假设 $G(V,E)$ 是一个图，V 是图中的节点，E 是边的集合，则节点 i 到节点 j 的转移概率为：

$$M_{ij} = \frac{A_{ij}}{d(i)} \qquad (1)$$

其中 A 表示图 G 的邻接矩阵，$d(i)$ 是节点 i 的出度。随机游走的矩阵 M 可以描述为如下的形式：

$$M = D^{-1}A \qquad (2)$$

其中 D 是对角矩阵（$\forall i, D_{ii} = d(i)$ and $D_{ij} = 0$, for $i \neq j$）。设 P_s 是初始向量，向量中的第 i 元素的值表示在第 s 步图中其它的节点移动到节点 i 的概率，则第 $s+1$ 步概率可以表示为：

$$P_{s+1} = M^T P_s \qquad (3)$$

初始向量 P_0 的每个元素赋予相同的概率，且所有节点的概率之和等于 1。类似于游动子以相同的概率移动到每个节点，整个随机游动过程可以形式化定义为：

$$P_{s+1} = uM^T P_s + (1-u)P_0 \qquad (4)$$

在等式(4)中，M 是图的转移矩阵，P_0 是一个 $N*1$ 向量，每个元素的初始值为 $1/N$，u 是调节因子，取值范围为 0 到 1，本文取 0.8（经验值）。经过足够长的时间和游动步数后，第 P_s 步和第 P_{s+1} 步的每个元素的概率差值小于 10e-10，此时向量收敛到一个稳定的状态。

给定待排序句子集合 $C = \{S_i\}$，构建一个链接关系图，图中每个节点表示一个候选句子，候选句子 S_i 到 S_j 之间的链接权重可以表示为:

$$P_{ij}(i \to j) = \frac{Sim(s_i \to s_j)}{\sum_k s_i \to s_k} \qquad (5)$$

其中 $Sim(S_i \to S_j)$ 是 S_i 和 S_j 之间的 n 元相似度值，将 nbest 重排序过程视为一个随机游走过程，根据随机游走的形式化定义即可计算句子间的相对排序。句子间的相似度计算，本文选择了 BLEU[6]评测标准，主要是计算句子间的 N 元字面相似度，其计算方法如下：

$$BLEU = BP(\bullet) * \exp(\sum_{n=1}^{N} \frac{\log p_n}{N}) \qquad (6)$$

与机器翻译评测不同的是，本文计算的是单个句子间的 BLEU。

3 实验

3.1 实验环境

本次评测实验环境如下：

CPU：Intel Xeon(R) E5405 2.0GHz*8CPU

内存：23.3G

操作系统：ubuntu-12.04.2-desktop-amd64

3.2 数据处理与使用

我们对中文部分的预处理包括：使用 Stanford-segmenter[7]进行中文分词、A3 区全角转半角；对蒙文的处理包括：转内大拉丁、分离标点符号；对维文的处理包括：使用乌鲁木齐领先科技公司提供的爱革维文编辑器 IlgharPad 进行传统维文到拉丁维文的转换、分离标点符号。

用于训练语言模型的 Sogou 语料为 08 年 SogouCA 新闻语料去除文件名中含有

“sport”的部分，使用 Stanford-segmenter 进行中文分词、A3 区全角转半角及去除 Sogou 语料中重复的部分。

参评系统所使用的语料如下表所示：

表1 参评系统语料使用情况

参评系统	语料使用情况	
	训练集	开发集
蒙汉主系统	76886 句对	4*1000
蒙汉对比系统	76886 句对+22 万蒙汉词典	4*1000
维汉主系统	109895 句对	4*700
维汉对比系统	149146 句对	4*700

其中，蒙汉、维汉主系统所使用语料均为受限语料。蒙汉主系统所用语料是从评测方提供的语料中去除新闻和政府文献语料及蒙汉句对中空行得到的；维汉对比系统语料是主系统语料加上 CWMT2011 中 5 万维汉训练语料中不重复的 39251 条得到的。

3.3 实验设置

词对齐采用两种开源词对齐工具：GIZA++[8]和 Berkeleyaligner[9]。其中 GIZA++对齐采用 grow-dial-final-and[10]的启发式对齐方式获得词对齐结果。Berkeleyaligner 对齐采用配置文件默认设置。将生成的两个词对齐结果合并，在合并的词对齐文件上训练翻译模型。

语言模型训练采用 SRILM[11]工具，并使用 Modified Kneser-Ney[12]进行平滑。采用两种语言模型相结合方法，在对数线性框架下同时使用两种语言模型：以训练集目标端单语为语料训练得到 5 元语言模型；以 Sogou 新闻语料为训练集训练得到 5 元语言模型。

对 CN_PageRank 输出结果合并空格及去除未登录词。

3.4 实验结果与分析

维汉新闻和蒙汉日常用语采用的评分标准均为评测组织方提供的多种自动评价标准，具体包括：BLEU-SBP、BLEU-NIST、GTM、mWER、mPER、ICT、METEOR、TER。除此之外，维汉新闻领域还对主系统的 current 测试集译文进行人工评测，人工评测中每个翻译结果被评价三次，忠实度和流利度的最终评价结果取所有相应打分结果的算术平均。

维汉新闻的 BLEU5-SBP 结果如下表所示：

表2 维汉评测结果

提交系统	BLEU5-SBP	
	progress	current
primary-a	0.4549	0.4733
contrast-b	0.5507	0.4732

其中，系统 primary-a 是评测组织方提供语料经 Moses 层次短语解码，并经过 CN_PageRank 输出的结果；系统 contrast-b 是评测组织方提供语料加上 CWMT2011 维汉语料经 Moses 层次短语解码，并经过 CN_PageRank 输出的结果。

通过主系统和对比系统之间比较，我们发现对比系统 BLEU5-SBP 在 progress 测试集有很大的提高，但是在 current 测试集上却没有提高。我们分析的原因是：由于 CWMT2011 维汉开发集和 CWMT2013 维汉开发集相同，但 CWMT2011 训练集与 progress 测试集相关度要比与 current 测试集相关度高，同时开发集与 CWMT2011 训练集相关度比与 CWMT2013 训练集高，导致

调参后参数更适合 progress 测试集。

维汉人工评测得分如下表所示：

表3 维汉人工评测结果

提交系统	Current		
	忠实度	流利度	总体平均值
primary-a	2.832	3.1227	2.9773

蒙汉日常用语的 BLEU5-SBP 结果如下表所示：

表4 蒙汉评测结果

提交系统	BLEU5-SBP	
	progress	current
primary-a	0.3775	0.1137
contrast-b	0.3752	0.1134
contrast-b*	none	0.1163

其中，系统 primary-a 是评测组织方提供语料经 Moses 短语解码，并经过 CN_PageRank 输出的结果；系统 contrast-b 是评测组织方提供语料加上我们已有的 22 万蒙汉词典经 Moses 短语解码，并经过 CN_PageRank 输出的结果。

通过主系统和对比系统之间比较，我们发现对比系统的结果反而没有主系统的好。分析原因时发现，我们把 contrast-b 系统的参数弄错了，contrast-b*是在线评测平台开放后改正参数的对比系统结果。因为 primary-a、contrast-b 系统的 progress 得分是评测组织方从 500 句 progress 测试译文中选取了与训练集不同的 440 句评测返回给我们的，我们的 contrast-b*在 progress 的得分没有可比性，就没有列出来。contrast-b*在 current test 上的得分相比 primary-a 也没高多少，我们分析的原因是：我们所拥有的词典并非日常用语领域，而且与开发集、测试集的相关度很低。

4 总结

本文介绍了中国科学院合肥智能机械研究所参加 CWMT2013 评测相关情况。我们参加了六项任务中的维汉新闻领域评测和蒙汉日常用语评测两项，取得了较好的成绩。但是我们缺乏命名实体、音字转换及数字时间词的识别与翻译模块。同时，我们对系统融合的研究还不够深入。在本次评测中，蒙汉日常用语领域未登录词较多，如何有效地结合形态信息和其它技术来获取更好的译文也是值得我们研究的问题。下一步我们将对现有系统加以改进。

参考文献

[1] Philipp Koehn, Franz Och, and Daniel Marcu. Statistical phrase-based translation[C]. Proceedings of NAACL, 2003:81-88.

[2] David Chiang. Hierarchical phrase-based translation[J]. Computational Linguistics, 2007, 33(2):201–228.

[3] 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, and Evan Herbst. Moses: Open source toolkit for statistical machine translation[C]. Proceedings of ACL, 2007:177-180.

[4] 李文，李淼，张建，朱海，陈雷. 基于混淆网络和 PageRank 的 Nbest 重排序[C]. 2010 年第三届全国少数民族青年自然语言信息处理、第二届全国多语言知识库建设联合学术研讨会.

[5] S. Brin and L. Page. The anatomy of a largescale hypertextual web search engine [J]. Computer Networks and ISDN Systems, 1998:30(1-7).

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

[7] Pi-Chuan Chang, Michel Galley and Chris Manning. Optimizing Chinese Word Segmentation for Machine Translation Performance[C]. Proceedings of WMT, 2008:224-232.

[8] F. J. Och and H. Ney. A comparison of align-ment models for statistical machine translation[c]. Proceedings of the 18th conference on Computa-tional linguistics, 2002:1086–1090.

[9] P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement[C]. Proceedings of NAACL, 2006:104-111.

[10] Andreas Stolcke. Srilm - an extensible language modeling toolkit[C]. Proceedings of ICSLP, 2002:901-904.

[11] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling[C]. Proceedings of ACL, 1996:310-318.

[12] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models[J]. Computational Linguistics, 2003, 29 (1):19–51.