第十一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，卫林钰，李淼，陈雷，孙凯，陈晟
(中国科学院合肥智能机械研究所，安徽 合肥 230031)

摘要：本文是中国科学院合肥智能机械研究所参加第十一届全国机器翻译研讨会（CWMT2015）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。我们对蒙古语和维语进行形态切分，构造了不同粒度的统计机器翻译系统。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

IIM Evaluation Technical Report for the 11th China Workshop on Machine Translation

YANG Zhenxin, WEI Linyu, LI Miao, CHEN Lei, SUN Kai, CHEN Sheng
(Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China)

Abstract: This paper describes an overview of evaluation systems for the 11th China Workshop on Machine Translation (CWMT2015), 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. To alleviate the data sparsity in our two translation tasks, we segmented Mongolian and Uyghur into morphemes, constructing multi-granularity phrase-based and hierarchical phrase-based SMT systems. This paper introduces the model, the system framework and the implementation method of the two tasks.
Keywords: CWMT2015; evaluation; technical report

1 引言

第十一届全国机器翻译研讨会（CWMT2015）评测中一共有六个子项目：英汉新闻、汉英新闻、蒙汉日常用语、维汉新闻、藏汉政府文献、双盲测试。中国科学院合肥智能机械研究所参加了其中的两个项目：蒙汉日常用语机器翻译和维汉新闻领域机器翻译，并取得了蒙汉日常用语领域第一名、维汉新闻领域第二名的好成绩。本文从参评系统、系统关键模块、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组合。

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
- 短语长度惩罚
- 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

我们[4] [5]将 PageRank[6]重排序融入混淆网络。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候选结果，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。

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。

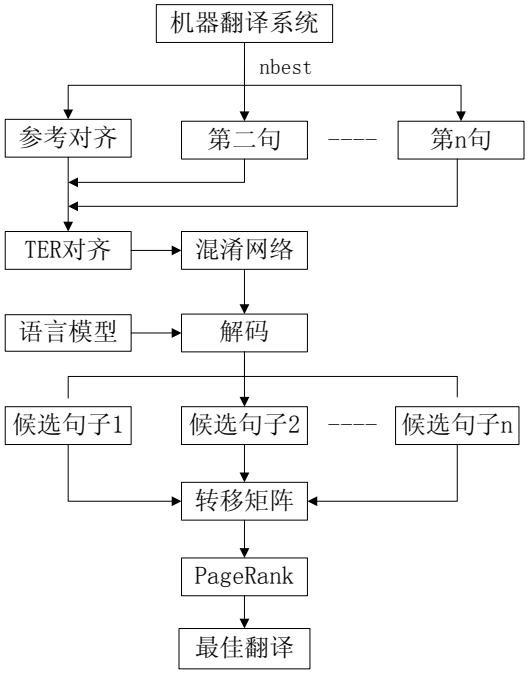

图 1 系统整体流程

3 系统关键模块

3.1 数据预处理

由于蒙古语和维语字形较为独特，我们处理起来较为困难，因此我们将两种语言转换成拉丁形式。我们采用实验室开发的传统蒙文转内大拉丁模块对蒙古语进行传统蒙文到拉丁蒙文的转换，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。

我们对中文部分的预处理包括：使用 NLPIR[1]进行中文分词、A3 区全角转半角、空格处理、对Sogou新闻语料去除重复部分。我们对民语进行标点符号分离和空格处理。另外，在蒙汉语料中存在标点符号不统一问题，即蒙语句子结尾和汉语句子结尾只有一个有标点符号的情况，我们对此进行添加标点符号处理。

3.2 词法分析技术

蒙古语和维语均为形态丰富的黏着语，在有限语料统计机器翻译系统中数据稀疏严重。我们对蒙古语进行联合词性标注的词法分析，并采用基于规则的方法对切分结果进行后处理。由于我们缺乏维语相关资源，

[1] http://ictclas.nlpir.org/

因此我们仅对维语进行词干提取。我们构建词级、词干级、词干词缀级统计机器翻译系统。

3.3 词对齐

我们采用多种词对齐相结合的方法增强翻译模型。词对齐采用两种开源词对齐工具： GIZA++ [2] 和 Berkeleyaligner [3] 。其中GIZA++对齐采用grow-dial-final-and[7]的启发式对齐方式获得词对齐结果。Berkeleyaligner对齐除了采用10轮迭代对齐以外，其余均为默认设置。将生成的词对齐结果合并，在合并的词对齐文件上训练翻译模型。我们使用GIZA++和Berkeleyaligner对词级、词干级、词干词缀级语料进行词对齐的训练。

3.4 语言模型

我们采用两种语言模型相结合的方法，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的5元语言模型；以 Sogou 新闻语料为训练集训练得到的 5 元语言模型。语言模型训练采用SRILM[8] 工具，并使用 Modified Kneser-Ney[9]进行平滑。

3.5 受限的蒙古语命名实体翻译

由于本次评测受限系统均要求仅使用评测组织方提供的数据进行训练，因此，这次评测我们开发了受限的蒙古语命名实体翻译模块。我们从评测组织方提供的蒙汉训练和开发集中抽取包括人名、地名、机构名在内的命名实体，并作为翻译知识融入机器翻译系统中。我们使用Stanford NER[4]识别汉语命名实体，蒙古语命名实体识别则采用联合近似形态切分的蒙古语命名实体识别方法[10]。

3.6 蒙古语数字时间词翻译

蒙古语数字时间词的翻译和识别采用基于规则的方法[11]。主要涉及 7 类数字时间词，包括源语言模板、目标语言模板、调序操作、基本翻译对四个模块。源语言模板、目标语言模板是针对7类数字时间词总结而来，根据源语言模板识别数字时间词，经过调序以及基本翻译对的翻译，最终将其翻译成符合目标语言模板的数词时间词。

4 实验

4.1 实验环境

本次评测实验环境如下：

CPU：Intel Xeon(R) E5-2687W V2 3.40GHz 2*16 核

内存：31.4G

操作系统：ubuntu-12.04.2-desktop-amd64

4.2 实验数据及实验设置

本次评测我们提交的所有系统均为受限系统，只使用了评测组织方提供的语料。由于传统蒙文转拉丁蒙文结果结果存在错误，因此我们对训练语料进行了筛选。经过预处理，蒙汉日常用语的训练集包含 112019 句对，开发集包含 1000 个源语言句子和4*1000 个参考译文；维汉新闻的训练集包含139399 句对，开发集包含 700 个源语言句子和 4*700 个参考译文。

在蒙汉日常用语中，由于部分开发集的源语言和训练集源语言相同，在调参过程中会产生严重的过拟合。因此，在调参阶段，我们采用滤除开发集的训练集作为调参的训练集，调参的语言模型采用滤除开发集的训练集目标端训练。在测试阶段，我们将训练集和开发集合并作为实际训练集，并以此训练语言模型。在调参过程中我们过滤出354 句训练集句子。采用最小错误率算法调参[12]。

在词干级、词干词缀级机器翻译系统中，我们采用 GIZA++和 Berkelryaligner 两种词对齐结果，但在词级的机器翻译系统中我们使用三种词对齐结果：词级 GIZA++、词级 Berkelryaligner、词干级 GIZA++。

我们对最终提交的评测结果进行了合并空格及去除未登录词处理。

4.3 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、

[2] http://code.google.com/p/giza-pp/
[3] http://sourceforge.net/projects/berkeleyaligner/
[4] http://nlp.stanford.edu/software/CRF-NER.shtml

GTM、mWER、mPER、ICT、METEOR、TER。蒙汉、维汉领域的评测结果均采用基于字符的评价方式。

维汉新闻的 BLEU5-SBP 评测结果如下表所示：

表1 维汉评测结果

提交系统	单系统	CN_PageRank
primary-a	0.4064	0.4108
contrast-b	0.3846	0.3918

由于我们对维语了解不够深入，因此，我们仅构造了两个维汉机器翻译系统。其中，系统 primary-a 是词级的层次短语系统，并经过 CN_PageRank 输出的结果；系统 contrast-b 是词干级的层次短语系统，并经过 CN_PageRank 输出的结果。我们给出了单系统及 CN_PageRank 的结果，最终提交的系统均为 CN_PageRank 输出的结果。

通过主系统和对比系统之间比较，我们发现词级的层次短语系统比词干级的效果要好。我们分析的原因是：虽然词干提取可以缓解数据稀疏，但是由于我们在维语词法分析上缺乏相关资源，因此我们的维语词法分析效果不好，导致最终词干级系统没有发挥出优势。

由于蒙古语词的一部分是命名实体，另一部分不是，因此我们在词级翻译系统中没有使用命名实体翻译模块，在词干级和词干词缀级系统中我们根据训练集、开发集中抽取出来的双语命名实体，对测试集中识别出来的命名实体进行匹配并以 XML 的形式融入翻译系统中。

蒙汉日常用语的 BLEU5-SBP 评测结果如下表所示：

表2 蒙汉评测结果

提交系统	单系统	CN_PageRank
primary-a	0.6492	0.6559
contrast-b	0.6335	0.6390
contrast-c	0.6044	0.6091

其中，系统primary-a是词干级的短语系统，并经过CN_PageRank输出的结果；系统contrast-b是词级的短语系统，并经过CN_PageRank输出的结果；系统contrast-c是词干词缀级的短语系统，并经过CN_PageRank输出的结果。我们给出了单系统及CN_PageRank的结果，最终提交的系统均为CN_PageRank输出的结果。

通过表2可以看出蒙汉测试集得分很高，因此我们对测试集进行了分析。我们将训练集、开发集、测试集的源语言去除标点的符号后发现1001句测试集蒙古语中有438句和训练集、开发集是一样的，造成了评测结果很高。

5 总结

本文介绍了中国科学院合肥智能机械研究所参加 CWMT2015 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，取得了蒙汉领域第一、维汉第二的成绩。我们发现将形态信息引入蒙汉、维汉评测是有必要的，多种词对齐结合及双语言模型、数字时间翻译模块等都能有效提高译文质量。但是本次评测仍然存在一些不足之处，比如我们没有结合形态信息对翻译模型进行更为细致的处理；在语言模型上，我们也没有尝试神经网络语言模型的效果；因为我们对维语了解不够深入，因此没有针对拼写错误进行校对等。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].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, 2010, 乌鲁木齐, 132-136.

[5] 杨振新，李淼，朱泽德，陈雷，张建. 第九届全国机器翻译研讨会评测技术报告[C].第九届全国机器翻译研讨会, 2013, 昆明.

[6] S. Brin and L. Page. The anatomy of a

largescale hypertextual web search engine [J]. Computer Networks and ISDN Systems, 1998:30(1-7).

[7] 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.

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

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

[10] 吴金星. 蒙古语语料库加工集成平台的构建 [D] 呼和浩特: 内蒙古大学 2015

[11] Mei Tu, Yu Zhou and Chengqing Zong. 2012. A Universal Approach to Translating Numerical and Time Expressions. Proc. of International Workshop on Spoken Language Translation 2012, 209-216.

作者联系方式：杨振新 安徽省合肥市蜀山区科学岛中国科学院合肥智能机械研究所
xinzyang@mail.ustc.edu.cn