CWMT2009 基于混合策略的汉蒙机器翻译系统介绍

王斯日古楞[1、2] 那顺乌日图[2] 斯琴图[3]
1 内蒙古师范大学计算机与信息工程学院 010022 呼和浩特
2 内蒙古大学蒙古学学院 3 内蒙古师范大学网络中心
Email:siriguleng@imnu.edu.cn

摘要:本文介绍了我们参与 CWMT2009 机器翻译系统评测的基于混合策略的汉蒙机器翻译系统。它是以基于短语的统计机器翻译系统为主，在训练时对于蒙古文进行了部分形态切分，用句法规则进行调序，用模板的方法处理了汉蒙量词翻译问题。本文简单介绍了系统的基本流程及其参与 CWMT2009 的评测情况。
关键字:汉蒙机器翻译系统、混合策略

Description for Chinese-Mongolian Hybrid Machine Translation System of CWMT09

Wang.siriguleng[12] Nasun-urtu[1] siqintu[3]
1. Inner Mongolian Normal University Computer and Information Engineering College ,010022
2. The Institute of Mongolian Studies, Inner Mongolia University
3. Inner Mongolian Normal University network center
Email:siriguleng@imnu.edu.cn

Abstract: *In this paper, we describe our Chinese-Mongolian Hybrid Machine Translation system. The central to this system is phrase-based statistical machine translation system, and it makes some morphological segmentation for Mongolian in training , makes use of syntactic rule to solve the reordering problem in the Chinese-Mongolian, and uses the template approach to solve the Chinese-Mongolian quantifier translation problem. We will introduce the flowchart of system and evaluation report for the CWMT2009.*
Keywords: *Chinese-Mongolian machine translation system, Hybrid Approach*

1 引言

对于汉蒙机器翻译我们曾经做过基于规则的研究和基于实例的研究，随着机器翻译技术的发展，我们正在开展基于混合策略的汉蒙机器翻译研究。在 2009 年中国机器翻译研讨会(CWMT2009)机器翻译评测中我们参加了汉蒙机器翻译系统的评测,在评测项目中只使用了评测组织方提供的资源.下面我们主要介绍基于混合策略的汉蒙机器翻译系统基本框架、实验过程和结果。

2 系统描述

我们开发的基于混合策略的汉蒙机器翻译系统是在基于短语的汉蒙统计机器翻译基础

上，通过对于蒙古文的名词格、名词复数形式和领属格等附加成分的形态切分，基于蒙古语语序的汉语句子调序和汉蒙量词对应翻译等方法构建了一个基于混合策略的汉蒙机器翻译系统。系统流程如图 1 所示.

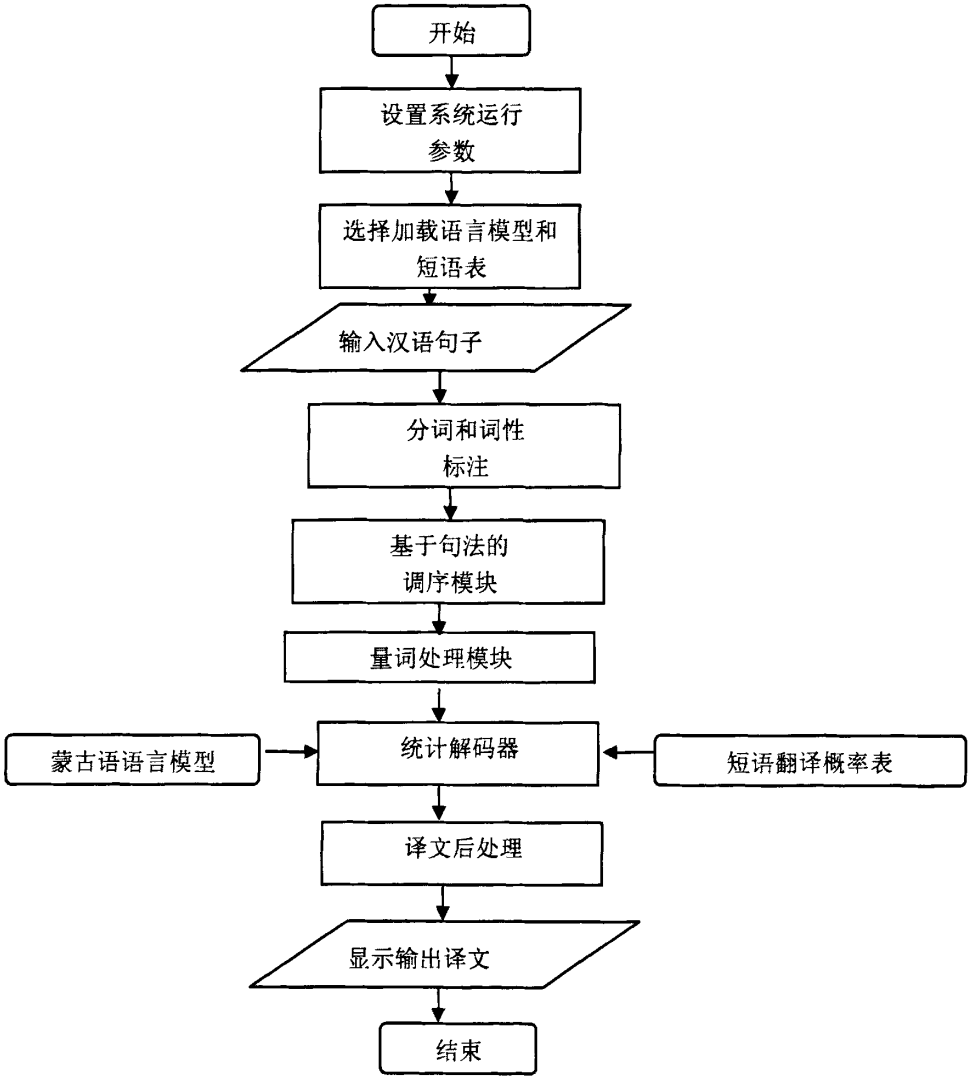

图 1 基于混合策略的汉蒙机器翻译系统流程图

基于混合策略的汉蒙机器翻译系统是一个以基于短语的统计机器翻译系统为主，在训练时对于蒙古文进行了部分形态切分，用句法规则进行调序，用模板的方法处理了汉蒙量词翻译问题。

外部技术说明：用GIZA++进行词语对齐，短语抽取和解码器使用了“丝路”1.0中的抽取模块和CAMEL解码器，用SRILM训练了蒙古语语言模型训练，汉语分析中使用了中科院计算所网上免费资源分词系统ICTCLAS和概率句法分析器ICTPROP。

3 实验

我们使用了CMMT2009机器翻译评测所组织方提供的汉蒙训练语料。需要说明的是系统训练过程中蒙古文语料全部使用了蒙古文内大拉定转写语料，我们编写了蒙古文拉丁转写和蒙古文UTF-8编码的转换程序，最后将拉丁形式的翻译结果通过转写程序转化成传统蒙古文UTF-8格式。

对于蒙古文语料进行形态切分处理后训练了语言模型和短语表。译文后处理时主要完成形态生成和译文中的未登陆词的删除。参数设置直接使用了 CAMEL 解码器中的设置。测试集上的评测结果如表 1.

表 1 评测结果

BLEU4-SBP	BLEU4	NIST5	GTM	mWER	mPER	ICT
0.1432	0.1517	4.8815	0.5185	0.6759	0.5954	0.3838

为了比较结果，我们将测试集中的蒙古文分别以拉丁转写格式和UTF-8进行了实验. 即用拉丁转写格式语料进行训练，翻译后得出的拉丁形式的蒙古文译文，通过转换程序转换成UTF-8格式的译文. 参考答案分别生成拉丁格式和UTF-8格式两种，对于转换前后的译文进行了评测。评测工具直接使用了评测组织方提供的mteval_sbp. 表2给出了实验结果。

表 2 传统蒙古文 UTF-8 编码格式和蒙古文内大拉丁转写格式文本的评测结果表

蒙古文文本格式	NIST	BLEU	BLEU_SBP	GTM	mWER	mPER	ICT
传统蒙古文 UTF-8 编码	4.9041	0.1539	0.1453	0.5209	0.6725	0.5904	0.3862
蒙古文内大拉丁转写	5.1138	0.2356	0.2282	0.5770	0.6314	0.5541	0.3711

从实验结果可以看到蒙古文的拉丁转写结果和蒙古文UTF-8编码评测结果之间存在着很大的差异。其可能的原因:

(1)现有语料中的蒙古文部分还存在一些拼写上的错误。

(2)蒙古文UTF-8文本中存在着大量的控制字符，使得两种类型文本中蒙古文单词的长度不一样。

(3)拉丁到UNICODE转写程序中可能存在漏洞。

今后我们对于汉蒙双语语料进一步校对，不断地完善转换程序，来提高训练语料的质量。

4 总结

本文简单介绍了基于混合策略的汉蒙机器翻译系统及其参与 CWMT2009 的评测情况。我们通过评测学到了很多，也看到很多我们系统中存在的不足。今后，我们将在此基础上，通过对系统中各个环节进行升级和完善，不断地提高系统性能。

参考文献

清格尔泰，蒙古语语法，内蒙古人民出版社，1991年。

俞士汶 等著，现代汉语语法信息词典详解，清华大学出版社，1998年

侯宏旭，刘群，那顺乌日图，基于实例的汉蒙机器翻译，中文信息学报，2007，第4期，P65-72。

刘洋，树到串统计翻译模型研究，中国科学院研究生院2007年博士学位论文。

那顺乌日图，蒙古语语法信息词典的框架设计，内蒙古大学2004年博士论文。

那顺乌日图、刘群、巴达玛放德斯尔，《关于汉蒙机器辅助翻译系统》，《阿尔泰学报》第11号，2001年，汉　城 。

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL 2005, pages 263-270, Ann Arbor, Michigan, June.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In Proceedings of ACL 2001, pages 523-530.

Philipp Koehn. (2004). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas,* pp. 115-124

Philipp Koehn and Hieu Hoang, 2007. Factored Translation Models. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868 - 876, Prague, June 2007.

D. Wu. 1995. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In Proc. of the 14th International Joint Conf. on Artificial Intelligence(IJCAI), pages 1328 - 1334, Montreal, August.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In HLT-NAACL.

http://www.nlp.org.cn，基于短语的统计机器翻译系统“丝路”1.0 版（SilkRoad V1.0）设计与使用说明，2006年10月。