内蒙古师范大学 CWMT2015 蒙汉机器翻译系统评测技术报告

乌丹牧其尔，藏丹，白慧琨，宁静，玉霞，王斯日古楞
（内蒙古师范大学计算机与信息工程学院 呼和浩特 010022）
E-mail：siriguleng@imnu.edu.cn

摘要：本文介绍了内蒙古师范大学计算机与信息工程学院蒙古文信息处理与机器翻译实验室参加 CWMT2015 机器翻译评测中的蒙汉日常用语机器翻译评测项目的情况。为了提高蒙汉机器翻译的性能，我们融入蒙古文的词干、词缀、词性等因子，采用基于短语的因子化模型进行翻译并提交了三组实验的翻译结果，其中实验二是主系统、实验一和实验三分别是对比系统 c 和对比系统 b。本文对参加评测的机器翻译系统和对语料进行的处理步骤进行了详细说明。

关键词：蒙汉机器翻译；因子化模型

Technical Report of Mongolian-Chinese Machine Translation

Abstract: This paper describes our Mongolian-Chinese Machine Translation of daily texts system for the machine translation evaluation task of the CWMT2015.In order to improve the performance of the Mongolian-Chinese machine translation system, we integrated factors like Mongolian stems, suffix, part-of-speech into the Mongolian-Chinese machine translation. In this paper ,we gave a detailed description of the our machine translation system and preparing works to the corpus. The second experiment is the primary system and the first and third one respectively correspond to contrast system c and b.

Keywords: Mongolian-Chinese machine translation; factored translation model

1 引言

CWMT2015 机器翻译评测开展了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、双盲评测机器翻译等六项评测项目，我们参加了其中的蒙汉日常用语机器翻译评测，并提交了一个主系统翻译结果和两个对比系统翻译结果。下面对系统和采用的方法进行详细说明。

2 系统

2.1 翻译模型

目前,基于统计的方法在机器翻译领域占据着主导地位,并出现了多种不同类型的统计机器翻译方法,例如基于词的、基于短语的、基于句法的以及基于混合策略的翻译方法。其中,基于短语的翻译模型仍然是当前统计机器翻译领域的一个研究热点。基于短语的统计机器翻译中统计模型的粒度是词级别的，但从前人的研究经验可知，形态丰富语言的机器翻译中融入形态学信息会在一定程度上解决数据稀疏问题，从而能够提高机器翻译系统的性能。

本文利用基于短语的因子化翻译模型(Factored Translation Model),把蒙古文的词干、词缀、词性等形态信息以因子的形式融入到基于短语的统计机器翻译模型中，进行机器翻译。

2.2 基本系统

本系统的运行环境:操作系统为 ubuntu 12.04 版的 Linux 平台;四核; CPU 均为 Intel(R)Xeon（R）CPU E7-4830@2.13GHz;内存为 32G。

该系统由若干模块组成,分别为语言模型、词语对齐模块、解码器及评测工具。下面分别介绍各个模块。

2.2.1 语言模型

语言模型是采用统计分析和模拟自然语言的结构规律,用统计概率来衡量某句子符合语法的程度。语言模型从大量的训练语料中获取语言结构知识。因此采用涉及领域广泛、规模足够大的语料库作为训练语料,使得训练结果充分反映目标语言句子的特点,这是建立统计语言模型的重要基础。本系统使用了语言模型工具 SRILM,并进行了 4-gram 和 5-gram 语言模型训练。

基金项目：内蒙古自治区自然科学基金项目：“蒙汉机器翻译研究”（2012MS0918），内蒙古师范大学计算机与信息工程学院蒙古文信息处理与机器翻译创新团队建设项目资助。

2.2.2 词语对齐模块

因子化翻译模型的建立与传统的基于短语的翻译模型类似，也在词语对齐的基础上进行,但其不同之处在于根据设计好的实验要求对除了表面词形以外的其他因子如词干、词缀、词性也要进行对齐。本系统使用了开源工具 GIZA++。利用 GIZA++进行蒙语到汉语、汉语到蒙语两个方向的训练,获得双向词语对齐结果。采用 grow-diag-final-and 对双向对齐结果进行优化。

2.2.3 解码器

因子化模型的解码与传统的基于短语的翻译模型类似，同样采用柱搜索算法，不同的是传统的基于短语的模型完成词到词的翻译，而因子化模型的翻译是因子到因子的，且源语言和目标语言的因子数目可以不同，所以因子化翻译模型适合用于蒙古文到汉语的翻译。解码器部分,本系统使用了 Moses 开源解码工具。

2.2.4 评测工具

本系统使用了由 CWMT2015 组织方提供的在线自动评测工具。

3 数据

本系统采用的训练集和开发集均为 CWMT2015 提供的语料,双语语料未使用其他外部资源，本文用到的语言模型的训练语料是对《搜狗全网新闻语料库》进行适当的筛选和预处理后与训练语料的汉语部分合并起来得到的汉语单语语料库。《搜狗全网新闻语料库》是由 CWMT2015 组织方提供的 08 版。

为了完成翻译,对训练语料进行了相应的预处理。预处理分为蒙古文语料预处理和汉语语料预处理两部分。

蒙古文语料的预处理的第一步是利用蒙古文拉丁转写工具将传统蒙古文转写成内大拉丁形式,采用蒙古文拉丁转写形式有运行速度快,处理方便等优点,而且采用传统蒙古文形式进行实验时由于丢失控制符，从而容易出现文字变形的情况，所以我们选择了蒙古文拉丁转写形式。第二步，将蒙古文拉丁转写形式中的单词与标点符号进行分割。第三步，由于语料规模较大，收集和整理过程难免出现一些差错。因此，在双语对齐语料中有几处蒙汉内容不对应、串行的现象，对此进行了更正之后用 Moses 下的 clean-corpus 工具对双语对齐语料进行长度大于 50 的句子和蒙汉对应句中的词数比例过大的部分进行过滤并进一步规范语料格式，对过滤的长句人工进行切分成若干个短句之后加入到训练集中。得到了实验三中使用的训练语料。

实验一和实验二中使用的训练语料是在上述预处理的基础上进行词性标注和切分后得到的，实验一中使用的语料是用基于隐马尔科夫模型的词性标注系统进行词性标注，用于词性标注的单语训练语料是内蒙古大学 100 万词级的《现代蒙古文数据库》和本实验室 80 万词级的语料。形态切分是用基于词缀词典的切分系统对分写附加成分和动词的连写附加成分进行切分。

实验二中的词性标注和切分都是用基于短语的统计机器翻译方法实现的，将蒙古文的词性标注和形态切分问题视为基于短语的统计机器翻译问题，以词性标注为例：训练语料是没有标注词性的蒙古文语料和与其对应的已人工标注词性的蒙古文语料，使用了语言模型训练工具SRILM在已标注词性的蒙古文语料上进行 3 元语言模型训练，使用的实验平台是本文中介绍的进行蒙汉机器翻译的实验平台，所用的语料是 CWMT2011 提供的训练集和我们实验室收集的语料总共约 8 万句。实验二中使用的原语料和派生语料分别是原语料和实验二、实验三中使用的语料。词性标注中的未登录词是用基于隐马尔科夫模型的词性标注系统进行标注，切分中的未登录词用基于词缀词典的切分系统对分写附加成分和动词的连写附加成分进行切分。实验一和实验二中使用的语料格式是：表面词形|词干|词缀|词性。基本语料：“BI AMI-BAR-IYAN YABVBA .”。派生语料：“ BI|BI|NoSuf|R AMI-BAR-IYAN|AMI|-BAR-IYAN|N YABVBA|YABV|+BA|V .|.|NoSuf|W”。

我们通过在部分训练语料上进行测试，发现用基于词缀词典的切分系统切分动词后进行机器翻译的结果优于切分所有符合

切分规则的词之后进行机器翻译的结果，由于这两组实验是在部分训练语料上进行的，其结果除了相互对比之外与其他实验的结果没有可对比性，因此没有给出实验结果。所以对于通过机器翻译方法进行形态切分后出现的未登录词，我们进一步对其中的动词和分写附加成分进行了切分。只切分动词的翻译效果比另一组实验结果好的原因主要有两点。一，采用基于词缀词典的切分系统进行切分的动词与动词词干词典进行比对，进一步确认切分准确性并还原词干没有匹配成功的词，因此动词的切分准确率远远高于其他词类。二，由于动词的形态变化形式最为复杂，翻译结果中的未登录词大多数是动词，对动词以外的其他词进行切分后反而会影响翻译效果。

汉语语料预处理是采用ICTCLAS2015对汉语进行分词和词性标注。对搜狗语料库进行的筛选和预处理步骤具体如下:从原文件中提取题目和正文内容、按句切分、删除一些重复率太多但对语言模型训练贡献不大的内容（如：作者XX)、删除一部分便于识别的乱码、删除连续的空格等一系列规范化格式的操作。最后采用ICTCLAS2015进行了分词。

4 实验

通过上述预处理步骤得到规范的训练语料后进行训练，对CWMT2015组织方提供的蒙古文测试语料库进行的预处理步骤与训练语料相同，不同的是在转换成拉丁形式后人工进行拼写校对。最终对校对和预处理过的1001条蒙古文测试集进行了翻译。对翻译结果进行的后处理包括未登录数词的翻译以及删除未登录词两个步骤。通过基于规则的方法对译文中的未登录蒙古文数词和时间词进行自动翻译，之后删除剩余未登录词，用CWMT2015提供的在线评测工具进行评测。

实验一采用因子化翻译模型利用蒙古文词干、词缀到汉语词，蒙古文词性到汉语词的映射关系进行翻译，采用了4元语言模型。实验一的译文评测结果如表1所示：

实验二在语料预处理方面采用的方法与实验一不同之外，采用的翻译模型和参数都相同。采用因子化翻译模型利用蒙古文词干、词缀到汉语词，蒙古文词性到汉语词的映射关系进行翻译，采用了4元语言模型。实验二的译文评测结果如表二所示：

实验一和实验二的译文评测结果在BLEU5-SBP值上只有0.0031的差距，但是足以证明基于短语的机器翻译模型的词性标注系统和切分系统都以短语为基本分析单位，因此考虑了词的上下文关系，这样能更好的解决兼类词的标注问题。

实验三采用传统的基于短语的翻译模型进行翻译，采用5元语言模型。实验三的译文评测结果如表三所示：

实验二和实验三的翻译结果在BLEU5-SBP值上有0.0074的差距，但是通过分析翻译结果后发现，这两种翻译模型在不同情况下各有各的优点。比如对源语言句子：“YAGAHIGSAN SAYIHAN UJEMJITEI GAJAR YVM E !”和“ENE AJIL-I ARBAN EDUR-TU BARAGAY_A GEJU TOLOBLEJU BAYIN_A .”,传统的短语模型的翻译结果是:“YAGAHIGSAN、美丽的地方啊!”和“这项工作一天BARAGAY_A。”,因子化模型的翻译结果是:“多么美丽的地方,啊!”和“这项工作计划十天之内完成。”,由此可见，因子化模型能够更好地解决训练集数据稀疏的问题。而对源语言句子：“B0D0GSAGAR B0D0GSAGAR SAyI B0D0JV 0LJAI .”和“TAN-DV BAYARLAGAD BARAHV UGEI BAYIN_A .”，传统的短语模型的翻译结果是：“想来想去才想到。”和“对您感谢不尽。”，因子化模型的翻译结果是：“想来想去才想。”和“您感谢不尽。”，由此可见，在没有未登录词的情况下，传统的短语模型在一定程度上要好于因子化模型，因子化模型在翻译过程中考虑多个因子，有时反而会影响翻译效果。

既然这两种翻译模型各有优点，将两种模型融合在一起，使其发挥各自的优点，应该能够得到更好的翻译结果，其中融合两个系统的最简单直接的方法就是通过短语模型翻译后，翻译结果中存在未登录词的句子用因子化模型翻译，用该方法进行了实验

四，译文的评测结果如表四所示。虽然实验结果提升的幅度较小，仍然可以发现合并后的结果比两个模型分别进行翻译的结果要好。

表一　实验一的译文评测结果

BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
0.5673	0.6234	0.5978	10.0207	10.0666	0.7921	0.3303	0.2720	0.7324	0.7046	0.3020

表二　实验二的译文评测结果

BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
0.5704	0.6279	0.6013	10.1126	10.1583	0.7931	0.3289	0.2697	0.7324	0.7083	0.3013

表三　实验三的译文评测结果

BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
0.5778	0.6291	0.6032	10.0884	10.1357	0.7971	0.3151	0.2630	0.7438	0.7117	0.2887

表四　实验四的译文评测结果

BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT	METEOR	TER
0.5803	0.6337	0.6072	10.1731	10.2202	0.7975	0.3154	0.2611	0.7394	0.7141	0.2892

5 总结

本文介绍了内蒙古师范大学蒙古文信息处理与机器翻译实验室参加 CWMT2015 蒙汉日常用语翻译评测的机器翻译系统情况。为了提高蒙汉机器翻译性能，我们融入蒙古文的词干、词缀、词性等因子，采用基于短语的因子化模型进行翻译，并将传统的基于短语的翻译模型和因子化翻译模型的翻译结果相互结合。

在实验过程中，我们发现蒙汉语料的训练集、开发集和测试集的蒙古文拼写上存在一些不一致和拼写错误的现象，所以，我们需要进一步提高蒙汉机器翻译的语料质量。目前，我们的蒙古文词性标注和切分工具还不够成熟，蒙古文语料的加工预处理的准确率不是很高，从而导致机器翻译的性能受到阻碍。所以进一步完善词性标注系统和切分系统是非常有必要的。另外，如何融合传统的基于短语的翻译模型和因子化翻译模型，使其能够更好地发挥各自的优点，进一步提高译文质量也是我们以后的研究工作的重点内容。

参考文献：

玉霞.蒙古文词法分析及其在蒙汉统计机器翻译中的应用[D].呼和浩特：内蒙古师范大学，2015.5

包美荣.融入形态学分析的汉蒙统计机器翻译研究[D].呼和浩特：内蒙古师范大学，2012.4

作者联系方式：

王斯日古楞，内蒙古呼和浩特市赛罕区昭乌达路 81 号，010022，13674889799，siriguleng@imnu.edu.cn

乌丹牧其尔，内蒙古呼和浩特市赛罕区昭乌达路 81 号，010022，15248083181，wuda_000@163.com