CWMT2009 统计机器翻译研讨会

内蒙古大学技术报告

侯宏旭，宋美娜，姜鑫，辛强，明玉
内蒙古大学计算机学院，内蒙古呼和浩特 010021

摘 要：本文主要介绍了内蒙古大学参评系统参加CWMT2009研讨会的技术报告，我们参加的项目评测任务是汉蒙日常用语统计机器翻译任务。这里主要介绍了该系统的主要框架、模型、实现细节及其评测结果。
关键字：统计机器翻译；调序模型；trigger对；语言模型

Technical Report of Inner Mongolia University on the Statistical Machine Translation Evaluation Task of CWMT2009

Hou Hongxu, Song Meina, Jiang Xin, Xin Qiang, Ming Yu
College of Computer Science, Inner Mongolia University, Hohhot 010021,China

Abstract: *This paper describes our statistical machine translation system used in the evaluation campaign of CWMT'09. In this year's evaluation, we participated in one task: Chinese-to-Mongolian translation. Here, we mainly introduce the overview of our system, the primary modules, the key techniques, and the evaluation results.*
Keywords: *statistical machine translation; reordering model; trigger pair; language model*

1 引言

2009年全国统计机器翻译研讨会(CWMT2009)机器翻译评测一共包括五个评测任务，即新闻的汉英和英汉翻译任务、科技的英汉翻译任务、新闻的汉英融合任务及日常用语的汉蒙机器翻译。内蒙古大学作为参加单位之一参加了其中的汉蒙机器翻译评测任务，这里主要介绍该系统的主要技术内容和相关评测参数。

2 参评系统描述

在这次机器翻译评测中，我们参加评测的系统是开源基于短语的汉蒙机器翻译系统（Moses），现在就详细介绍一下该系统的整体设计及各个模块的实现原理。

2.1 系统流程

该系统包含以下四个主要部分：短语翻译模型的训练、语言模型的训练、解码、翻译结果的评价。在模型训练过程中，除了利用（Moses）的默认设置来进行训练解码外，还加入了我们自己开发的调序模型以及Trigger对长距离的蒙语语言模型作为特征，从而可以提高机器翻译的质量。

2.1.1 短语翻译模型的训练

通过短语翻译模型的训练，从汉蒙句子对齐的语料库中学习到汉语短语到蒙语短语的翻译概率表，流程图参见图1.

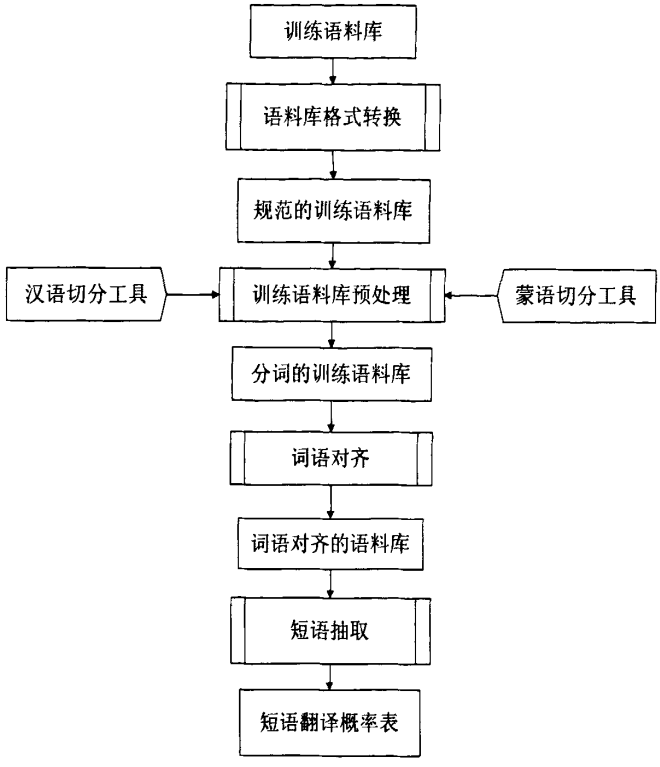


图 1：短语翻译模型训练的流程

在统计机器翻译翻译中，由于调序扮演着非常关键的作用。现有的基于短语的统计机器翻译方法中，通常采用的是 IBM 调序模型。这种模型比较适合于词序变化不大的语言之间的翻译。而汉语和蒙古语词序的差异是非常大的。汉语是 SVO 型的语言，宾语总是出现在谓语后面，而蒙语是 SOV 型的语言，谓语动词总是出现在句子尾部。因此在汉蒙翻译中采用 IBM 调序模型或者 HMM 调序模型都会面临比较严重的问题。为此，我们采用了一种基于词序变化概率分布的调序模型。因为目前还没有出现比较完善的蒙古语句法分析器，而且句法上调序的开销比较高。因此，这种模型并没有考虑句法信息，只是涉及到翻译中词序变化的概率分布，我们通过以下公式，利用目标短语相对位置与源短语相对位置的位置差来描述短语的这种调序关系。

$$P\left(\frac{j}{Len\ (F)} - \frac{i}{Len\ (E)} \mid e, f\right)$$

给出的概率是短语 e 在 E 中出现的位置与 f 在源语言句子中出现位置的相对差，在解码时，我们需要对目标句子的可能长度进行估计。通过实验我们可以知道，句子的长度比的分布基本满足正态分布曲线，也就是说这个长度比 len(e)/len(f)是满足正态分布的，并在解码过程中去修正这个估算的目标句子长度。

通过一些分析，我们发现对于汉语词和它的蒙古语翻译的距离存在着比较明显的正态分布关系。因此我们利用这一正态分布曲线来拟合这一分布，并得到下面的概率计算公式：

$$p\ (\ x\ ) = \frac{1}{\sqrt{2\ \pi\sigma}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

$$P(i \mid f, e, j) = \int_{-0.5}^{+0.5} p\left(\frac{i}{Len\ (E)} - \frac{j}{Len\ (F)}\right)$$

在这个公式中，我们得到的是当源语言短语 f 中的第一个词的位置为 j，目标短语 e 的第一个词应该出现的位置。在解码时将这个概率作为一个特征。

2.1.2 语言模型的训练

统计语言模型在机器翻译、文字处理、文字检索等领域有着广阔的应用。作为机器翻译的一项基础性工作，蒙古语语言模型的建立不能照搬汉语、英语等语言中使用的语言模型方法。所以针对蒙古语语言的特点我们采用了一种新的蒙古文统计语言模型，即基于 trigger 对的长距离蒙古语语言模型，并应用于汉-蒙机器翻译系统。流程图如图 2 所示：

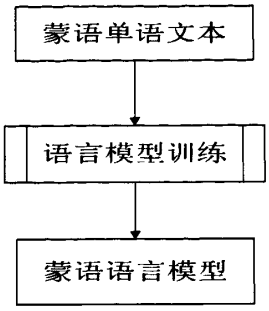


图 2：语言模型训练的流程

因为蒙古语属于黏着型语言，蒙古语的构词，构形都是通过在词干后缀接不同的词尾而实现的，并且它们还可以层层缀接，这使得蒙古语词法形态变化丰富且复杂。为了能够反映出更长距离的相关信息，利用 trigger 对来描述更长距离上的关联信息。

如果词 A 的出现使得后文中词 B 出现，则称（A→B）为一个 trigger 对。其中 A 称作触发者（trigger），B 称作被触发者（trigger word）。在自然语言中，这种情况是非常普遍的，也就是通常所说的词的习惯搭配现象。从 trigger 对的定义及选取标准可以看到，trigger 对能够表达长距离的词之间的相关程度，而这种对距离限制很少的词之间的搭配是非常符合人们的语言习惯的，这恰好弥补了传统 N 元文法语言模型描述距离小于 N 的缺点。因此如能恰当地将基于 trigger 的语言模型与 N 元文法语言模型结合起来必将有助于更好地描述语言的统计特性，进而提高机器翻译系统的性能。

而在构建一个基于 trigger 对的语言模型时就需要选择一个合适的度量标准并据此保留所需数目的 trigger 对。

一个最简单的控制 trigger 对数目的方法就是给历史加窗，即限制 trigger 对的最长约束距离。一般说来，这个参数并无精确要求，可以根据经验在合适范围内选取。大量文献认为：在历史中最近的六个词已包含了绝大部分信息。在选择 trigger 对的实验中，我们选取的窗长限制为 9，即只考虑当前词的前 9 个词作为历史。

trigger 对选取的距离由于 trigram 的存在，最短距离从 4 开始，最长距离用 9。

基于 trigger 对的长距离蒙古语语言模型认为第 i 个符号的出现是由于第 i 个符号做为被触发者所构成的 trigger 对来决定的，而一个句子第 i 个符号做为被触发者可能会与前面多个符号构成 trigger 对，这时则认为强度由其中最强的来决定。

基于 trigger 对的长距离蒙古语语言模型的得分标准采用如下方法：

$$\log\ P_{trigger}\ (w_1 w_2 \ldots w_n) = \sum_{i=1}^{n} \log\ MAX\ [P(w_{i+3} \mid w_i),\ldots,\ P(w_{i+L} \mid w_i)]$$

其中 $\log P_{trigger}(w_1 w_2 \ldots w_n)$ 表示的是句子的 trigger 对得分；$P(w_{i+3} \mid w_i)$ 表示 w_i 出现条件下 w_{i+3} 出现的条件概率；L 表示 trigger 对窗口的最大距离，设置为 9。

因此，在解码时将 trigger 对的长距离蒙古语语言模型的得分作为另外一个特征。

2.1.3 解码

在汉蒙机器翻译系统中，我们除采用 Moses 里面默认设置的特征外，还加入了自己的调序模型以及 Trigger 对语言模型的特征，并利用这些特征计算出总的翻译概率。

在解码时，我们求解候选翻译结果的期望值，并选取概率最高的作为最终的翻译

$$P(e|f)=\frac{\exp\left[\sum_{m=1}^{m}\lambda_m h_m(f,e)\right]}{\sum\exp\left[\sum_{m=1}^{m}\lambda_m h_m(f,e)\right]}$$

这里，我们用这些概率的对数形式作为特征。这些参数可以人工指定，也可以通过训练得到。

在我们的系统中，我们采用了基于 BLEU 值的最小错误率训练。训练集在和测试集类型都是来源于 CWMT09 发布的训练语料和测试语料，并利用最小错误率算法对参数进行训练。

2.1.4 翻译结果的评价

在这次评测过程中，我们采用 CWMT09 发布的评测工具（mteval_sbp）对测试集进行 NIST-BLEU 打分。

2.2 系统性能

在这次评测中，机器翻译评测采用计算机配置如下表所示：

CPU	内存	操作系统
P4 2.0GHz	1G	Linux 6.0 以上平台上

3 实验

汉蒙机器翻译用到了 Moses 系统。

3.1 数据准备

我们的训练数据主要来源于CWMT09发布的训练语料，其中CWMT09发布的训练语料有67288句对。对于翻译模型的训练，我们并非用所有的全部训练语料，而是过滤了一部分语料来生成我们最终的训练语料。对于语言模型的训练，我们采用CWMT09发布的62400个句子蒙语语料用来训练。

表1：训练语料资源列表

语种	领域	规模	说明
汉语-蒙古语	政府文献和法律法规 日常对话、文学	CWMT2009发布的训练语料 总共67288句子对	UTF-8编码
		对CWMT2009发布的训练语料进行处理最终用于翻译模型的训练语料规模为67186句子对	UTF-8编码
蒙古语	教材、政治、 文学、新闻	CWMT2009发布的单语语料包括100万词总共62400个蒙语句子	UTF-8编码

获得上述所有的训练语料之后，我们对中文和蒙文分别做了如下的处理：对中文数据进行的处理有：中文的分词和全角变半角：对蒙文数据进行的处理为：标点符号的分离、蒙文空格以及蒙文元音间隔符的处理。

其中中文的分词是利用开源工具ICTCLAS。

3.2 短语表的获取

所参评的汉蒙机器翻译项目的短语表是利用工具包（Moses）进行训练获取的，其中参数是利用Moses工具包的默认设置。但是在我们所用到的调序模型中，对Moses所生成的短语

表又进行了处理，此外，在我们所用到的Trigger对语言模型中，我们又对62399句对蒙语语料进行处理，生成Trigger对语言模型文件。

3.3 开发集的获取

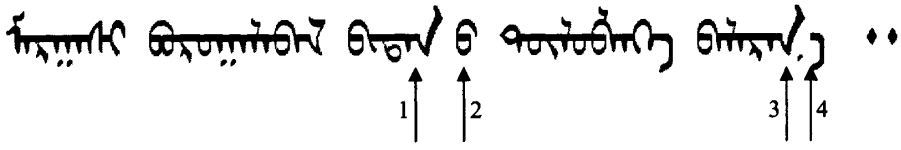
在这次机器翻译评测任务中，我们的开发集直接来源于 CWMT09 发布的开发集。其中汉蒙评测任务的开发集的规模如表 2 所示：

表 2 汉蒙机器翻译评测任务的开发集规模

评测任务	汉蒙机器翻译
开发集规模	400 个汉语句子，其中每个汉语句子有 4 个参考答案

3.4 对测试语料的特殊处理

在这次评测过程中，由于最初对蒙古文语料进行了错误的处理，导致评测结果很低。尤其是没有对蒙文空格以及蒙文元音间隔符进行处理。例如



在蒙文字编码中，一个蒙文字和它的附加成分之间是用窄宽度无间断空格 NARROW NO-BREAK SPACE(U 202F)来间隔的，该间隔的 UTF8 编码为 202F。如图所示箭头 1 所指向的字和箭头 2 所指的附加成分之间的空隙为窄宽度无间断空格。

另外一种情况是一个蒙文字最后一个字节（多为辅音+元音的情况）分开来写，此时该蒙文字中间的间隔为蒙文元音间隔符。如图中箭头 3 和 4 中间的空隙为蒙古文元音间隔符（180E）。由于最初对蒙古文语料处理上的错误，把蒙文的窄宽度无间断空格与元音间隔符按普通的空格（2000H）进行了处理，导致了评测结果很低。

在此基础上，我们对汉语语料进行了分词以及全角变半角，对蒙文语料进行了标点符号的分离、蒙文空格以及蒙文元音间隔符的处理。最后，我们又对汉语合并空格，对蒙文进行了标点符号的合并。

3.5 实验结果

由于我们先前对蒙文语料处理上的错误，导致评测结果非常低，我们在对蒙古文语料重新进行处理后，又分别从测试集4000句摘出一部分语料进行了测试，下面的表格分别是人工的评测结果以及经过我们改进后的测试结果。但是由于时间上的原因，我们只利用了测试集中的一部分，相对来说测试集的规模非常小，NIST得分上可能没有明显的提高。表4中系统primary-systema是加入调序模型以及Trigger对语言模型之后的汉蒙机器翻译基本系统，系统contrast-systemb是在Moses默认设置下的对比系统，其中测试集上的打分我们利用了大小写敏感进行打分。

表3 汉蒙日常用语人工评测结果

participants	Average Score of Adequacy	Average Score of Fluency
imuc	2.32	2.41

表 4 汉蒙日常用语系统经过改进后的测试结果

systems	BLEU4	NIST5
zh_mn_dail_trans-imuc-primary-systema.result	0.1719	3.1469
zh_mn_dail_trans-imuc-contrast-systemb.result	0.1533	2.6922

4 结论

本文主要介绍了内蒙古大学参加第五届统计机器翻译评测活动的机器翻译系统的实现情况和测试结果。在这次评测活动中，我们在开源的基于短语的统计机器翻译系统（Moses）的基础上实现了汉蒙机器翻译。

目前，我们参评的系统所采用的调序模型非常简单，模型参数很少，但效果好于传统的IBM模型，比较适合于汉语和蒙古语这类语序差别较大的语言之间的翻译。但是，由于时间限制，我们还有其它自主开发的模块没能用到这次评测活动当中，而且，在这次活动中，我们对蒙古文语料处理上出现了失误，以致评测结果出乎我们的意料。在评测结果出来后，我们分析原因，发现在对蒙古文语料处理上没有对蒙文空格以及蒙文元音间隔符进行处理，在此基础上，我们对汉语语料进行了分词以及全角变半角，对蒙文语料进行了标点符号的分离、蒙文空格以及蒙文元音间隔符的处理。从而经过我们改进后，测试结果明显得到了提高。但是仅针对我们参评的系统，还是存在很多问题有待解决。例如在翻译模型中并没有利用词干、词缀等语言学信息。如何通过将翻译的基本单位从词改变到词干词缀，从而可以得到更深层次的语言信息，也就有可能能够进一步提高机器翻译的质量。利用这些信息，以及句法信息是我们下一步需要进行深入研究的内容。

此外，参评系统所采用的trigger对的长距离蒙古语语言模型可以很好的解决N-gram在长距离上的信息缺乏问题。将基于trigger对的长距离蒙古语语言模型与传统的3-gram语言模型进行结合，非常有效的提高了机器翻译的准确性。但是随着语言模型研究领域的不断扩充和深入，trigger对的长距离蒙古语语言模型还存在一些问题，比如trigger对距离范围的变化是否对语言模型的性能产生影响，还有同调序模型一样，语言模型仅利用了蒙语表面词形这一方面信息，并没有利用词干、词缀信息，这些问题在今后的工作中将做进一步研究，从而使蒙古语语言模型更加完善。

总之，希望通过这次评测，能够跟其它的研究机构和参评单位进行一次很好的沟通，努力学习其它参评系统的特长，总结经验，从而能够取长补短，进一步改进和完善我们目前的系统。

参考文献

侯宏旭，刘群，李锦涛. 一种基于短语的汉蒙统计机器翻译与调序模型，高技术通讯 CHINESE HIGH TECHNOLOGY LETTERS 2009年 第05期

确精扎布. 蒙古文编码. 内蒙古大学出版社，2000.

苏韬，孙甲松，王作英. 基于Trigger的长距离语言模型[J]，计算机工程与应用，2002，(18)：59～61

王小捷，常宝宝. 自然语言处理技术基础[M]，北京:北京邮电大学出版社. 2002.

杨攀，张建，李淼，乌达巴拉，雪艳，汉蒙统计机器翻译中的形态学方法研究，中文信息学报 2009 年01期

伊·达瓦；张玉洁；上园一知等，蒙古语语言-文字的自动化处理，中文信息学报，2006，20（4）：68-74.

Andreas Stolcke. SRILM – an extensible language modeling toolkit[C]. In Proceedings of International Conference on Spoken Language Processing, 2002，2：901～904.

Moses – a factored phrase-based beam-search decoder for machine translation. 13 April 2007, URL: http://www.statmt.org/moses/ .

Peter F. Brown, Stephan A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter Estimation. Computational Linguistics, 19(2):263-311.

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. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, demonstation session.*