高技术通讯 2009 年 第 19 卷 第 5 期:475 ~ 479

doi:10.3772/j.issn.1002-0470.2009.05.007

一种基于短语的汉蒙统计机器翻译与调序模型①

侯宏旭②* ** *** 刘 群** 李锦涛**

(* 内蒙古大学计算机学院 呼和浩特 010021)
(** 中国科学院计算技术研究所 北京 100190)
(*** 中国科学院研究生院 北京 100190)

摘 要 根据蒙古语的一些特点,为基于短语的汉蒙统计机器翻译提出了一种适合于汉蒙统计机器翻译的调序模型,并给出了相应的训练和解码算法以及初步实验的结果。汉蒙双语语料库规模很小,数据稀疏问题严重,而在汉蒙翻译中,词序变化又非常明显,在汉英等机器翻译中使用的调序方法难于应用到汉蒙统计机器翻译中。通过对汉蒙翻译过程中词语顺序变化的正态分布假设,建立了一种概率调序模型。实验表明,这种概率调序模型好于 Moses 系统中采用的调序方法。

关键词 机器翻译,统计方法,蒙古语,调序,概率

0 引 言

调序模型是统计机器翻译研究的一个重点,近几年取得了一些重要的成果[1,2]。但是,对于汉蒙统计机器翻译,调序模型并没有专门进行研究,目前可知的系统都是采用汉英等其它语种的统计机器翻译中使用的调序模型,这些模型并不适应于汉蒙机器翻译。虽然对于语序上比较接近的语言,例如汉语和英语,采用短语翻译可以解决大多数的词序调整,但是调序仍然是不可避免的,对汉英机器翻译来说这往往体现在形容词短语或者副词短语上。对于汉语和蒙古语之间的翻译,这样的情况则更为复杂。蒙古语的基本语序是主-宾-谓结构,典型的蒙古语句子的结构中谓语动词处在句子的最后,而汉语则是主-谓-宾结构[3]。这样长距离的调序在汉蒙机器翻译中是非常常见的。因此,我们必须在汉蒙统计机器翻译中考虑调序模型。

目前应用在汉英等语种的统计机器翻译的调序模型主要有 IBM 的调序模型[4,5]、利用句法信息的调序模型[6]、最大熵调序模型[7]等。IBM 的调序模型是一种非词汇化的调序模型,这种方法的效率比较低,因此通常需要限定调序的范围[8]。这使得这种方法不能处理长距离的调序,而处理长距离的调序,正是汉蒙机器翻译里面必需的。而后两种模型都引入了显式或者隐式的句法结构信息,因此可以达到较好的效果。由于蒙古语语法分析的不成熟和语料库的不足,使得这两种方法很难在汉蒙统计机器翻译中得到应用。在日语相关的翻译中很多方法也大都是建立在句法或者依存树分析的结果上的。

本文试图利用蒙古语的特点,研究一种基于短语的统计机器翻译中的调序模型,希望在不考虑引入句法结构信息的情况下,通过词语顺序变化的概率分布特点建立调序模型,提高汉蒙翻译中词语调序的准确率。

1 基于词序变化概率分布的调序模型

由于汉语和蒙古语的句法差异是比较大的,因而调序模型的好坏对机器翻译质量的影响非常大。由于蒙古语自然语言处理的研究还处在初级阶段,相关的研究还在进行中,目前还没有出现比较完善的蒙古语句法分析器,而且句法上的调序的开销比较高。因此,在本文的研究中我们没有考虑引入句法信息,而是对翻译中词序变化的概率分布进行了研究。

现有的基于短语的统计机器翻译方法中,通常采用 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。汉语是主-谓-宾(SVO)型的语言,

① 973 计划(2007CB316503)和内蒙古自然基金(200607010805)资助项目。
② 男,1972 年生,博士,副教授;研究方向:自然语言处理;联系人,E-mail: cshhx@imu.edu.cn
(收稿日期:2008-03-31)

宾语总是出现在谓语后面，而蒙古语是主-宾-谓(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 调序模型或者隐马尔科夫模型(hidden Markov model，HMM)调序模型都会面临比较严重的问题。

为此，我们提出了一种基于词序变化概率分布的调序模型。在这种模型里面，我们利用汉蒙词序变化的概率分布来描述调序的长度和好坏。

我们试验了一些方法，例如基于目标短语的绝对位置的方法、基于目标短语在目标句子中的绝对位置的方法、基于目标短语在目标句子中的相对位置的方法、基于目标短语与源短语的位置差的方法、基于目标短语相对位置与源短语相对位置的位置差的方法，其表示式如下：

(1) 绝对位置：$P(i \mid e,f)$

(2) 相对位置：$P(\frac{i}{\text{len}(E)} \mid e,f)$

(3)短语绝对位置差异：$P(j-i \mid e,f)$

(4)短语相对位置差异：$P(\frac{j}{\text{len}(F)}-\frac{i}{\text{len}(E)} \mid e,f)$

上面公式中的 e 为源语言短语，f 为目标语言短语，j 为源短语在源句子中的位置，i 为目标短语在目标句子中的位置。文献[9]是针对日英翻译的，它就是采用了方法(3)。

由于句子的长度是不定的，所以通过绝对位置是不能很好地表达短语的偏移的。单纯依靠目标短语的位置又不能良好地表达源短语和目标短语之间的相对关系。所以我们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调序关系。

问题是，由于在解码时目标句子的长度并不知道，所以我们只能估计出目标句子的可能长度。因此，我们对 12000 个汉蒙对照的平行语料进行了统计。通过实验我们知道，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，并在解码过程中修正这个估算的目标句子长度。

如果以短语相对位置作为调序的依据的话，那么用一个什么样的模型来描述这个位置关系呢？和句子长度的估计所采用的方法一样，我们也希望通过一个简单的概率模型去描述相对位置的分布。

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。而相对来说汉语和英语则关系不是特别明显。以名词为例，对汉语和英语来说，并不会因为这个词是做主语还是宾语而改变词的形式。例如，“我”在做主语和宾语时形式是不会改变的。同样，英语中的“student”也不会因为它的角色改变而改变词形。但是，蒙古语中的词却会因为角色的变化而产生词形的变化，通常是添加某些不同的词缀。我们正好利用蒙古语的这一特点，希望发现不同的汉蒙短语对的分布接近一个比较容易描述的概率分布。

通过对较高频率的短语调序分布规律进行分析，我们发现它们都能较好地符合正态分布的规律。我们不能给出更详细的分析结果，下面给出的是两个例子。

以短语对“我”-“BI”为例，在实验语料中共出现 1689 次，相对距离在 $-1 \sim +1$ 之间，每间隔 0.05 计数(即横坐标为 0～40)，绘制的曲线图如图 1。

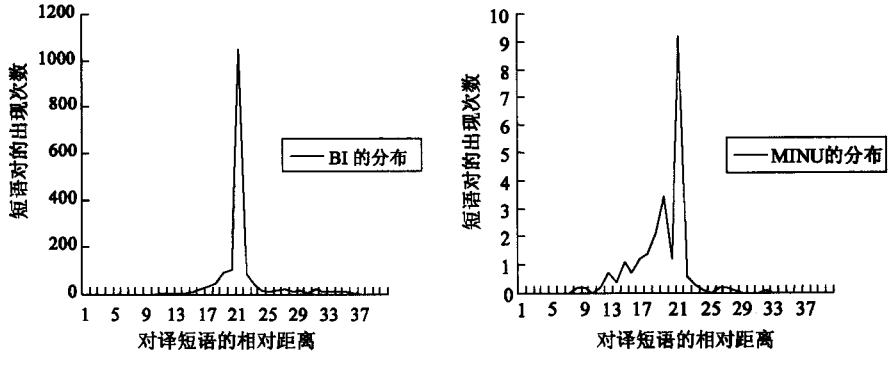

图 1 典型的相对距离分布

图 1 中横坐标为对译短语的相对距离(为了清晰起见，实际是相对距离乘以 40 再加 20)。从这个图可以看到，“我”和“BI”对译时，它们在源语言句子和目标语言句子中出现的平均相对距离约为

0.0238,两个词基本上都出现在相同的相对位置上。又例如,“我”和“BIDE”对译时,它们的平均相对距离约为 -0.1075,也就是说,“BIDE”在目标句子中的相对位置会前移 11% 左右的距离。这样我们可以确定大多数情况下目标短语的位置。从图中看出,我们可以用一个正态分布曲线来近似描述这个分布。

第二个短语对(“我”,“MINU”)的情况要比“我”和“BI”略为复杂一些。在这个短语对的分布中不再是平均值附近出现一个高峰,而是出现了两个高峰。但是,尽管如此,仍然呈现出一定的正态分布趋势。

因此我们利用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:

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

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

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 i 的概率,这个概率作为机器翻译模型的一个特征。

2 参数训练和解码

在汉蒙机器翻译系统中,我们采用了若干个不同的特征,利用这些特征计算出总的翻译概率[10,11]。这些特征如下:

(1) 短语翻译概率 P_T;

(2) 短语调序概率 P_D;

(3) 语言模型概率 P_{lm};

(4) 句长概率 P_l。

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[12,13]。翻译概率的计算公式如下:

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

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。训练在和测试集类型相近的开发集上进行。首先,我们需要生成开发集翻译结果的 N-best 结果。然后利用文献[14]中的最小错误率算法对参数进行了训练。

3 实验结果

由于我们目前可能获得的汉蒙双语语料比较少,所以我们只能够利用规模不太大的语料完成实验。后面的所有实验都是在一个 19000 个句对的汉蒙双语语料库上进行的。这个语料库中的语料为口语语料。语料是已经进行过词语对齐的。

由于汉蒙机器翻译的研究还处在一个初级阶段,目前还没有其它的统计汉蒙机器翻译系统。因此,我们只能自己设计开发集和测试集。

在实验中,我们从其它口语语料中随机抽取了 400 个汉语句子。然后由以蒙古语为母语的人为这 400 个句子提供了 4 个参考答案(图 2)。

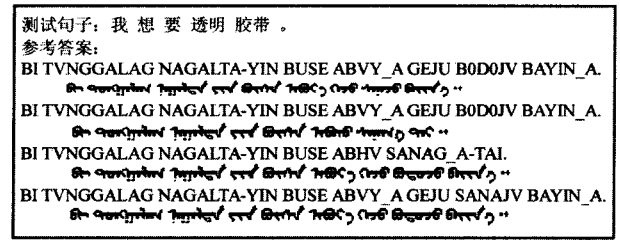

图 2 汉蒙机器翻译参考答案

评价工具采用的是 NIST 评测工具 mteval-v11.pl。给出的评价指标是 BLEU(4 元)[15]。

我们针对不同的调序模型进行了实验(表 1)。

表 1 调序模型的比较实验

调序方法	BLEU
不调序	0.2011
IBM 调序模型	0.2017
绝对位置	0.2155
相对位置	0.2036
绝对距离	0.2151
相对距离	0.2373

在调序模型的对比实验中,我们可以看到相对于不调序的翻译模型,增加了调序以后得分会有一定的提高。其中对于 IBM 的调序模型和相对距离的调序模型都能得到比较好的性能提升。但是实验中 IBM 的调序模型由于搜索空间比较大,因此翻译速度要比相对距离方法速度慢很多。

另一个对比实验中我们和小组开发的另一个汉英统计短语翻译系统 Confucius 以及 Moses 进行了对比(表 2)。

表2 调序模型的比较实验2

对比系统	BLEU
对比系统1：不调序	0.2011
对比系统2：IBM调序	0.2017
对比系统3：Confucius	0.2315
对比系统4：Moses	0.2354
采用概率化的调序	0.2373
Confucius+概率化调序模型	0.2443

对比系统1完全没有进行调序。

对比系统2采用了非概率的调序算法，在这个算法中，调序被限定在以源短语为中心的范围之内，短语出现在这个范围内的概率是处处相同的。由于汉英翻译时这样短语级的调序变化比较少，而且距离一般也比较小，所以这个方法对于汉英等机器翻译的效果是比较好的，能够带来性能比较大的提升，而这些提升其实是由语言模型提供的。但是，对于汉蒙机器翻译来说，由于语序的变化是比较大的，所以这样的算法能够解决的顺序问题是非常有限的，只有通过增大限定范围的方法来提高效果，但是这样又会使效率急剧下降。从实验结果上来看，利用这种方法的BLEU并没有明显的提高。

对比系统3是采用了目前我们在汉英机器翻译中使用的一个系统，现对于汉蒙统计机器翻译系统来说，这个系统采用了更多的特征，系统也更加完备，所以得分相对来说比较高。而汉蒙系统仅采用了短语概率，语言模型，调序模型和句长模型4个特征。作为后续的工作，我们会将更多的特征加入到汉蒙机器翻译系统中去。

对比系统4是一个开源的系统，它的调序模型采用了一个固定的线性调序距离的惩罚。

而对我们提出的概率化调序模型来说，由于采用了正态分布的模型描述，给出了调序的中心长度，也就是说给出了一个短语可能被最大概率调到的位置，这样，我们就更可能得到最好的结果。从实验结果上来看，相对于IBM模型的方法，BLEU成绩有大约3个百分点的提高，这得益于高元的n-gram匹配次数的增加。

4 结 论

目前汉蒙机器翻译的研究，尤其是汉蒙统计机器翻译的研究仍然处于起步阶段。充分利用蒙古语的特点，是提高汉蒙统计机器翻译性能的好方法。通过对汉蒙机器翻译中调序模型的研究，我们给出了一种基于相对偏移概率的调序方法。通过实验看到，这种方法能够有效地提高汉蒙机器翻译的性能。

本文主要为基于短语的汉蒙机器翻译提出了一种基于词语语序变化分布特点的调序模型。该模型比较简单，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。

虽然我们在汉蒙统计机器翻译方面做了一些工作，但是，在词干、词缀层面的工作还停留在语言模型方面，翻译模型中并没有利用词干、词缀信息。通过将翻译的基本单位从词改变到词干词缀，可以得到更深层次的语言信息，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。

参考文献

[1] Och F J, Tillman C, Ney H. Improved alignment models for statistical machine translation. In: Proceedings of the Conference on Empirical Methods of Natural Language Processing, College Park, Maryland, USA, 1999.20-28

[2] Koehn P, Och F J, Marcu D. Statistical phrase-based translation. In: Proceedings of the Homan Language Technology/North American Chapter of the Association for Computing Linguistics 2003, Edmonton, Canada, 2003. 127-133

[3] 侯宏旭，刘群，那顺乌日图等. 基于实例的汉蒙机器翻译. 中文信息学报，2007，21(4)：65-72

[4] Brown P F, Cocke J, Della Pietra S A, et al. A statistical approach to machine translation. *Computational Linguistics*, 1990, 16(2):79-85

[5] Brown P F, Della Pietra S A, Della Pietra V J, et al. The mathematics of statistical machine translation parameter estimation. *Computational Linguistics*, 1993, 19(2):263-311

[6] Liu Y, Liu Q, Lin S. Tree-to-string alignment template for statistical machine translation. In: Proceedings of the International Conference on Computing Linguistics/Annual Meeting of the Association for Computing Linguistics, Sydney, Australia, 2006. 609-616

[7] Xiong D, Liu Q, Lin S. Maximum entropy based phrase reordering model for statistical machine translation. In: Proceedings of the International Conference on Computing Linguistics/Annual Meeting of the Association for Computing Linguistics, Sydney, Australia, 2006. 521-528

[8] Tillmann C, Zhang T. A localized prediction model for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computing Linguistics, Ann Arbor, Michigan, USA, 2005. 557-564

[9] Watanabe T, Sumita E. Bidirectional decoding for statistical

machine translation. In: Proceedings of the 19th International Conference on Computational Linguistic(COLING), Taipei, China, 2002. 1078-1085

[10] 侯宏旭, 刘群, 刘志文等. SKIP-N 蒙古文统计语言模型. 内蒙古大学学报(自然科学版),2008, 39(2):220-224

[11] Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computing Linguistics 2002, Philadelphia, PA, USA, 2002. 295-302

[12] Koehn P. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, Washington D C, USA, 2004. 115-124

[13] 徐波,史晓东,刘群等. 2005 统计机器翻译研讨班研究报告. 中文信息学报,2006,20(5):1-9

[14] Och F J. Minimum error rate training in statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computing Linguistics, Sapporo, Japan, 2003. 160-167

[15] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computing Linguistics, Philadelphia, PA, USA, 2002. 311-318

A phrase based statistical Chinese-Mongolian machine translation and reordering model

Hou Hongxu[*] [**] [***], Liu Qun[**], Li Jintao[**]

([*] College of Computer Science, Inner Mongolia University, Hohhot 010021)

([**] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([***] Graduated University of Chinese Academy of Sciences, Beijing 100190)

Abstract

Based on the phrase-based statistical Chinese-Mongolian machine translation, an ordering model is put forward according to the Mongolian features, together with the corresponding drills, the decoding algorithm as well as the results of the primary experiments. Currently, the Chinese-Mongolian bilingual corpus is on a relatively small scale and its data are seriously sparse. In addition, the word order changes are dramatic and prevalent in Chinese-Mongolian translations. Consequently, the reordering method used in Chinese-English translation can not be optimally applied to the Chinese-Mongolian translation. By the assumption of the normal distribution of word-order changes after the analyses of these changes in Chinese-Mongolian translations, a probabilistic reordering model is established in the paper. According to the experimental results, the probabilistic model is superior to the ordering method in the Moses.

Key words: machine translation, statistical method, Mongolian, reorder, probability