文章编号:1001 - 9081(2007)08 - 1973 - 03

一种基于短语统计机器翻译的高效柱搜索解码器

罗 毅[1,2],李 森[1,2],张 建[1,2]
(1. 中国科学院 合肥智能机械研究所,合肥 230031; 2. 中国科学技术大学 信息科学技术学院,合肥 230027)
(luoyi@mail.ustc.edu.cn)

摘 要:描述了一种基于短语统计机器翻译的柱搜索解码器。搜索算法的效率是解码的关键,基于传统的柱搜索解码算法,提出了提高搜索效率的改进措施:动态剪枝策略改进了原来固定地剪枝对搜索当前情形反应不足的问题,提高了剪枝精度;预剪枝策略限制了较差的扩展,减少了不必要的扩展,提高了搜索速度;在研究了当前主要位置重排限制的基础上,提出了一种快速位置重排限制策略,加快了位置重排时的解码速度。此外,针对领域术语翻译唯一性问题提出了专门处理方法以提高翻译的准确度。分析对比实验结果,证明了算法的有效性。

关键词:统计机器翻译;解码器;柱搜索;动态剪枝;位置重排限制
中图分类号: TP391.2 **文献标志码**:A

Efficient beam search decoder for phrase-based statistical machine translation

LUO Yi[1,2], LI Miao[1,2], ZHANG Jian[1,2]
(1. *Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei Anhui* 230031, *china*;
2. *School of Information Science and Technology, University of Science and Technology of China, Hefei Anhui* 230027, *china*)

Abstract: An efficient beam search decoder for phrase-based statistical machine translation was described. The efficiency of search algorithm is the key to decoding process. After introducing the conventional beam search decoding algorithm, some efficiency improving measures were proposed. Dynamic pruning strategy enhanced the accuracy of pruning by improving that the original fixed pruning had not enough response to the current situation of search. Pre-pruning strategy was used to limit the poor sprawl, reduce unnecessary expansion and improve search speed. A rapid reordering constrains strategy was presented based on the research of the current major reordering constrains. In addition, the domain term always has the only translation, so a special process approach was put forward to improve the accuracy of the translation. Comparative analysis of the experimental results proves the effectiveness of the algorithm.

Key words: statistical machine translation; decoder; beam search; dynamic pruning; reordering constrains

0 引言

目前统计机器翻译采用基于短语[1]的方法将句子从源语言翻译成目标语言。这个方法在原来 IBM 模型词对词翻译基础作了重大的改进[2]。解码是统计机器翻译的关键。解码是指在给定语言模型、翻译模型和其他模型的基础上,根据输入的源语言句子找到概率最大的译文。无论模型如何解码过程都将是一个求解离散最优解的 NP 问题[3]。

柱搜索算法是一种在人工智能领域中广泛使用的搜索算法。它采用宽度优先的方式构建搜索树,在搜索树的每层采用启发式函数对扩展的状态进行评分,通过剪枝选取前 N 个最优的状态进行扩展。

柱搜索算法结合了启发式规则具有高效的剪枝特性,又能有效地减小收敛到局部最优解的风险,可以在解码速度和质量取得良好的折中,得到了广泛的重视[4-6]。

在使用柱搜索算法进行解码时,好的启发式函数可以降低陷入局部最优的风险,同时改进剪枝策略也可以提高搜索的精度和效率。此外,位置重排的限制也是影响解码器性能的一个重要因素。本文在研究常用位置重排限制方式基础上,提出了位置重排限制的改进方法。还针对领域术语翻译的特殊性提出了一种专门的处理方法,提高其翻译效果。

1 系统概述

本解码器采用动态规划的柱搜索算法。首先需要从双语语料库中训练出短语翻译模型和目标语言模型,根据对数线性模型[7]的思想我们还加入了其他一些附加模型。在开始解码时将输入的源语言句子划分为多个重叠的词语片段(短语),并计算它们的模型评分和目标短语,然后根据这些短语信息对源语言句子短语进行逐个翻译,从左至右生成译文。考虑源语言句子短语的位置重排,搜索出特定源/目标句子短语对齐下的最佳译文。

1.1 语言模型和短语翻译模型

我们采用文献[5]中描述的短语翻译模型,使用 GIZA++[1]进行源语言与目标语言的双向词语对齐,利用启发式的规则在双向对齐求交集和并集的基础上进行对齐扩展[5,6],然后根据扩展得到的词语对齐进行短语抽取来构建短语翻译模型。

对抽取的短语使用极大似然法计算抽取出来短语翻译对的双向翻译概率,并在双向词语对齐抽取的词典翻译概率上计算短语的词典概率[5]。

语言模型用于评价译文的忠实度和流利度。在解码过程中我们使用了 SRILM 语言模型训练工具[8]训练的 N-gram 的

收稿日期:2007 - 02 - 15;**修回日期**:2007 - 05 - 14。 **基金项目**:中国科学院知识创新工程重要方向项目(KGCX2-SW-511)。
作者简介:罗毅(1983 -),男,湖南耒阳人,硕士研究生,主要研究方向:自然语言处理、机器翻译; 李森(1955 -),女,安徽庐江人,研究员,博士生导师,主要研究方向:人工智能与农业知识工程; 张建(1954 -),男,陕西延安人,副研究员,主要研究方向:人工智能及其应用。

语言模型。

1.2 附加模型

根据对数线性模型[7]的思想,可以轻松加入多个模型。本文讨论的解码器加入了扭曲模型和词语惩罚模型[6]。扭曲模型对源语言短语在翻译的位置重排进行惩罚。扭曲模型的计算方法如下:

$$Pr(e,f) = -\left|\sum_{i=1}^{I} last\_word_{i-1} + 1 - first\_word_i\right|$$

为了防止目标语言句子过长,通过加入词语惩罚模型对短句子进行补偿,加入的词语惩罚模型为 $Pr(e) = \exp(I)$,其中 I 为译文句子长度。

1.3 柱搜索解码算法

解码算法是解码器的核心。解码时从一个没有翻译的初始状态开始,每翻译一个短语生成一个新的状态直至所有源语言句子的所有短语翻译完成。本解码器所采用的柱搜索算法的解码过程如下:

1)加载短语翻译模型和目标语言模型;
2)构建翻译候选项列表;
3)构建将来概率表;
4)生成初始状态,并加入到第 0 栈 stack[0]中;
5)从 $i = 0$ 到 $i = n - 1$ 执行如下循环:{
6)对于栈 stack[i] 每个状态 state 执行如下循环:{
7)在 state 上生成新的状态,加入相应栈中;}
8)对 stack[$i + 1$] 进行剪枝};
9)选择 stack[n] 最好的状态回溯译文。

步骤 1)中以哈希表的结构加载短语翻译模型,该哈希表以源语言短语为键,目标短语和翻译概率信息为值。这样对短语翻译模型的访问接近常数时间。步骤 2)中,根据短语翻译模型收集源语言句子的所有可用源短语信息,避免扩展新状态时的重复查找和计算。

在栈剪枝前,需要对栈中的状态进行评分。评分函数 $f(i) = g(i) + h(i)$,$g(i)$ 表示从初始状态到当前状态的所有模型的加权评分和,$h(i)$ 为当前状态到目标状态的估计值。与 Pharaoh 解码器[7]类似,我们采用动态规划的方法计算最大的未翻译短语评分值(包括语言模型和翻译模型)构建步骤 2 的将来概率表。此外还对将来概率信息进行扩展,加入了扭曲模型的最佳估计值。这样 $f(i)$ 更好地反映了当前状态的质量。

2 解码器的优化实现

在柱搜索算法进行解码时,通过剪枝可以大大减少搜索空间提高解码速度。剪枝的效率直接影响到搜索性能,通过使用高效的动态剪枝方法和预剪枝策略来提高剪枝的效率。此外改进了的位置重排限制策略提高位置重排扩展的速度。

专业术语在翻译过程具有领域唯一性和稳定性,针对这些特殊性对其进行特殊处理可提高对它的翻译效果。

2.1 动态剪枝

在解码的步骤 7)中对栈中状态进行剪枝,选择评分较好的状态进行下一步扩展。在以往的解码器中[4-6],都是根据经验设定固定的栈大小和评分阈值进行剪枝。由于搜索过程中栈中状态的个数在不断变化,这样固定的剪枝策略很难反映当前搜索的情形。

由于栈中状态的个数很容易确定,可以根据当前栈中状态的个数确定剪枝的数量。最简单的方法就是设定一个固定的标尺,比如 1/5,但是这样在状态数目比较少时(比如 100 个),删除了些好的状态,而在状态数目较多时(比如 1 000 个),却又保留了较多的差的状态。为此可以通过设定多个不同标尺解决这个问题,每个标尺对应一个状态数目的范围。对于标尺 S 的描述如下:

$s = S_i$; 当 $num \in [n_{i-1}, n_i]$, $i \in [0, N]$

其中 num 为当前状态的个数,N 为标尺的个数,S_i 为 n_{i-1} 到 n_i 范围所对应的标尺值。这样就可以灵活控制标尺了。显然,所有标尺组成的集合 $T = \{S_i\}$ 是个降序序列。

2.2 预剪枝策略

虽然短语翻译模型和目标语言模型都采用哈希表存储,能够进行快速的访问,但是很多生成的状态最终又会被剪枝删掉。对这些较差状态的扩展和删除,提出了一种预剪枝的策略,即在生成这些状态之前就停止对它们的扩展来提高搜索效率。

在扩展过程中记录每个栈中最好状态的评分,当有更好评分的状态出现时,更新这个栈的最好评分值记录。在最好的评分基础上设置一个栈阈值,对扩展前超过栈阈值的新状态不进行扩展。新状态采用如下公式估计评分:

$$Escore = preScore + \lambda_t \cdot tmScore + \lambda_l \cdot lmScore + \lambda_d \cdot dScore + fScore$$

式中 $Escore$ 为新状态的估计评分值,$preScore$ 为其父状态除 $h(i)$ 外的评分值,$tmScore$,$lmScore$ 分别为新状态所用翻译备选项的翻译模型与语言模型评分,$dScore$ 和 $fScore$ 是新状态的扭曲模型和将来概率评分值,λ_t、λ_l、λ_d 为模型的权重。如果 $preScore$ 与 $fScore$ 之和超过栈的阈值时就可以停止对新状态的扩展,否则进一步估算 $Escore$ 的值决定是否进行扩展。

为了提高预剪枝全局搜索时的准确性,应该尽量先扩展评分较好的翻译备选项。这样使栈的最好评分值能够比较稳定,采用的栈阈值能够更准确地进行预剪枝。预先对所有不同源短语的翻译备选项按将来概率进行排序,其中相同源短语的翻译备选项之间进行再排序,这样对排好序翻译备选项依次扩展就可以了。

2.3 位置重排限制

对源语言句子短语进行位置重排反映了源/目标语言之间词语位置的差异。若位置重排没有任何限制,搜索就变成 NP 问题[3],一般通过对位置重排的限制来降低搜索复杂度。当前使用最多的位置重排限制方式有三种:受限的自由排序、ITG 限制和 IBM 限制[9]。

受限的自由排序允许任何源短语在所限制的重排范围内都可以跟在任何其他短语后。这种重排限制保证所指定范围内重排位置的完全性,但是它会导致很多扩展的浪费。比如一个四个词语的源句子序列 1234,设定限制范围为 2。如果已按 132 的重排顺序进行扩展,使用扭曲模型计算公式,从 1 到 3 的扭曲值为 $D(1,3) = |1 + 1 - 3| = 1$,从 3 到 2 的扭曲值 $D(3,2) = |3 + 1 - 2| = 2$,而对于剩余 4 位置单词来说,从 4 到 2 的扭曲值 $D(4,2) = |4 + 1 - 2| = 3$ 超过了限制范围导致这条扩展路径失败。越长的句子这样的情形会越多,扩展效率很低。

根据文献[10],ITG 限制的扩展速度比受限的自由排序慢很多,但它可以限制交叉重排情况,如 2431 和 3142 排序。IBM 限制不允许源短语重排的位置超出 i 个当前没翻译的源短语的范围。IBM 限制的扩展比 ITG 快[10],当 i 越大时包含的重排情况也越多,其扩展速度也会随之下降。

为了提高位置重排的精度、缩短解码时间,提出了一种新的位置重排限制方法。

在此方法中,设定 $lmin$ 为源句子中左边第一个未翻译的词位置,C 为扩展后翻译单词总数,当前扩展的源短语 P 的第一个词位置为 $Pfirst$,最后一个词位置为 $Plast$。D 为位置重排的扭曲值。在限制范围为 k 时,规定不允许下面任意条件成立时的位置重排:

1) $lmin > C$;
2) $C < Pfirst$;
3) $D > k$;
4) $|Plast - lmin| \geqslant D$;

条件 1、2 保证位置重排从未翻译的短语开始,条件 3、4 使得重排满足扭曲限制。在上面的重排限制下,对于 4 个词语输入可以得到图 1 所示的扩展顺序图。

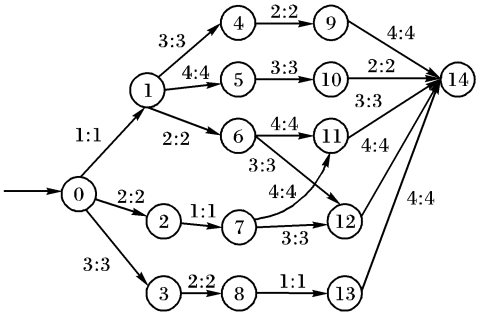

图 1　长度为 4 的源句子位置重排扩展图 ($k = 2$)

改进的重排限制方法要求必须扩展当前第一个未扩展的位置,同时也要求新的扩展要满足扭曲限制。这种位置重排限制包含受限的自由排序方法的所有扩展路径,而且不会出现扩展浪费。此外它像 ITG 限制一样不允许交叉重排情况。

2.4　术语翻译

在翻译的过程中常常要面对领域术语的问题。由于翻译模型中的短语都是从语料库中进行抽取过来的,对于每一个源语言短语一般会出现多个目标语言短语,且其翻译概率不同。领域术语要求源/目标语言的短语对翻译具有一致性,对于领域术语的翻译信息我们不能再直接依靠翻译模型。

进行领域术语翻译时,要构建一个术语翻译表,表中存放术语的源语言短语和目标语言短语对,以及其所对应的领域标识。进行解码时首先要加载领域术语翻译表,然后查找源语言句子中的领域术语,并对领域术语的目标短语进行唯一指定,同时忽略该术语对应的翻译模型信息。这样保证在搜索过程中领域术语只存在唯一的翻译选择,提高了领域术语翻译的准确性。

3　实验结果

实验使用了资助项目中 12000 句对的汉蒙双语语料库为基础。以蒙语为目标语言训练 3-gram 的蒙语语言模型和汉蒙短语翻译模型。选用 100 句长度为 10 到 20 个词语的汉语句子作为测试输入数据,进行了动态剪枝、预剪枝、位置重排限制的对比实验。

表 1　动态剪枝结果对比

剪枝策略		速度/词语·s^{-1}	错误率/%
固定栈	200	12.3	30.17
大小	400	14.8	29.24
剪枝	800	16.1	27.37
动态剪枝		11.2	26.14

表 1 为动态剪枝方法与固定栈大小的实验结果,可以看出,动态剪枝提高了剪枝的精度,但剪枝的速度有所降低。表 2 是固定栈为 200 时,预剪枝策略的实验对比结果,可以看出预剪枝在速度上提高了将近一倍。表 3 为不同位置重排限制方式下的实验对比结果,改进的限制策略能够快速进行重排扩展,并且由于包含了更多的重排路径使正确率得到一定提高。

表 2　预剪枝结果对比

剪枝策略	速度/词语·s^{-1}	错误率/%
没有预剪枝	12.3	25.41
预剪枝	20.8	26.14

表 3　位置重排限制方式对比

位置重排限制($k = 2$)	速度/词语·s^{-1}	错误率/%
自由位置排序	11.2	25.81
IBM 限制	10.8	26.23
ITG 限制	8.6	26.47
改进的限制策略	16.3	25.73

4　结语

解码器的研究是统计机器翻译研究的关键。在解码过程中遇到庞大的搜索空间,如何快速准确的进行搜索是评价解码器的重要标准。在介绍传统的柱搜索解码算法后,提出了动态剪枝、预剪枝和改进的位置重排限制策略以提高搜索速度和精度。此外,针对领域术语翻译唯一性问题提出专门处理方法以提高翻译的准确度,当然,对于如何使搜索结果达到全局最优或更接近全局最优以及位置重排限制时如何提高全局搜索的完全性还要进行进一步的研究。

参考文献:

[1] OCH F J, TILLMANN C, NEY H. Improved alignment models for statistical machine translation[C/OL]// Proceedings of EMNLP/WVLC'99, June 1999 [2007-01-16]. http://acl.ldc.upenn.edu/W/W99/W99-0604.pdf.

[2] BROWN P, COCKE J, PIETRA S, *et al*. A statistical approach to machine translation[J]. Computational Linguistics, 1990, 16(2): 79-85.

[3] KNIGHT K. Decoding complexity in word-replacement translation models[J]. Computational Linguistics, 1999, 25(4): 607-615.

[4] TILLMANN C, NEY H. Word reordering and a dynamic programming beam search algorithm for statistical machine translation[J]. Computational Linguistics, 2003, 29(1): 97-133.

[5] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[EB/OL]. [2007-01-16]. http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/phrase2003.pdf.

[6] KOEHN P. Pharaoh: A beam search decoder for phrase-based statistical machine translation models[C/OL]// Proceedings of the Association of Machine Translation in the Americas (AMTA-2004) (2004-10) [2007-01-16]. http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/pharaoh-amta2004-slides.pdf.

[7] OCH F J, NEY H. Discriminative training and maximum entropy models for statistical machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, PA: [s.n.], 2002: 295-302.

[8] STOLCKE A. SRILM-an extensible language modeling toolkit [EB/OL]. [2007-01-16]. http://www.speech.sri.com/cgi-bin/run-distill?papers/icslp2002-srilm.ps.gz.

[9] ZENS R, NEY H, WATANABE T, *et al*. Reordering constraints for phrase-based statistical machine translation[C/OL]// The 20th International Conference on Computational Linguistics, Geneva, Switzerland, August 23-27, 2004 [2007-01-16]. http://pfstar.itc.it/public/publications/rwth-Coling_2004.pdf.

[10] ZENS R, NEY H. A comparative study on reordering constraints in statistical machine translation[C]// Annual Meeting of the ACL: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, Japan. Morristown, NJ: Association for Computational Linguistics, 2003, 1: 144-151.