基于层次短语的统计翻译系统中规则冗余的高效约束方法

方李成 宗成庆

中科院自动化研究所模式识别国家重点实验室 北京 100190

{lcfang, cqzong}@nlpr.ia.ac.cn

摘 要： 基于层次短语的统计机器翻译模型是近年来比较流行且翻译质量较好的一种模型。层次短语翻译系统有效地将同步上下文无关文法的重排序能力构建于成熟的普通短语翻译系统之上，得到了在重排序和捕捉上下文信息方面都具有优势的模型。然而，层次短语翻译系统在计算复杂度方面远高出普通短语翻译系统，使用的规则存在大量的冗余。本文分析了基于层次短语的翻译系统的规则冗余问题，提出了一种基于重排序分割点的约束方法，使得学习重排序规则的训练过程集中在训练语料中重排序真实发生的片段。实验证明这种方法大幅度减少了规则数量和解码时间，且使训练时间减少了一个量级，而翻译质量仅有微小损失，并保持了基于层次短语的翻译系统和普通短语翻译系统相比翻译质量的优势。

关键词： 统计机器翻译，层次短语，同步上下文无关文法，重排序分割点，重排序

An Efficient Constraint to Reduce the Redundancy of Rules in Hierarchical Phrase-Based Translation Systems

FANG Licheng, ZONG Chengqing

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

{lcfang, cqzong}@nlpr.ia.ac.cn

Abstract: Hierarchical phrase-based translation model is a popular statistical translation model which yields high quality translation by combining the reordering power of synchronous context free grammars and the proved wisdom of conventional phrase-based translation models. However, the model suffers from a significant high computational cost and a large redundancy of rules compared with conventional phrase-based systems. This paper analyzes the rule redundancy in hierarchical phrase-based systems, and proposes a rift-based constraint that forces the rules with reordering power to focus on where reordering has actually happens. Experimental results show that our method greatly reduces the number of rules extracted and used in the system, the decoding time, and reduces the training time by an order of magnitude. The sacrifice in translation quality is little and the advantage over conventional phrase-based systems is maintained.

key words: Statistical machine translation. hierarchical phrases, synchronous context free grammars, rift, reordering

1 引言

统计机器翻译研究始于 IBM 的词对齐翻译模型[1]。为了克服基于词的翻译模型的局限性，更好地消岐和结合局部上下文信息，引入了基于短语的翻译模型，其中，Franz Och 提出的构造在双向词对齐基础上的短语提取方法[2]广为流行。然而，基于短语的翻译模型仍然无法处理长距离的依赖关系和调序问题，而长距离的依赖关系和调序问题在中英互译的过程中出现的非常频繁。因此，研究者们提出了多种基于句法的模型，这些模型多采用句法分析器对两种语言的一侧或者两侧进行分析得到分析树，然后在此基础上进行对齐。

另一种基于句法的翻译模型是以吴德凯提出的 Inverse Transduction Grammar（ITG）[3]和蒋伟提出的层次短语模型为代表的[4]。这类模型的训练的解码过程都不依赖于句法分析器，而直

接从双语语料库中学习文法。层次短语模型由于同时结合了同步上下文无关文法（Synchronous Context Free Grammar, SCFG）和普通短语翻译模型的优势，在翻译质量上与普通短语翻译系统相比有了明显的提高。而且，该模型仅仅通过一个非常简单的出发点，即“自然语言存在层次结构”，在未引入任何额外信息的情况下做到翻译质量的提升，也是受到关注的原因之一。然而，与普通的基于短语的统计翻译系统相比，基于层次短语的翻译系统的计算代价的增长也异常明显。在使用的规则数量、训练时间、解码时间上，都需要极大的额外开销，尤其是规则的冗余异常明显。

减少基于层次短语的翻译系统中的规则冗余，可以探讨两个方向：首先，为了得到更加准确也更加少量的规则，一个标准的方法就是采用 EM 算法来进行训练，然后再使用一个训练句对的最可能的 Viterbi 推导过程作为实际的观测数据，重新进行规则的计数。然而，EM 作为一种迭代方法，首先存在着训练计算量的问题，其次，为了最大化训练语料的似然度，EM 方法始终需要面对学习到的规则粒度过大、过拟合严重的问题。对于同步上下文无关文法的训练，目前尚无有效的 EM 类算法。

另一种可能的方法就是对规则加以约束，要求规则提取算法所提取出的规则形式良好。文献[5]尝试用浅层句法分析的结果作为层次短语翻译系统规则提取的出发点，要求同步上下文无关文法中的变量属于某个浅层分析的成分。文献[6]提出一种基于 Bigram 互信息的源语言句子切分方法来约束普通短语翻译系统中的规则，减少了规则数量。本质上，这两种方法都是引入了量度规则中源语言一端是否形式良好的判据，达到过滤一部分规则的目的。

本文提出的针对层次短语的翻译系统中规则冗余问题的解决方案，是基于一个简单的假设：在层次短语翻译系统中，初始短语（不具有变量）和带变量的规则包含着不同的信息，初始短语的主要功能是提供上下文信息，而带变量的规则的主要功能是提供重排序信息。因此，在我们的方法中，我们不去寻找更为良好的规则形式，而是引入重排序分割点的概念，使得带变量的规则的提取只发生在训练语料中重排序真实发生的片段。实验证明，这种方法可以大幅度减少训练得到的规则，训练时间得到了一个数量级的减少，解码时间也相应下降。而只付出了很小的翻译质量上的代价。

本文第 2 部分简要介绍层次短语的翻译模型。第 3 部分分析基于层次短语的翻译系统中的规则冗余。第 4 部分介绍重排序分割点的约束方法。第 5 和第 6 部分分别是实验和总结。

2 层次短语翻译模型

层次短语翻译模型的训练过程得到如下形式的规则：

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle$$

其中，X 是层次短语翻译模型中使用的唯一一个代表短语的非终结符，γ 和 α 都是由终结符和非终结符构成的串。~代表 γ 和 α 中非终结符的一对一的对齐。

规则通过以下步骤学习得到：

(1) 双语语料用 GIZA++进行双向对齐，对齐的结果取并集。

(2) 使用文献[2]的算法提取初始短语（Initial phrases）。

(3) 如果某个初始短语的源语言侧和目标语言侧分别能被另一个初始短语的源语言侧和目

标语言侧所覆盖。那么我们可以把两个短语相减，得到一个带有变量（非终结符）的规则。使用一个有效的 phrase-subtract 算法，以某个句对的初始短语作为输入，所有可能的带有变量的规则都被提取出来。其中有一些启发式的约束来限制规则数量（见第 5 部分）。

然后使用两个条件概率互译特征，两个词汇化特征来给规则打分，即 $P(\alpha|\gamma)$，$P(\gamma|\alpha)$，$lex(\alpha|\gamma)$，$lex(\gamma|\alpha)$。解码器利用对数线性模型把这些特征组合，用 CYK 形式的算法使用学习得到的同步上下文无关文法的源语言侧对测试集句子进行句法分析，同时生成目标翻译。具体用到的剪枝和其他参数见第 5 部分。

3 规则冗余的问题分析及约束策略

与普通的基于短语的翻译系统相比，在基于层次短语的翻译系统中，带有变量的同步上下文无关文法规则带来了很强的重排序能力，但同时也带来了很大的计算代价。通常带有非终结符的规则占到规则总数的绝大部分（见第 5 部分表 2）。同时，同 Och 的短语提取算法相比，提取带变量规则的 phrase-subtract 算法的时间复杂度也大大增加，占据了训练时间的绝大部分。

所以一个很自然的问题就是，在基于层次短语的翻译系统中，我们是否需要投入如此多的计算资源，才能达到现有的重排序能力呢？我们认为，在基于层次短语的翻译系统中，带有非终结符的规则集至少存在着以下两类冗余：

首先，通过检查提取出来的规则我们可以发现，在带有变量的规则中，存在相当多的冗余，而且类似如下形式的规则占了相当部分的比重。

$$X \rightarrow \langle 我\ X, \mathrm{I}\ X \rangle \quad X \rightarrow \langle X\ 是, X\ \mathrm{is} \rangle \quad X \rightarrow \langle X_1\ 这\ X_2, X_1\ \mathrm{this}\ X_2 \rangle$$

这些规则所包含的信息，实际上仅仅是顺序翻译。在基于层次短语的翻译系统中，理论上这些规则所完成的翻译推导完全可以由初始短语和已经存在的粘贴规则（glue rule）

$$S \rightarrow \langle SX, SX \rangle \quad S \rightarrow \langle X, X \rangle$$

所代替。

同时，带有非终结符的规则都是词汇化的，系统永远依赖某种词汇化信息来做出重排序的判断。然而，在文献[7]提出的基于 ITG 和最大熵（maximum entropy）的重排序模型中，重排序的能力来自于一个预测相邻短语是否需要重排序的最大熵分类器。而文献[7]的实验指出只需要使用两个短语的源语言端和目标语言端的首词作为特征，即可得到很好的重排序能力。这实际上证明了我们可能使用很少的词汇化特征来捕捉重排序信息。

我们提出的约束方法是，给予不带变量的初始短语和带有非终结符的规则不同的分工。由初始短语来捕捉上下文信息，由带有非终结符的规则来捕捉重排序信息。我们使用重排序分割点的概念，把训练语料划分为双语语块。由于重排序只发生在语块内，我们把提取带有非终结符的规则的过程限制在这样的双语语块内，从而达到约束规则冗余的目的。不过需要指出，在现在的基于层次短语的翻译系统中，实际上带有终结符的规则实际上是包含一部分上下文信息的，如规则

$$X \rightarrow \langle 这 是 X 吗?, \text{is this } X ? \rangle$$

就表达了一种远距离的依赖关系。

4 重排序分割点

IBM 的早期机器翻译研究中就曾经提出过重排序分割点（rift）的概念[8]。最初的目的是为了提高解码器效率，在待翻译句子中寻找“安全”的分割点来将句子分割为片段，由解码器顺序翻译。这里面“安全”就意味着在分割点的两侧重排序没有发生。

给定一个训练句对及其对齐$\langle E, F, A \rangle$，其中A中元素$a_j = i$意味着F中的第j个词f_i对应着E中的第i个词e_i。文献[8]中定义重排序分割点为F中满足如下条件的位置j：对于所有的$k < j$，有$a_k \le a_j$；对于所有的$k > j$，有$a_k \ge a_j$。也就是说，所有f_j左边的词由e_{a_j}左边的词生成，所有f_j右边的词由e_{a_j}右边的词生成。这种重排序分割点的定义反映了当时单向对齐的情况。为了使用基于层次短语的翻译系统中的双向对齐，我们扩展这个定义如下：

给定一个句对及其对齐$\langle E, F, A \rangle$，其中$A(i, j) = 1$表示e_i和f_j对齐，否则$A(i, j) = 0$。定义**重排序分割点**为二元组$\langle k, l \rangle$，对于所有满足$A(i, j) = 0$的i和j，如果$i < k$，那么$j < l$；如果$i > k$，那么$j > l$。图 1 可以很清楚地看出，这些分割点就是同时划分源语言和目标语言句子的一条不与任何对齐链接交叉的线段。需要注意的是，重排序分割点和初始短语边界是不同的，图 1 中虚线是一个初始短语的边界，然而不是重排序分割点。

图 1 重排序分割点

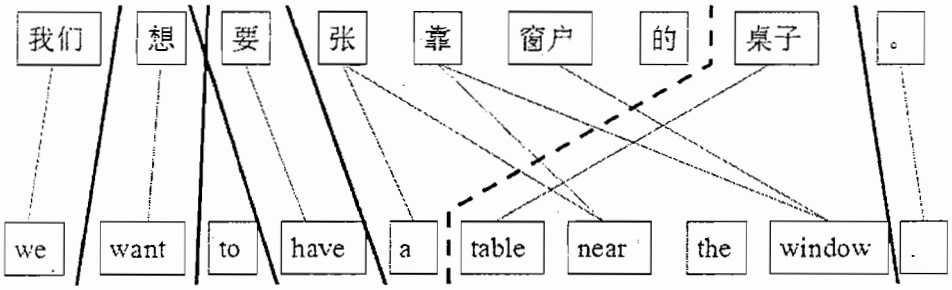

图 2 所示算法能够快速找出一个句对的所有重排序分割点。其中的S数组保存了某个源语言词以及其左侧所有的词所对应的最右侧的目标语言词，T数组刚好相反。然后在一个从左至右的扫描过程中找出所有满足条件的i和j。按照前文所述定义，连续对空的词组成的片段中可以提取多个重排序分割点，在实际中我们只提取最边缘处的重排序分割点（如上图英文一端“to”

的两侧)。

很容易看出，重排序分割点把句对分成了一些双语片段，而重排序只发生在这些片段之内，因此我们可以把提取规则的修改成下面的样子，其中带有变量的规则提取只发生在可能存在重排序的片段之内。

(1) 从句对中提取初始短语

(2) 确定句对的所有重排序分割点

(3) 标记所有跨越重排序分割点的初始短语

(4) 只把未标记的初始短语作为 phrase-subtract 算法的输入，提取带有非终结符的规则

用如上步骤得到的最终规则集合，含有和第 2 部分所述过程提取的规则集同样的初始短语，但是带非终结符的规则数量会大幅减少。

图 2 给定源语言和目标语言长度 m，n，以及对齐 A，计算重排序分割点

```
S — NULL array of length m
T — NULL array of length n
for all i such that S[i] is not null-aligned do
    S[i] = max{j|A(i, j) = 1, 0 ≤ k ≤ i}
end for
for all j such that T[i] is not null-aligned do
    T[j] = max{i|A(i, k) = 1, 0 ≤ k ≤ j}
end for
rift = [], i = 0, j = 0
while i < m and j < n do
    append (i, j) to rift
    if S[i] = NULL then
        i — i + 1
        continue
    end if
    if t[j] = NULL then
        j — j + 1
        continue
    end if
    while t[j] ≠ i or s[i] ≠ j do
        i, j — t[j], s[i]
    end while
    i — i + 1, j — j + 1
end while
return rift
```

5 实验与讨论

5.1 实验设置

我们使用一个重新实现的层次短语翻译引擎来进行分割点约束的实验。同时我们也给出我们开发的一个柱搜索解码的普通短语翻译系统在这批数据下的表现作为比较。

表 1 是我们的实验数据统计，所用语料是 IWSLT 2006 的中英翻译任务评测语料，其中包含的是旅游领域的对话语料。包括训练集 39953 句对，我们使用 IWSLT2006 的开发集 2（500 句）和开发集 3（506 句）分别作为实验的开发集和测试集，其中开发集和测试集句子都含有 16 个参

考译文。

在提取规则的过程中。初始短语的最大长度为 10，并且允许在另一种语言中没有对应的语言成分出现在初始短语的两端（软边界），因为我们发现这些没有对应的语言成分对于翻译质量有很大的帮助。带非终结符规则的非终结符个数不超过 2 且不能在规则源语言端连续出现，带非终结符规则中非终结符和终结符的个数总和不能超过 5。

所有实验使用的 Ngram 模型为在训练语料的英语端上训练的 Trigram 模型，采用 SRILM[9]生成。普通短语翻译系统的柱搜索过程中，栈中翻译假设的最大值为 200。基于层次短语的翻译系统每个栈中翻译假设的最大值为 30，同时加 $\beta = 10$ 的阈值剪枝。

规则特征的权重参数使用最小错误率训练[10]在开发集上得到，最优化 BLEU 打分[11]。结果打分使用大小写不敏感的 BLEU 打分。结果由表 1 给出，对于基于层次短语的翻译系统图中给出了在测试集上过滤前和过滤后的规则数，而解码时间是过滤后的。

表 1　实验结果比较

	BLEU	规则数量（过滤后/过滤前）	训练时间	解码时间
普通短语系统	0.5096	608,006	1'48"	2'37"
层次短语系统	0.5791	484,953/4,333,306	172'11"	17'34"
层次短语系统+约束	0.5715	128,571/1,180,768	8'57"	7'30"
加约束后变化百分比	-1.3%	-73%/-73%	-95%	-57%

5.2　讨论

实验结果表明，加入分割点约束后，BLEU 打分只相对下降了 1.3%，说明翻译质量的损失并不大。而且，重要的是，新加入的约束并没有使基于层次短语的翻译系统失掉在与普通短语翻译系统相比时的明显优势。

而在计算代价方面的降低相当显著。我们去掉了规则总数中的 73%。表 2 把规则按含有的非终结符数分类，给出了加约束和不加约束的比较。我们可以看出，加分割点约束前，带终结符规则占据了规则总数的大部分，而加约束之后规则形式的分布更符合我们的理想情况：带有重排序信息的规则应该是少量而准确的。

表 2　规则分类统计

	无约束	有约束
初始短语	618,960	618,960
带一个终结符的规则	1,900,372	336,780
带两个终结符的规则	1,813,974	225,028
总数	4,333,306	1,180,768

在时间代价方面，新加入的约束使训练时间下降了一个数量级。这种极为显著的下降可以归因于提取带非终结符规则的过程远慢于普通短语的提取，占据了规则提取的大部分时间。而约束

虽然使解码时间折半，但不如训练时间方面的提高显著，是因为解码器栈中的翻译假设数量会随着可用规则的增加指数级增长，在有无分割点约束的情况下，都需要依赖严厉的剪枝来组织搜索空间。

6 总结与展望

本文提出了一种简单有效的方法处理基于层次短语的翻译系统中存在的巨量规则冗余问题。实验证明我们能够在只对翻译质量做微量牺牲的情况下大幅降低基于层次短语的翻译系统的规则数和训练、解码时间。更重要的是，海量数据的存在使得计算代价永远是机器翻译系统的一个瓶颈，任何在某一个方面的计算代价的大幅降低都能够使得其他方面的提高翻译质量的尝试变得可行，例如：提取更长的初始短语，增大搜索空间，等等。有利于我们走向更好的质量和性能之间的平衡。

然而本文的内容仅仅表明我们可能向正确的方向迈了第一步，还有很多不足和可能的探讨。首先，仅仅在重排序真实发生的片段上进行重排序规则的训练，我们面临着过高估计重排序规则的参数的风险。我们需要更多的实验来深入理解这个问题。另外，由于规则的减少意味着搜索空间的减少，发生搜索错误的的可能降低；更少的规则也使得 EM 类的复杂度高的算法变得可行，帮助我们更好的估计规则，也就是说，更少的规则也可能是同时是更为准确的规则；所以我们相信本文描述的方法有潜力同时得到翻译质量的提高和计算代价的下降。

参考文献

[1] P.F. Brown, S.D. Pietra, V.J.D. Pietra, and R.L. Mercer. 1994. The Mathematic of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2):263-311.

[2] F.J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. Computational Linguistics, 30(4):417-449.

[3] D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics, 23(3):377-403.

[4] D. Chiang. 2007. Hierarchical phrase-based translation. Computational Linguistics, 33(2):201-228.

[5] W. Wei and B. Xu. 2007. Hierarchical chunking phrase based translation. Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007. International Conference on, pages 268-273.

[6] 周玉. 2008. 面向统计机器翻译的双语对齐研究. 中国科学院自动化研究所博士论文.

[7] D. Xiong, Q. Liu, and S. Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pages 521-528.

[8] A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39-71.

[9] A. Stolcke. 2002. SRILM-an Extensible Language Modeling Toolkit. Seventh International Conference on Spoken Language Processing.

[10] F.J. Och. 2003. Minimum error rate training in statistical machine translation. Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL), pages 160-167.

[11] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311-318.