文章编号：1003-0077(2011)02-0055-06

利用依存限制抽取长距离调序规则

涂兆鹏，刘 群，林守勋

（中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190）

摘 要：长距离调序是统计机器翻译领域的一个重要问题。层次短语模型提供了一个很好的解决方案，它使用层次短语规则可以很好地表示局部调序和长距离调序。但是，使用传统的算法抽取长距离层次规则将会导致规则表数量急剧增加，从而加大解码内存和时间消耗。为了解决这个问题，该文提出了一种利用依存限制抽取长距离调序规则的新方法。实验表明，该文的方法可以比基准系统高出 0.74 个 BLEU 点。

关键词：统计机器翻译；层次短语模型；长距离调序；依存限制

中图分类号：TP391 **文献标识码**：A

Extract Long Distance Reordering Rules with Dependency Restriction

TU Zhaopeng, LIU Qun, LIN Shouxun

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Long distance reordering is a key problem in statistical machine translation (SMT). Hierarchical phrase-based model offers an alternative to address this problem by using hierarchical rules that could characterize both local and long distance reordering. However, extracting long distance reordering rules with traditional algorithm will cuase heavy cost in decoder time-and-memory. We propose a new algorithm to extract long distance reordering rules with an extra dependency restriction. Our experiments show that our method achieves 0.74 point improvement in BLEU score.

Key words: statistical machine translation; hierarchical phrase-based model; long distance reordering; dependency restriction

1 前言

过去十年，我们见证了机器翻译领域的快速发展。短语模型[1-2]通过使用短语翻译替代字翻译来提高翻译质量，句法模型[3-5]通过加入句法信息进一步提高翻译质量。两类模型各有优缺点，具体如表 1 所示。

表 1 短语模型和句法模型的优势和不足

模型	优 势	不 足
短语模型	1. 很好地刻画短语内部的调序 2. 不依赖其他句法信息	短语间的调序较差
句法模型	很好地刻画短语内部和短语间的调序	依赖句法分析

层次短语模型[6]使用上下文无关语法规则来综合基于短语模型和基于句法模型的优势，能够很好地刻画短语内部和短语间的调序，并且不依赖于句法分析。Chiang 表明使用层次短语模型可以比当前最好的短语模型高出 1 到 3 个 BLEU 点[6]。

层次短语模型通过层次规则来表示短语间的调序。由于层次规则是从初始规则中泛化而来的，如果要抽取隐含长距离调序信息的规则，则必须先抽取长跨度的初始短语。这将会生成巨大的规则表，

收稿日期：2010-08-25 **定稿日期**：2010-11-30

基金项目：国家自然科学基金重点资助项目(60736014, 60873167)；国家自然科学基金青年基金资助项目(60903138)

作者简介：涂兆鹏(1988—)，男，博士生，主要研究方向为自然语言处理；刘群(1966—)，男，研究员，主要研究方向为机器翻译和自然语言处理；林守勋(1948—)，男，研究员，主要研究方向为多媒体技术和分布协同计算。

从而导致极大的解码系统内存和时间消耗。为了避免这个问题,Chiang限制了初始短语的最大跨度的阈值[6]。但是,这样会削弱模型的长距离调序能力,因为规则无法表示跨度大于阈值的短语间的长距离调序。

依存树能在一定程序上反映调序信息。Quirk et al.在源端使用依存树以训练一个调序模型[7];Shen et al.通过引入依存语言模型来刻画目标端依存结构中的长距离词之间的关系[8];Ding and Palmer使用依存树上定义的概率同步依存插入语法[9]。

受上述工作的启发,我们提出了一个基本但有效的方法以在层次短语模型上抽取长距离调序规则。首先,我们对训练语料的源端进行依存分析。然后,我们抽取源端为一棵完整依存子树或几棵完整依存子树集合的长距离调序规则。实验表明,我们的方法可以得到0.74个BLEU点的提高,并且规则表数量增加不大。

剩余的章节安排如下:第2节,先简单介绍短语的调序及分析为什么短语模型在短语的调序方面表现较差;第3节,介绍层次短语模型,并分析它的优势和存在的问题;第4节,描述如何利用依存限制抽取长距离调序规则,以解决层次短语存在的问题。为了解决由此带来的解码速度过慢的问题,提出了利用前缀树快速匹配规则的方法;第5节,展示实验结果及分析;最后一节,给出总结和展望。

2 短语的调序

图1中给出了一个中文句子,它对应的英文翻译和句对间的对齐。我们可以从中抽取如下短语:

打击 → the crackdown on (1)

的 成果 → results of (2)

这两个短语间的调序关系,便是短语的调序。

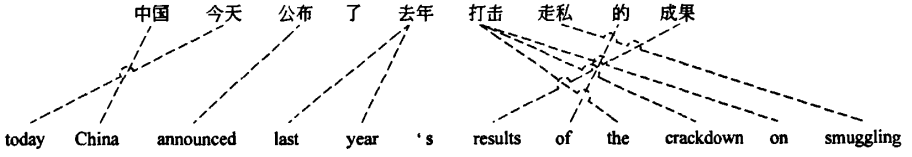

图1 一个中文句子,它的英文翻译和它们之间的对齐

短语模型可以很好地刻画短语内部的调序信息,但是对于短语间的长距离调序,短语模型表现较差。比如为了表示短语(1)和(2)的调序,短语模型可以抽取短语(3),通过短语内部的调序,来刻画短语(1)和(2)间的长距离调序。

打击 走私 的 成果→results of the crackdown on smuggling (3)

但是Koehn et al.发现当短语长度超过3的时候,对于系统的性能提高便有限,表明训练语料可能由于数据稀疏问题所以无法学到更长的规则[1]。比如解码时如果遇到下面这个词组,由于训练语料中没有出现过该词组,我们便无法找到相应的短语,这便是数据稀疏问题。对于这个词组,短语模型只能分别翻译里面的各个短语"打击","犯罪"和"的 成果",

打击 犯罪 的 成果 (4)

犯罪 → crime (5)

调用短语(1),(2)和(5),再将之顺序拼接起来,得到翻译"the crackdown on crime results of"而无法利用训练语料中短语(1)和(2)的调序信息。所以,短语模型对短语间的长距离调序能力表现较差。

为了解决这一问题,Chiang使用包含变量的层次短语规则来刻画短语间的调序[6]。

3 层次短语模型

3.1 介绍

层次短语模型是基于上下文无关语法的[6]。正式地,层次短语模型的规则可以定义如下:

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

其中,X是非终结符,γ和α是源端和目标端的字符串(由终结符和非终结符组成),$\sim$表示γ和α之间非终结符间的对齐。

层次短语模型的规则抽取可以分为两步。首先,抽取满足对齐一致性[2]的初始短语;然后,将初始短语中的子短语替换为非终结符得到层次短语。比如对于图1中所示的对齐句对,我们可以首先抽取一个满足对齐一致性的初始短语:

打击 走私 的 成果→results of the crackdown on smuggling

然后我们可以通过将子初始短语

走私→smuggling

替换为非终结符得到一条包含一个非终结符的规则：

打击 X_1 的 成果→results of the crackdown on X_1　　(6)

这里 X 表示非终结符，下标表示源端和目标端中非终结符的联系。

这样，层次短语便可以很好地表示短语(1)和(2)间的调序。当遇到词组(3)时，我们可以通过短语(5)和层次短语(6)来翻译，具体过程如下：

打击 X_1 的 成果→ results of the crackdown on X_1

⇓ 短语(5)

打击 犯罪 的 成果→results of the crackdown on crime

另外，层次短语包含了两条黏合规则：

$$S \rightarrow \langle S_1 X_2, S_1 X_2 \rangle$$
$$S \rightarrow \langle X_1, X_1 \rangle \quad (7)$$

粘合规则是用来将一系列部分翻译顺序拼接起来。

3.2 存在的问题

层次短语是通过将初始短语中的子短语替换成非终结符而得到的，这会产生极大的规则表。为了避免规则表规模过大，Chiang 限制初始短语的长度最多不能超过 L 个词[6]。但这样，对于长度超过 L 的初始短语，我们无法从中生成层次短语。那么层次短语模型就无法表示长度超过 L 的初始短语中的调序信息。

层次短语模型无法刻画长度超过 L 的两个短语间的调序，也就是长距离调序能力。下面我们将会给出长距离调序的定义，并提出一个解决方案。

4 长距离调序

长距离调序是指距离较长的两个短语间的调序，在本文中特指距离超过 Chiang 规定的最大长度 L[6]的两个短语间的调序。

4.1 利用依存限制抽取长距离调序规则

使用传统的规则抽取方法抽取长距离调序规则将会生成极大的规则表，从而影响翻译速度及所占内存。我们认为一个可能的原因是对齐一致性的约束较弱。对于长度超过 L 的初始短语，里面会包含很多满足对齐一致性的子短语，从而生成指数级的长距离调序规则。

一个解决方法是在抽取长距离调序规则时，对于子短语加入更强的限制，以减少满足条件的子短语，从而减少抽取的长距离调序规则。为了解决这一问题，我们在抽取长距离调序规则时加入依存限制，以抽取数量可以接受的高质量长距离调序规则。

图 2 显示了一个中文句子“中国 今天 公布 了 去年 打击 走私 的 成果”的依存树。箭头由子节点指向它的父节点，或称为头节点。比如在图 2 中，“公布”是“中国”的父节点或头节点。依存树可以反映词语间，尤其是较长距离的词语间的关系[7-9]。比如图 2 中，“成果”直接依存于“公布”。此外，我们观察到同时满足对齐一致性和依存结构完整性的初始短语是一个非常好的整体。比如从图 2 抽取的初始短语（去年 打击 走私 的 成果，last year's of the crackdown on smuggling）。

为此，我们限定长距离调序规则的源端必须是完整的依存结构。完整的依存结构是指一棵或多棵完整依存子树的集合。参考 Shen et al. 中对依存结构的定义[8]，我们对其严格定义如下：

定义 1：对于一个句子 $S=w_1 w_2 \cdots w_n$，$d_1 d_2 \cdots d_n$ 表示每个词的头节点（父节点），对于根节点 w_i，我们定义 $d_i=0$。一个依存结构 $d_i \cdots d_j$ 是头节点集合 H 的完整依存结构，当且仅当

- $\exists h \notin [i,j]$，　s.t. $\forall k \in H, d_k = h$
- $\forall k \in [i,j]$　and　$k \notin H, d_k \in [i,j]$
- $\forall k \notin [i,j]$，　$d_k \notin [i,j]$

图 3 给出了两个完整依存结构的例子，(a)和(b)的头节点集合分别是（中国，今天）和（成果）。我们可以发现(a)和(b)同样满足对齐一致性。

假设层次短语模型传统算法中初始短语的最大跨度 L 为 7（论文中为 10，这里为叙述方便作此假设），则对于跨度为 9 的源端“中国 去年 公布 了 去年 打击 走私 的 成果”，传统抽取算法无法处理。而我们可以通过将同时满足对齐一致性和完整依存结构限制的图 3 中(a)和(b)结构泛化成非终结符得到长距离调序规则（X_1 公布 了 X_2，X_1 announced X_2）。

由于长距离调序规则覆盖的词语较多，我们可以抽取包含多个终结符的规则。我们使用 LDDR_n 表示包含 n 个非终结符的长距离调序规则。此外，为了将长距离调序规则和普通规则区分开来，我们在解码时加入一个新的特征：长距离规则计数，计算

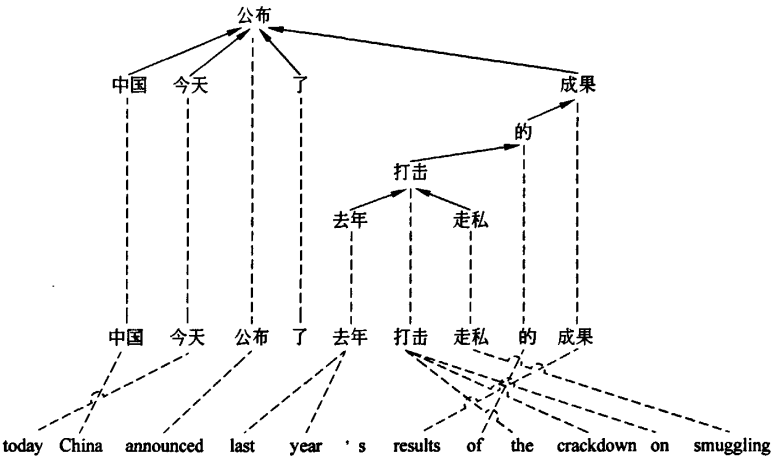

图2 一个中文依存树，它的英文翻译和它们之间的对齐(为了更清楚地表示中英文之间的联系，我们同样给出了中文句子)

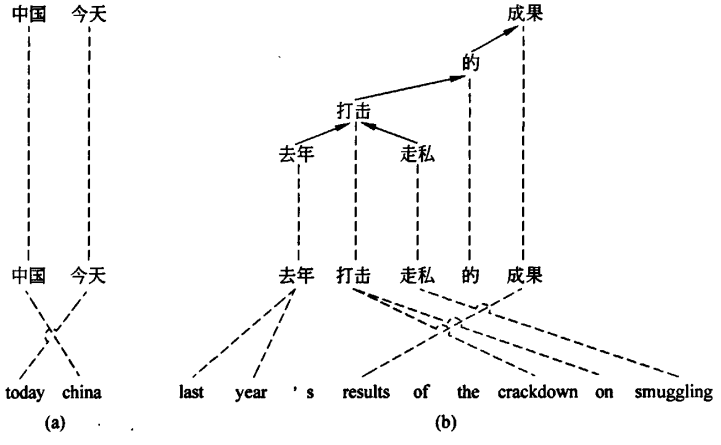

图3 完整依存结构的示例((a)和(b)的头节点集合分别是(中国，今天)和(成果))

解码时用到的长距离调序规则的数量，与普通规则计数相对应。

4.2 规则快速匹配

层次短语模型使用自底向上的CKY算法来生成推导。对于一个长度为 l 的跨度，传统的规则匹配算法是枚举出所有可能的候选规则，然后在规则表中查找。假设每条规则最多含有 m 个非终结符，则将会有 $O(l^{2m})$ 个候选规则。对于 $l>10$ 的跨度，枚举所有候选规则是非常耗时的。

受Lopez工作的启发[10]，我们使用前缀树结构存储规则，并构建词图表示候选规则。如图4所示，对于输入 $a\ b\ c\ d$，所有的候选规则只能以 a 或变量 X 起始。我们首先查找所有以 a 起始的候选规则，

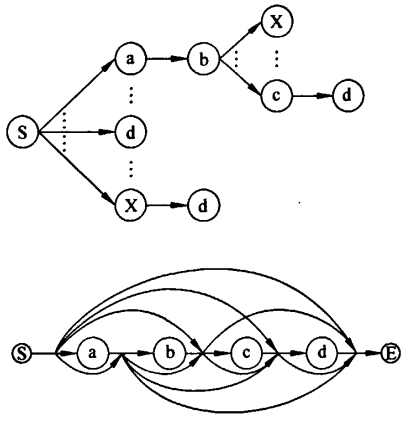

图4 前缀树规则表和词组候选规则(每条曲线箭头表示一个变量)

在规则表中我们找到了以 a 开始的规则；起始为 a 的候选规则后面只能接 b 或变量 X，然后我们在规则表中发现以 a 起始的规则后面只有接 b 的规则，所以所有 a X 起始的候选规则均不存在于规则表中。

5 实验

5.1 数据准备

我们使用 FBIS 语料（约 240K 句对）作为训练语料，并使用移进—归约的依存分析器[11]对源端进行依存分析。为了得到更好的依存分析结果，我们过滤源句子超过 40 的句对，则剩下的句对数为 190K。我们在训练数据上运行 GIZA++[12]以生成对齐句对。我们使用 SRI 工具[13]在新华语料的 GIGAWORD 部分训练一个四元的语言模型，训练中采用改进的 Kneser-Ney 平滑方法[14]。

所有的实验均是在汉—英测试集上执行的。我们用最小错误率训练[15]方法在 NIST 2002 数据集上调参，并在 NIST 2005 数据集上测试。使用大小写不敏感的 BLEU[16]测试翻译质量。

我们使用修改的层次短语模型来完成翻译，在层次短语模型上加入了一个新的特征——长距离调序规则计数，以将之和普通规则区分开。当跨度小于 10 时，我们使用传统抽取算法抽取规则；当大于 10 时，我们使用 3.1 节所定义的方法抽取长距离调序规则。

5.2 结果

表 1 列出了规则表大小和 BLEU 值。我们可以发现新增的长距离调序规则的数量是可以接受的（<10%）。当长距离调序规则所含的最大非终结符数目增加时，规则数量增加并不明显。一个可能的原因是仅有较少的初始短语同时满足对齐一致性和完整依存结构两个限制。我们发现使用长距离调序规则可以得到 0.74 个 BLEU 点的提高。

表 2　规则表大小和 BLEU 值。

	规则表	比率	BLEU 值
baseline	1.7M		30.11
LDDR_2	1.7M+190K	9.3%	30.68*
LDDR_3	1.7M+230K	10.2%	30.85**

注：比率表示 1-best 结果中长距离调序规则所占的比率。LDDR_n 表示使用最多含有 n 个非终结符的长距离调序规则。* 和 ** 分别表示 $p<0.05$ 和 $p<0.01$ 情况下的显著性提高。

表 3　不同规则匹配方法的平均时间（秒/句）。

方法	枚举规则/构建词图	规则匹配	总共
传统匹配方法	1.76	0.40	2.16
快速匹配方法	0.05	0.15	0.20

NIST05 测试集包含 1 082 个句子，平均长度为 28 个单词。规则表包含 1.7M 的普通规则和 190K 的长距离调序规则。表 3 显示了不同规则匹配方法消耗的时间。我们发现传统规则匹配方法的大部分时间花在枚举规则上。由于使用了长距离调序规则，传统方法需要枚举整个句子所有的候选规则，所以候选规则数量极其多。这也导致规则匹配所需时间稍长。而当我们使用快速匹配方法时，基本上不用花费时间构造词图，而规则匹配的时间也仅需要 0.15 秒/句，较之传统方法极大的减少了时间。这是由于我们在快速匹配时采用动态规则的方法，匹配过程舍弃了大部分不可能存在于规则表的候选规则。

6 总结与展望

本文提出了一个基本但有效的方法抽取长距离调序规则，利用依存限制减少子短语的数量，以抽取数量可以接受的长距离调序规则。相应地，我们设计了新的规则匹配算法以快速匹配长距离调序规则。实验表明使用我们的方法可以在生成较少数量长距离调序规则的情况下，得到 0.74 个 BLEU 点的提高。

尽管如此，我们的方法仍然依赖于词语对齐和依存分析。将来我们会设计新的算法以减轻对词语对齐和依存分析的依赖，比如，使用对齐矩阵[17]和依存森林[18]。

参考文献

[1] Philipp Koehn, Franz Joseph Och, and Daniel Marcu. Statistical phrase-based translation [C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003:48-54.

[2] Franz Joseph Och and Hermann Ney. The alignment template approach to statistical machine translation [J]. Computational Linguistics, 2004, MIT Press, Volume 30: 417-449.

[3] Yang Liu, Qun Liu, and Shouxun Lin. Tree-to-string alignment template for statistical machine translation [C]//Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, 2006: 609-616.

[4] Liang Huang, Kevin Knight, and Aravind Joshi. Statistical syntax-directed translation with extended domain of locality [C]//Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing, 2006: 66-73.

[5] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models [C]//Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, 2006:961-968.

[6] David Chiang. Hierarchical phrase-based translation [J]. Computational Linguistics, 2007, MIT Press, Volume 33: 201-228.

[7] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: syntactically informed phrasal SMT [C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005:271-279.

[8] Libin Shen, Jinxi Xuand, and Ralph Weischedel. A new string-to-dependency machine translation algorithm with a target dependency language model [C]// 46th Annual Meeting of the Association for Computational Linguistics,2008: 577-585.

[9] Yuan Ding and Martha Palmer. Machine translation using probabilistic synchronous dependency insertion grammars [C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005: 541-548.

[10] Adam Lopez. Hierarchical phrase-based translation with suffix arrays [C]//Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing , 2007: 976-985.

[11] Liang Huang, Wenbin Jiang, and Qun Liu. Bilingually-constrained (monolingual) shift-reduce parsing [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 1222-1231.

[12] Franz Joseph Och and Hermann Ney. Improved statistical alignment models [C]//Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 2000: 440-447.

[13] Andreas Stolcke. Srilm - an extensible language modeling toolkit [C]//Proceedings of Seventh International Conference on Spoken Language Processing, 2002: 901-904.

[14] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling [C]//Proceedings of Acoustics, Speech, and Signal, 1995: 181-184.

[15] Franz Joseph Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation [C]//Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, 2002: 295-302.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. Bleu: a method for automatic evaluation of machine translation [C]//Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

[17] Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. Weighted alignment matrices for statistical machine translation [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 1017-1026.

[18] Zhaopeng Tu, Yang Liu, Young-Sook Hwang, Liu, Qun Liu and Shouxun Lin. Dependency Forest for Statistical Machine Translation [C]//Proceedings of the 23rd International Conference on Computational Linguistics, 2010: 1092-1100.