基于强制对齐的层次短语模型过滤和优化 *

付晓寅 魏玮 卢世祥 徐波

（中国科学院 自动化研究所 数字内容技术与服务中心，北京 100190）

摘要：该文提出一种层次短语模型过滤和优化方法。该方法在采用传统方法训练得到层次短语规则的基础上，通过强制对齐同时构建源语言和目标语言的解析树，从中过滤并抽取对齐的层次短语规则，最后利用这些规则重新估计翻译模型的翻译概率。该方法不需要引入任何语言学知识，适合大规模语料训练模型。在大规模中英翻译评测任务中，采用该方法训练的模型与传统层次短语模型相比，不仅能够过滤 50%左右规则，同时获得 0.8~1.2 BLEU 值的提高。

关键词：统计机器翻译；层次短语；强制对齐；模型训练

中图分类号：TP391　　**文献标识码：**A

Filtration and Optimization for Hierarchical Phrase-based Model with Forced Alignment

Xiaoyin Fu, Wei Wei, Shixiang Lu and Bo Xu

(Interactive Digital Media Technology Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper proposes an effective method for filtering and optimizing hierarchical phrase-based (HPB) model. After obtaining the original HPB rules with traditional training method, we generate the bilingual derivation trees that represent source and target sentences with forced alignment, and then extract the HPB rules from derivation trees. At last, we re-estimated the probabilities of HPB rules with the extracted rules. This method does not need any linguistic knowledge, and it is suitable for large-scale training corpus. In the large scale Chinese-English translation tasks, our proposed method filters about 50% of the original HPB rules and improves the translation performance ranging from 0.8~1.2 BLEU on the test sets, comparing to the traditional training method.

Key words: statistical machine translation；hierarchical phrase-based model；forced alignment；model training

1 引言

层次短语模型[1-2]是目前最为实用的统计机器翻译模型之一。该模型采用单一的非终结符替换短语模型中的短语规则，不需要语言学上的标注和假设，具有良好的扩展能力。和短语模型[3]相比，层次短语模型不仅可以用短语规则描述局部的翻译信息，还可以利用层次短语规则进行短语之间的调序，因此具有更强的表达能力和长距离调序的处理能力。目前，层次短语模型已经被广泛地应用于构建统计机器翻译系统。

传统训练层次短语模型的方法一般采用启发式的训练方式[2]。该方法首先通过基于词对齐的平行语料抽取短语规则，然后从短语规则中利用子短语替换短语规则获取层次短语规则。这种启发式的模型训练方式存在两个主要问题。第一，启发式的抽取生成了大量冗余的层次短语规则，并随着语料规模的增加急剧膨胀。这样不仅造成规则的过度生成，同时在解码时容易引起搜索错误。第二，这种启发式的训练方式容易造成规则错误抽取和规则概率估

* **收稿日期：**　　**定稿日期：**

基金项目：互联网语言翻译系统研制（2011AA01A207）

作者简介：付晓寅（1986—），男，博士研究生，研究方向为统计机器翻译；魏玮（1982—），女，助理研究员，研究方向为统计机器翻译；卢世祥（1987—），男，博士研究生，研究方向为机器翻译、信息提取。

计偏差。首先，启发式规则抽取仅仅依靠基于词的对齐模型，当词对齐出现错误时，容易造成规则的错误抽取。其次，启发式抽取无法准确统计规则的频次，造成层次短语翻译模型概率估计出现误差。

针对上述层次短语模型中存在的问题，我们提出一种基于强制对齐的方法对层次短语翻译模型规则进行过滤和优化。该方法首先使用传统训练方式得到初始层次短语集合，然后采用强制对齐同时构建源语言和目标语言的双语解析树，并从中抽取出对齐的层次短语规则，最后利用这些规则重新估计翻译模型的概率。一方面，强制对齐使用层次短语规则的后验概率统计双语解析树的对齐程度，能够有效过滤层次短语中的冗余和错误。另一方面，强制对齐通过解析树统计层次短语规则频次的这个训练过程与实际解码过程相一致，从而更加准确地估计规则的翻译概率。该方法在过滤和优化层次短语模型的过程中不需要引入语言学知识，适合大规模语料训练模型。大规模中英翻译结果显示，采用本文方法和传统训练方法相比，不仅过滤了约 50%层次短语规则，同时显著提高了系统的翻译性能。

文章的组织结构如下：第 2 节简述层次短语模型过滤和优化的相关工作。第 3 节详细介绍基于强制对齐的层次短语模型过滤和优化方法。首先介绍层次短语模型的基本概念，然后介绍强制对齐的方法，并介绍规则过滤以及概率重估的方法。第 4 节描述实验设置和实验结果。最后是总结和展望。

2 相关工作

近年来，不少学者针对启发式层次短语模型训练过程中存在的规则冗余和错误，以及概率估计偏差等问题提出了各种方法进行过滤和优化。

为了减少层次短语规则中存在的冗余和错误，He[4]等人使用源语言端的关键短语过滤层次短语，不需要利用任何的语言学信息。Iglesias[5]等根据层次短语中非终结符的数目和类型对层次短语规则进行分类，并引入多种过滤策略提高翻译规则的质量。Shen[6]等人使用目标语言端依存结构大量过滤层次短语规则，造成翻译性能的降低。Wang[7]等在采用源语言和目标语言松弛依存结构对层次短语进行过滤，同时提高了系统的翻译性能。但是该方法在过滤层次短语时，依赖于双语语料的依存句法分析。由于缺乏足够的训练数据，依存分析容易引入新的错误。特别是对于训练大规模语料的翻译模型。

针对启发式规则抽取造成的规则错误和概率估计偏差，Wuebker[8]等人采用强制对齐的方式对源语言和目标语言进行短语对齐，并使用留一交叉验证估计短语翻译的概率，在短语系统上表现出良好性能。Heger[9]等在此基础上将短语对齐信息引入层次短语模型训练中，但该方法仅对短语规则进行优化，并没有对层次短语规则进行处理。Blunsom[10]等把层次短语的解析树作为隐变量，提出一种基于区分度的层次短语模型训练方法。Čmejrek[11]等人通过对语料进行双语解析从中直接抽取层次短语规则。虽然 Blunsom[10]等和 Čmejrek[11]等均采用了类似于强制对齐的方式训练层次短语模型，但是前者主要通过构建层次短语解析树来比较翻译假设与目标语言的不同进行区分度训练，训练复杂度非常大；后者则利用双语解析通过 EM 迭代重估层次短语翻译概率，并没有过滤冗余的层次短语规则。

本文通过强制对齐的方式同时对层次短语模型规则中存在的这两个问题进行有效处理。首先，强制对齐依靠翻译规则的后验概率描述解析树的对齐程度，在对齐过程中不需要引入语言学知识，同时有效过滤规则中的冗余和错误，适合大规模语料训练模型。其次，强制对齐训练过程与实际解码过程相一致，因此能够更加准确地估计层次短语规则的翻译概率。与 Blunsom[10]的方法不同，我们限定强制对齐同时匹配源语言和目标语言片段，从而大大缩小训练过程的搜索空间，提高模型训练效率和准确性。

3 基于强制对齐的层次短语模型

3．1 层次短语

层次短语模型是一种加权同步上下文无关文法（SCFG），其翻译规则可以表示为：

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle \tag{1}$$

其中，X 表示非终结符，γ 和 α 表示包含终结符或者非终结符的串，~表示 γ 和 α 中非终结符的对应关系。

为了提高模型的鲁棒性，层次短语规则中添加了两个粘贴规则。这两个规则使层次短语能够以顺序的方式进行合并。其形式如下：

$$\begin{aligned} S &\rightarrow \langle S_1 X_2, S_1 X_2 \rangle \\ S &\rightarrow \langle X_1, X_1 \rangle \end{aligned} \tag{2}$$

在翻译解码过程中，对于给定的源语言 f，层次短语规则一般会生成各种不同的解析树 D。这些解析树分别对应不同可能的目标语言翻译 e。通常采用对数线性模型[12]对解析树 D 的多个特征进行融合：

$$P(D) \propto \prod_i \varphi_i(D)^{\lambda_i} \tag{3}$$

式中 φ_i 表示定义在解析树上的特征，λ_i 表示特征的权重。常用的特征包括源语言到目标语言以及目标语言到源语言的双向短语翻译特征、双向词汇化特征、语言模型特征、单词长度特征、规则数目特征等。

3．2 强制对齐

基于层次短语的强制对齐过程可以看作是一个利用层次短语规则对训练语料进行双语解析的过程。该过程根据层次短语具有的同步上下文无关文法特性对平行句对进行解析，获得能够同时表示源语言和目标语言的双语解析树，并从该解析树中抽取得到相应的层次短语规则。

给定平行语料中的源语言和目标语言句子 $\mathbf{f} = f_1^M$ 和 $\mathbf{e} = e_1^N$。对于每个平行句对 **f** 和 **e**，强制对齐通过搜索同时匹配源语言和目标语言句子片段的规则，自底向上的构建训练语料的双语解析树。解析树上的每个节点对应一条层次短语规则。在强制对齐的过程中，采用递归的方式合并当前节点的层次短语规则和该节点的孩子节点对应的对齐片段，从而生成更长的对齐片段。这种递归方式类似于翻译过程的 CKY 解码。不同之处在于，强制对齐已知源语言和目标语言句子，因此解析树必须同时符合双语片段。此外，强制对齐并不要求双语解析树匹配整个源语言和目标语言句子，即使能够部分匹配双语句对，我们同样可以从中抽取有用的层次短语规则。

图 1 给出了一个从平行句子中利用强制对齐抽取层次短语的实例。假设层次短语集合中存在如图 1 左侧的层次短语规则，那么对于给定的源语言和目标语言句子通过强制对齐能够构建一棵如图 1 右侧的解析树。为了更清楚的说明，图中目标语言解析树中父节点的非终结规则和子节点的翻译假设进行了合并。

根据图中箭头的方向可以看到，强制对齐的过程实际上是通过层次短语规则自底向上同时构建源语言和目标语言解析树的过程。首先由于中文单词“中国”和“经济”根据短语规则能够被翻译成英文“China”和“the economy”，而且这两个翻译片段能够完全匹配目标语

言句子片段，因此我们保留这两条规则作为双语解析树的节点。然后根据规则集合中的层次短语规则对翻译片段组合，查看新的翻译片段是否能够匹配双语片段并保留该规则生成父节点。该过程持续进行直到双语句对“发展 中国 的 经济”和“developing the economy of China”完全解析。最后，我们可以得到图中所示的双语解析树。

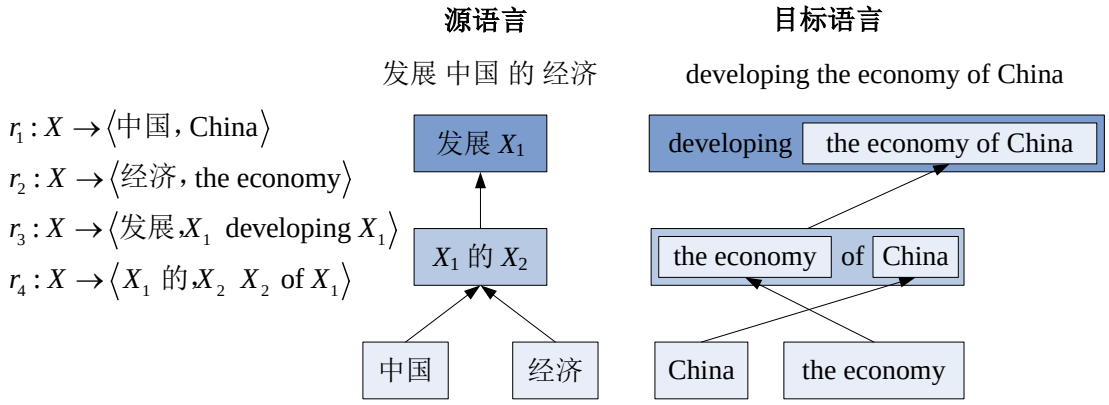

图 1 强制对齐抽取层次短语规则实例

3．3 规则抽取和概率重估

3.3.1 规则抽取

对于每一个双语平行句对，强制对齐通常会生成大量不同的双语解析树。特别是对于存在大量冗余和错误的层次短语规则，双语解析树的形式往往差别很大。为了更好的衡量双语解析树的对齐程度，我们 引入和翻译解码相似的对数线性模型对解析树进行打分。对于强制对齐形成的双语解析树，我们使用如下规则计算其权重：

$$P(D \mid e, f) \propto \prod_{i, X \rightarrow \langle \gamma, \alpha \rangle \in D} \varphi_i(X \rightarrow \langle \gamma, \alpha \rangle)^{\lambda_i} \quad (4)$$

式中 φ_i 和 λ_i 的含义与公式(3)相同，$X \rightarrow \langle \gamma, \alpha \rangle$ 表示双语解析树 D 节点上的层次短语规则。实验时，我们设定各个特征的权重 λ_i 为一个相同的值。而且在强制对齐中，我们使用的特征包括双向短语翻译特征 $p(f \mid e)$ 和 $p(e \mid f)$、双向词汇化特征 $lex(f \mid e)$ 和 $lex(e \mid f)$，规则数目特征。其中翻译概率特征和词汇化特征来自于启发式的模型训练。虽然启发式训练得到的后验翻译概率并不准确，但是能够有效衡量强制对齐解析树的对齐程度。这里我们并没有使用语言模型特征，因为强制对齐中目标语言已经给定。

当双语解析树生成后，我们通过递归的方式自顶向下回溯解析树上的每个节点，并从节点上获取用到的层次短语规则。抽取时，我们对解析树上的节点做如下约束：

1）解析树的节点打分大于阈值 τ 。

2）解析树的节点所表示的源语言跨度大于 l。

由于层次短语规则中存在一定的对齐错误，第一条约束要求强制对齐获得的对齐片段拥有较高的对齐得分，从而 在一定程度上减少规则错误造成的误差。实验中我们设定阈值 τ 为相同源语言跨度的 n 最优结果，并且取 $n=6$ 。第二条约束要求对齐的双语片段具有一定的长度。我们认为较长对齐片段的解析树打分准确性较高，实验中，我们取 $l=2$ 。通过这种方式，能够有效过滤层次短语规则中的冗余和错误规则。

3.3.2 概率重估

我们使用极大似然估计来重新计算层次短语翻译概率。和启发式的模型训练不同的是，

短语规则的频次的统计来自强制对齐的解析树，而不是词对齐信息。层次短语规则翻译概率的计算公式如下：

$$p_{FA}(f \mid e) = \frac{count_{FA}(f, e)}{\sum_{f'} count_{FA}(f', e)} \quad (3)$$

式中 $count_{FA}(f,e)$表示层次短语规则在训练语料所有双语解析树中出现的频次。在强制对齐中，层次短语规则和短语规则的概率都统一采用相同的方式进行估计，规则的频次的统计和实际解码的使用方式一致，因此能够有效避免启发式的方式造成概率估计的偏差。

4 实验结果与分析

4．1 实验设置

我们分别选取了口语和新闻两个领域的中文—英文的翻译任务来测试采用不同方法训练的层次短语模型及其翻译性能。翻译实验所用的语料规模如表 1 所示。我们使用 SRILM[13] 工具训练四元语言模型，并用 Kneser-Ney 平滑估计参数。对于口语领域，我们使用的开发集为 IWSLT07，测试集为 IWSLT08。对于新闻领域，我们使用的开发集为 NIST06，测试集为 NIST08。我们使用最小错误率训练[14]优化对数线性模型的各个参数。翻译结果的评价标准采用的是大小写不敏感 BLEU-4[15]。实验中使用的解码器是基于 CKY 方式解码的层次短语翻译系统，并使用 Cube-Pruning 裁剪搜索空间。

表 1 翻译模型和语言模型训练语料统计

翻译领域	模型	句对数	中文词数	英文词数
口语	翻译模型 [1]	382K	3.0M	3.1M
	语言模型 [2]	1.3M	——	15.2M
新闻	翻译模型 [3]	3.4M	64M	70M
	语言模型 [4]	14.3M	——	377M

4. 2 实验结果

我们首先采用传统启发式的方法训练得到初始层次短语翻译模型，并将该模型作为基准系统。训练过程中采用的约束条件与文献[2]相同。然后用本文提出的方法通过训练语料强制对齐对这些规则对进行过滤和优化。

首先，我们分别统计基准系统和强制对齐得到的层次短语模型规则数目，比较采用强制对齐的训练方式在模型过滤的方面性能。

从表 2 可以看出，虽然在启发式的模型训练过程中对规则抽取进行了约束，但是层次短语规则数目相比短语规则依然多出不少。而且在口语和新闻两个领域，采用强制对齐均能够有效过滤层次短语。和基准系统相比，口语和新闻领域规则总数目分别减少了 46%和 53%。值得注意的是，和短语规则过滤的数目相比，采用强制对齐的方式能够过滤更多的层次短语规则。这在一定程度上反映出层次短语规则存在更多的冗余和错误，通过强制对齐能够有效

[1] BTEC (Basic Traveling Expression Corpus)和 CJK (China-Japan-Korea corpus)双语语料。
[2] BTEC+CJK+CWMT2008 语料英文部分。
[3] NIST08 提供的 LDC 语料。包括 LDC2002E18，LDC2002T01，LDC2003E07，LDC2003E14，LDC2003T17，LDC2004T07，LDC2004T08，LDC2005T06，LDC2005T10，LDC2005T34，LDC2006T04。LDC2007T09。
[4] LDC2007T07 语料英文部分.

过滤这些规则。

表 2 基于强制对齐的层次短语模型过滤性能

系统	翻译领域	短语规则	层次短语规则	合计
基线系统	口语	4,643K	18,762K	23,405K
	新闻	57,272K	290,094K	347,366K
强制对齐	口语	3,807K	8,871K	12,678K
	新闻	45,817K	118,444K	164,261K

为了分别比较强制对齐在模型过滤和优化上的翻译性能，我们首先仅仅对基准系统的层次短语模型进行过滤，解码时规则的翻译概率依然采用和基准系统相同的概率进行解码。然后再对规则的翻译概率根据强制对齐结果进行重估。实验结果如表 3 所示，其中“**”表示在显著性测试中 $p < 0.01$。

表 3 基于强制对齐的层次短语模型翻译性能

系统	IWSLT 08	NIST 08
基线系统	35.39	26.43
+强制对齐（规则过滤）	36.12**	26.98**
+强制对齐（概率重估）	36.57**	27.20**

从实验结果可以看出，强制对齐在大量过滤层次短语规则的基础上，显著提高了系统翻译性能。进一步说明采用启发式模型训练得到的层次短语规则存在较多的冗余和错误，通过强制对齐能够对其进行有效过滤。仅仅通过过滤错误和冗余规则，翻译性能就已经得到显著提高。此外，采用强制对齐重新估计规则的翻译概率能够进一步提高系统的翻译性能。在 IWSLT08 测试集上，BLEU 值提高了 1.2 个点；在 NIST08 测试集上，BLEU 值提高了 0.8 个点。说明采用启发式的方法统计和估计层次短语规则得到的翻译概率并不准确。通过强制对齐能够有效估计层次短语翻译概率，并提高系统翻译性能。

5 总结与展望

本文提出一种基于强制对齐的层次短语规则过滤和优化方法。该方法利用初步训练得到的层次短语规则对双语语料进行强制对齐，构建源语言和目标语言句子的解析树，并从解析树上统计得到层次短语规则的频次，重估层次短语规则的翻译概率。强制对齐在过滤和优化层次短语规则的过程中不需要引入语言学知识，适合大规模语料训练模型。而且，强制对齐训练过程与解码过程相一致，能够更加准确地估计层次短语规则的翻译概率。实验结果显示，该方法能够过滤 50%左右的层次短语规则，同时在测试集上获得 0.8~1.2 BLEU 值的提高。

在目前工作的基础上，我们将进一步利用层次短语强制对齐从双语语料的解析树中获得的层次短语规则之间的上下文信息，用来指导解码过程中的规则选择。

参考文献

[1] David Chiang. A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of the 43rd Annual Meeting of the ACL. 2005:263-270.

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

[3] Philipp Koehn, Franz Joseph Och, Daniel Mareu. Statistical Phrase-Based Translation[C]//Proceedings of the 2003 Conference of the NAACL: HLT. 2003: 48-54.

[4] Zhongjun He, Yao Meng, Yajuan Lü, Hao Yu, Qun Liu. Reducing smt rule table with monolingual key phrase[C]//Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. 2009: 121-124.

[5] Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, William Byrne. Rule filtering by pattern for efficient hierarchical translation[C]//Proceedings of the 12th Conference of the EACL. 2009: 380-388.

[6] Libin Shen, Jinxi Xu, Ralph Weischedel. A new string-to-dependency machine translation algorithm with a target dependency language model[J]. Proceedings of ACL-08: HLT, 2008: 577-585.

[7] Zhiyang Wang, Yajuan Lü, Qun Liu, Young-Sook Hwang. Better filtration and augmentation for hierarchical phrase-based translation rules[C]//Proceedings of the ACL 2010 Conference Short Papers. 2010: 142-146.

[8] Joern Wuebker, Arne Mauser, Hermann Ney. Training phrase translation models with leaving-one-out[C]// Proceedings of the 48th Annual Meeting of the ACL. 2010: 475-484.

[9] Carmen Heger, Joern Wuebker, David Vilar, Hermaan Ney. A combination of hierarchical systems with forced alignments from phrase-based systems[C]//Proceeding of the IWSLT. 2010:291-297.

[10] Phil Blunsom, Trevor Cohn, Miles Osborne. A discriminative latent variable model for statistical machine translation[J]. Proceedings of ACL-08: HLT. 2008: 200-208.

[11] Martin Čmejrek, Bowen Zhou, Bing Xiang. Enriching SCFG rules directly from efficient bilingual chart parsing[J]. Proceeding of the International Workshop on Spoken Language Translation. 2009:136-143.

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

[13] Andreas Stolcke. SRILM – an extensible language modeling toolkit[C]//Proceedings of the 7th International Conference on Spoken Language Processing. 2002:901-904.

[14] Franz Joseph Och. Minimum error rate training in statistical machine translation[C]// Proceedings of the 41st Annual Meeting on ACL. 2003:160-167.

[15] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on ACL. 2002: 311-318.

作者联系方式：付晓寅　北京市海淀区中关村东路 95 号智能化大厦 801 室 邮编：100190
电话：15210651706 电子邮箱：xiaoyin.fu@ia.ac.cn