一种基于短语搭配和高频骨干词的层次短语约束抽取方法*

苏劲松[1,2] 吕雅娟[1] 刘群[1]
[1]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
[2]中国科学院研究生院 北京 100190
E-mail: sujinsong@ict.ac.cn

摘 要：层次短语翻译模型融合了短语模型和句法模型的优点，是当前统计机器翻译的主流模型之一。然而，传统的层次短语抽取方法在抽取过程中，只考虑了对齐一致性和规则长度的限制，没有考虑到句子中存在的短语搭配和高频骨干词，因此抽取出了大量冗余的泛化规则。对此，本文提出了相应抽取约束策略：1）采用对数似然比来确定短语搭配，并且在抽取过程中，把短语搭配当作一个整体语言单位；2）采用频率来识别句子中的高频骨干词，在抽取过程中不对完全由高频骨干词组成的子短语进行泛化。实验证明，我们提出的方法在保证翻译质量基本不变的情况下，可以大量减少冗余泛化规则的产生。

关键词：统计机器翻译，层次短语，短语搭配，对数似然比，高频骨干词

A constrained hierarchical rule extraction method based on phrase collocations and high-frequency backbone words

Jinsong Su[1,2], Yajuan Lv[1], Qun Liu[1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Beijing, 100190
[2]Graduate University of Chinese Academy of Sciences, Beijing, 100190
E-mail: sujinsong@ict.ac.cn

Abstract: *Hierarchical-phrase based machine translation model is a popular translation model which combines advantages of phrase-based translation models and syntax-based translation models. However, since there are no linguistic constraints in the procedure of current hierarchical phrase extraction, there are a large number of redundant generalized rules extracted. In this paper, we propose two strategies to limit the extraction of hierarchical rules and eliminate the number of redundant rules: first, we identify the phrase collocations with the log likelihood ratio, and then we require the phrase collocations should be packed as a whole during the extraction; second, we distinguish the backbone words using the frequency, and then set the limit during extraction that the sub phrases which consist of only backbone words can not be replaced with variables. Experimental results show that our methods substantially reduce the number of generalized rules and have no significant decrease in BLEU score.*

Key Words: *statistical machine translation, hierarchical phrase, phrase collocation, log likelihood ratio, high-frequency backbone word.*

---

* 本文的研究是在国家自然科学基金重点项目“融合语言知识与统计模型的机器翻译方法研究（60736014）”和 863 重点项目课题“面向跨语言搜索的机器翻译关键技术研究(2006AA010108)”的支持下完成的。

1 引言

层次短语翻译模型[1][2]是当今统计机器翻译[3]的主流模型之一，该模型融合了传统短语翻译模型和句法翻译模型的优点，使得翻译性能相比传统短语翻译模型有了较大幅度的提高。一方面，该模型有效地将同步上下文无关文法的重排序能力同短语翻译模型融合起来，使得模型具有较强的调序能力；另外一个方面，该模型是基于形式句法的，不需要像其它的句法翻译模型预先对源短语进行句法分析，避免了句法分析带来的分析错误和系统负担。鉴于该模型优越的性能，层次短语模型在各种统计机器翻译评测中得到了广泛的应用。

构造基于层次短语的翻译系统，我们需要从平行句对语料库中抽取层次短语（这里我们把词汇化规则和泛化规则统称为层次短语 hierarchical phrases）。与传统的短语翻译模型相比，层次短语翻译模型在提高了系统翻译性能的同时，抽取的翻译规则数量也大大增加了。如图 1 所示，当训练语料数量从 20 万句对增加到 100 万句对时，规则表所含有的层次短语数量从 2200 万增加到 8500 万。显然，在训练大量的平行句对时，层次短语的大量增加将使得翻译系统的计算代价和系统开销都大幅度增长，这给系统的实际应用带来了一定程度的困难。

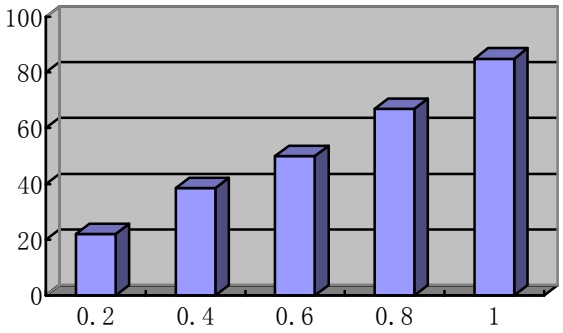


图 1　层次短语数量(Y 轴)和训练语料数量(X 轴)的关系图，单位：百万

针对该问题，国内外的学者展开了广泛研究。Shen[4]提出了首先对目标语的句子进行依存句法分析，然后用依存树 well-form 子片段结构来约束层次短语的抽取。Wei[5]通过把层次短语的抽取限制在符合语法的句法结构来减少抽取的层次短语数量。Fang[6]基于平行句对的对齐结果，提出了在抽取泛化规则时只抽取具有调序功能的泛化规则，以此大大减少了规则数量。

本文针对传统方法抽取出来的层次短语进行了分析，认为在传统方法在抽取泛化规则时，只考虑句对的对齐情况，而没有考虑句子中的其它信息：（1）句子中存在结合紧密，共现频率高的短语搭配，传统方法没有把该类短语搭配当作一个整体语言单位，抽取出大量冗余规则；（2）句子中高频词往往代表句子的骨干，对骨干词进行泛化而得到的规则泛化能力不强，也会产生大量的冗余规则。

基于以上思想，本文提出了从两个方面对传统抽取方法进行改进：（1）识别源句子中结合强度较强的短语搭配，在抽取泛化规则过程中，将其当作一个整体语言单位。（2）根据频率识别源句子中的高频骨干词，在抽取泛化规则过程中，对完全由高频骨干词组成的子短语不进行泛化。显然，该方法十分简单，并不需要像前面工作中需要依赖于句法分析或者词语对齐的结果。实验证明，我们的方法可以在保持翻译质量基本不变的情况下，大量减少抽取的层次短语数量。

本文第 2 章介绍了传统层次短语翻译模型和抽取方法。第 3 章分析了传统层次短语翻译模型抽取中存在的问题，并提出相应的约束抽取策略。第 4 章详细地描述了如何用统计方法来实现约束抽取策略。第 5 章和第 6 章是实验分析和总结展望。

2 传统层次短语抽取方法和规则形式

层次短语模型是基于同步上下文无关文法（SCFG）的翻译模型。在此，我们将对传统层次短语抽取方法和规则形式进行介绍。

传统层次短语抽取方法主要包括以下四个步骤：

（1）使用 GIZA++建立词语对齐语料库。

（2）确定初始短语对(initial phrase pairs)，该短语句对需要满足对齐一致性，边界词存在对齐，并且初始短语对的双语最大长度不超过 10。

（3）在步骤（2）的基础上，直接抽取源端长度不超过 5 的初始短语对作为词汇化规则。

（4）在步骤（2）的基础上，对初始短语对所包含的范围更小的子初始短语对进行泛化，抽取泛化规则。同样，泛化规则的源端最大长度也不超过 5，最大变量个数为 2，并且变量在源端不能连续出现。

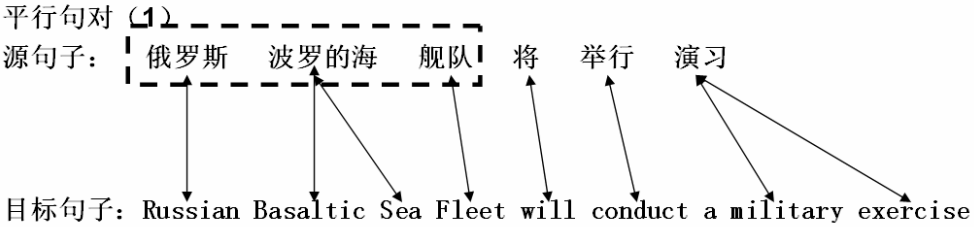
（5）采用最大似然估计的方法对规则进行打分，主要包括 2 个短语翻译概率和 2 个词汇化翻译概率。

经过以上步骤，抽取出来的层次短语规则具有如下形式：X -> < **a, r,** ~ >。其中，X 是一个代表短语的非终结符，**a** 和 **r** 都是由终结符和非终结符构成的串，~ 代表 **a** 和 **r** 中非终结符的一一对应关系。

3 传统层次短语抽取方法的问题分析和约束策略

层次短语模型之所以能超越传统短语模型，很重要的一点就是层次短语具有很强的泛化能力和调序能力。在传统抽取方法中，我们对所有可能的泛化进行了组合枚举，生成泛化规则。然而，在这个过程中，我们只考虑对齐一致性和规则长度的限制，而没有考虑语料库中其它信息。因此，抽取出来的层次短语泛化规则存在大量冗余。经过人工分析，我们认为传统方法在抽取泛化规则时存在以下几个方面的缺陷，可以通过考虑语料库本身中的语料信息，加入一定的约束策略来减少冗余规则的产生：

1）句子中存在大量的短语搭配和成对标点符号，这些短语搭配和成对标点符号在确定初始短语和抽取泛化规则时应该作为一个整体。考虑下面两个平行句对例子：



短语搭配是两个或者多个连续的词序列，该序列词与词结合紧密，并且具有句法和语义单位的特性[7]。例如：在平行句对（1）当中，“俄罗斯”、“波罗的海”和“舰队”两两结合紧密，经常作为一个整体出现，那么我们可以认为这三个词组成的序列是一个短语搭配。

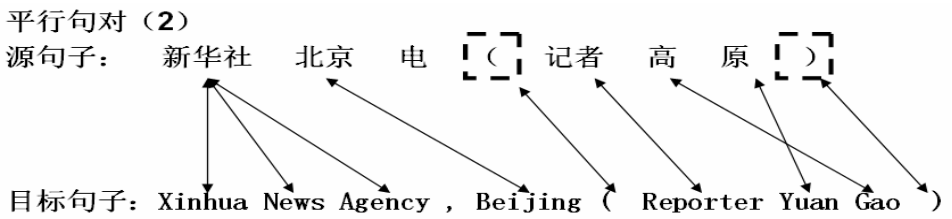
传统层次短语抽取方法并没有考虑到短语搭配现象。例如，从平行句对（1）当中，根据传统方法，我们可以抽取出以下泛化规则：

规则（1）： 俄罗斯 波罗的海 舰队 将 X1 ||| Russian Basaltic Sea Fleet will X1

规则（2）： 波罗的海 舰队 将 X1 ||| Basaltic Sea Fleet will X1

规则（3）： 俄罗斯 波罗的海 X1 将 X2 ||| Russian Basaltic Sea X1 will X2

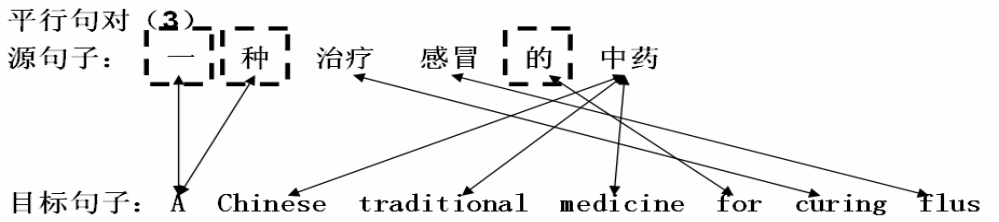
由于在语料库中，“俄罗斯 波罗的海 舰队”结合强度高，往往是作为一个整体出现，那么我们可以说从语法表达能力上来看，规则（2）是规则（1）的子串，规则（2）是冗余规则；同样，从规则变量的泛化能力上来讲，规则（3）中的变量 X1 基本上就只对应“舰队 ||| Fleet”，我们可以认为规则 3 的泛化能力不强，因此也是冗余规则。



除了上述的短语搭配，在语料库中不连续的词之间也存在结合紧密，共现频率高的现象。最为典型就是成对标点符号。例如，平行句对（2）的左括号和右括号在语料库中都是同时出现的。

对此，我们提出在抽取泛化规则时，应该考虑短语搭配和成对标点符号的整体共现性，在确定初始短语和泛化子短语的过程中，把短语搭配和成对标点符号都当作一个整体：或者同时在规则中出现，或者同时被包含于泛化子短语中。

2）句子中的高频词往往是是句子骨干，决定着一个句子的语境含义，不宜进行泛化。考虑以下句对：



在以上句对中，“一”、“种”和“的”都是语料库很经常出现的词，这类词往往构成了句子的骨干。如果对这类词进行泛化，那么我们抽取到泛化规则中的变量往往不具有普遍性。例如，从平行句对中，我们可以抽取出以下泛化规则：

规则（4）： 一 种 治疗 流感 X1 中药 ||| A Chinese traditional medicine X1 curing flus

对此，我们提出在抽取泛化规则时，必须充分考察被泛化子短语是否完全由高频骨干词构成，如果是，我们认为抽取的泛化规则泛化能力不强，是冗余规则。

4 层次短语泛化规则抽取的约束限制策略

章节 3 对传统层次短语规则抽取中冗余规则产生的原因进行了分析，并且提出在抽取过程中考虑短语搭配、成对标点符号和高频骨干词的语言特性，相应地采用一定的约束策略来减少冗余泛化规则的产生。下面，我们将对如何确定短语搭配，成对标点符号和高频骨干词进行描述。

4.1 短语搭配

短语搭配抽取一直是自然语言的研究热点。近年来，采用基于统计的方法来抽取短语搭配成为了研究主流，综合考察各种方法[8]，主要有以下几种：Dice 系数（Dice Formula），

Phi 平方系数（Phi-Square Coefficient），对数似然比（Log-likelihood Ratio，LLR）和加权同现测度（Weighted Score）。其中，采用对数似然比来抽取短语搭配的方法得到了广泛应用，该方法综合考虑 Bigram 同时出现，单独出现和不出现的情况，具有良好的抽取效果。因此，我们采用对数似然比来衡量相邻词的粘合强度，确定候选短语搭配。

采用 LLR 抽取出来的短语搭配并不是完全正确的，例如：“中国 的”。这类词对共现频率很高，相应的 LLR 值也很高，该词串并不具有合理的语言意义。我们对这类短语进行了人工分析，发现该类短语往往包含一个出现频率很高的虚词，对此，我们统计每个词的出现频率，根据词性来建立停用高频虚词词表。如果候选短语搭配包含高频虚词，则这该词串对不是一个正确的短语搭配。此外，长的词串成为短语搭配的可能性较低，对此，我们对抽取的词串进行长度限制，超过了最大长度限制的词串就不认为是正确的短语搭配。通过以上方法，我们进一步提高短语搭配抽取的准确率。

本文采用的方法步骤如下：

步骤 1、统计语料库词频，取频率最高的 topN1 个词，根据词性来建立高频虚词词表。在实验中，我们设置 topN1=100。

步骤 2、遍历语料库中所有源语言句子的相邻二元词组，计算相应的 LLR 值。在此，LLR 计算公式如下：

	词 W 出现	词 W’不出现
词 W 出现	a	b
词 W’不出现	c	d

其中：a = freq(W, W’), 词 W’和 W’相邻共现的二元词组数

b = freq(W) - freq(W, W’), 只包含词 W 的二元词组数

c = freq(W’) - freq(W, W’), 只包含词 W’的二元词组数

d = N+a-b-c，既不包含 W，也不包含 W’的二元词组数，其中 N 表示语料库中二元词组的总数。

$$LLR(W, W') = 2[\ Log\ L(p1, a, a+b) + Log\ L(p2, c, c+d) - Log\ L(p, a, a+b) - Log\ L(p, c, c+d)\ ]$$

其中，$Log\ L(p, k, n) = k \log(p) + (n-k)\log(1-p)$

$p1 = a/(a+b)$, $p2 = c/(c+d)$, $p = (a+c)/(a+b+c+d)$, $\log(0) = 0$

步骤 3、对所有二元词组根据 LLR 值进行排序。取 LLR 值最大的 topN2 个二元词组作为候选短语搭配，然后根据是否含有高频虚词和词串长度来判断抽取的词串是否是正确的短语搭配。实验中，我们设置 topN2 = 0.05*N，短语搭配最大长度为 4。

步骤 4、将确定的短语搭配加入最终的短语搭配表。如果迭代次数小于设定阈值，则转步骤 2，把前面抽取出来的短语搭配当作一个“词”，重新计算对数似然比和进行短语搭配的抽取；否则，抽取短语搭配完毕。实验中，我们设置最大迭代次数为 3。

这里需要说明的是，如果在一个句子中，我们判断重叠的两个二元词组都是正确的短语搭配，那么我们就认为含有两个二元词串的三元词串是正确的短语搭配，把这个三元词串加入最终的短语搭配表。例如，在平行句对（1）中，我们经过统计可以发现“俄罗斯 波罗的海”和“波罗的海 舰队”都是短语搭配，那么我们就认为“俄罗斯 波罗的海 舰队”是一个短语搭配，把它加入到短语搭配表中。

采用以上方法，我们可以从语料库中迭代抽取短语搭配形成短语搭配表。在抽取泛化规则的过程中，我们考察初始短语和泛化子短语的边界词和外部相邻词是否构成短语搭配，如果是，则不能抽取泛化规则。反之，按照传统的方法抽取泛化规则。

4.2 成对标点符号

成对标点符号可以看成一种不连续的短语搭配，具有和短语搭配一样的整体共现特性。语料库中的成对标点符号种类有限，我们考虑的成对标点符号主要包括：前引号和后引号，左括号和右括号，书名号等。对于此类成对出现的标点符号，我们要求成对标点符号的左右符号或者都在初始短语中出现或者都包含于泛化子短语当中。实验结果表明，该约束策略可以在一定程度上减少冗余泛化规则的产生。

4.3 高频骨干词

高频骨干词的确定方法十分简单。我们在语料库中统计词频，然后根据出现频率对所有词进行排序。取前面出现频率最高的 topN3 个词构成高频骨干词词表，在抽取泛化规则时，如果泛化短语中的每个词都是高频骨干词，那么我们对该子短语就不进行泛化。在实验中，我们取 topN3 = 100。

5 实验

5.1 实验设置

实验中，我们用 C++重新实现了著名的层次短语解码器 Hiero [1][2]。解码器是基于对数线性模型框架，采用 CKY 方式进行解码，为了加快解码速度，解码器采用 cube-pruning 方法[9]来减少搜索空间。规则特征权重参数我们采用最小错误率训练[10]在开发集上得到，其中使用的特征主要包括：

1）双向短语翻译概率
2）双向词汇化翻译概率
3）语言模型分数
4）目标译文词个数
5）粘合规则使用个数
6）规则使用个数

此外，其它参数设置如下：

候选翻译个数：50
最终翻译 N-best 个数：100

5.2 实验数据

实验中，为了测试我们的方法在不同规模的语料库上的可行性，我们分别在两个不同规模的训练语料上做了相同设置的实验。小规模的训练语料为机器翻译实验中常用的 FBIS 训练集，该语料含有 24 万句对，大规模训练语料主要来自 863LDC 平行句对语料库，共含有 155 万平行句对。表 1 是我们对两个训练数据集的统计。在实验中，我们使用 Nist02 评测集作为实验的开发集，使用 Nist05 和 Nist08 评测集作为实验的测试集。实验中所采用的语言模型是 4 元 giga 语言模型。

表 1　训练语料统计结果

	句对数	中文词数	英文词数

FBIS	239,357	6,909,918	8,972,606
863LDC	1,548,447	42,334,463	48,152,966

5.3 实验结果和讨论

对于实验结果，我们采用国际机器翻译评测中通用的大小写不敏感的 BLEU-4[11]来对系统性能进行评测。采取以上的参数设置，两个实验的结果如下：

表 2　24 万句对语料的实验结果

24 万句对	Nist05	Nist08	规则	泛化规则
传统方法	29.93	22.73	28,857,443	24,920,196
新方法	29.59	22.81	10,507,750	6,570,503
变化百分比	-1.14%	+0.35%	-63.59%	-73.63%

表 3　155 万句对语料的实验结果

155 万句对	Tst05	Tst08	规则	泛化规则
传统方法	32.77	23.75	133,171,766	114,245,530
新方法	32.33	24.13	46,322,047	27,395,811
变化百分比	-1.34%	+1.6%	-65.22%	-76.02%

表 2 和表 3 分别列出两个不同规模语料上的实验结果。从上面的数据，我们可以清楚地看到，在规模不同的训练语料上，我们的约束抽取方法在保持翻译质量变化不大（-1.34%和+1.6%）的前提下，在实验中分别减少了 73.63%和 76.02%的泛化规则，这些规则分别占规则总量的 63.59%和 65.22%。因此，可以说我们的方法是有效的。

显然，我们的约束抽取方法也可以扩展到词汇化规则。然而，对词汇化规则进行约束抽取具有较大的风险。这主要是考虑到词汇化短语本身在层次短语中占的比例并不大，仅占不到 15%，而冗余规则更多是在抽取泛化规则的过程中产生的。因此，我们将约束抽取方法只应用于抽取泛化规则。

6 总结与展望

本文针对传统层次短语中泛化规则的抽取存在的缺陷进行了初步分析，提出了导致该问题的很重要的一个原因就是传统方法只考虑了平行句对中的对齐，而没有考虑句子本身存在的语言特性。本文提出了从两个方面来改进泛化规则的抽取：1）采用对数似然比来识别源句子中的短语搭配，充分考虑短语搭配整体共现的语言特性，在泛化规则抽取过程中，把短语搭配看成一个整体语言单位。2）根据词频识别源句子中的高频骨干词，在泛化规则抽取过程中，不对完全由高频骨干词组成的子短语进行泛化。实验结果表明，我们的方法是有效的，在保持翻译质量没有显著不变的情况下，有效地减少了冗余规则的产生。

显然，我们的方法仍然存在继续改进的余地：

1）在本文中，我们主要是考察相邻短语搭配对泛化规则抽取的影响，而实际上语料库中也存在大量的非连续的短语搭配，如何准确地识别这种搭配来改进我们的抽取方法，将是我们下一步的研究工作之一。

2）在本文中，我们根据源句子中的短语搭配，成对标点符号和高频骨干词来约束泛化规则抽取，举得了不错的效果。我们相信同时利用双语中的语言特性将有潜力进一步改进我们的方法。

3）层次短语数量的减少使得采用 EM 算法来准确估计短语概率变得可能。

参考文献

[1] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In proceedings of ACL 2005, pages 263-270.

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

[3] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1994. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2):263-311.

[4] Libin shen and Jinxi Xu and Ralph Weischedel. 2008. A new String-to-Dependency machine translation algorithm with a target dependency language model. In the Proceeding of ACL 2008, pages 577-585.

[5] Wei Wei and Bo Xu. 2008. Hierarchical chunking phrase based translation. In the proceedings of NLP-KE 2007

[6] Licheng Fang and Chengqing Zong. 2008. An Efficient Approach to Rule Redundancy Reduction in Hierarchical Phrase-Based Translation. In the proceedings of NLP-KE 2008

[7] Choueka. 1998. Looking for needles in a haystack or locating interesting collocational expressional in large textual databases. In the proceedings of the RIAO 1988, pages 38-43.

[8] Yajuan Lv. 2003. Research on Automatic Translation Knowledge Acquisition Based on Bilingual Corpus Alignment. Phd Thesis. Harbin Institute of Technology

[9] Liang Huang and David Chiang. 2005. In the Proceedings of the 9th International Workshop on Parsing Technologies (IWPT)

[10] Franz Joseph Och. 2003. Minimum error rate training in statistical machine translation. In the Proceedings of ACL 2003, pages 160-167.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In the Proceedings of ACL 2002, pages 311-318.