文章编号：1003-0077(2012)01-0031-06

层次短语翻译模型的介词短语调序

冯 洋[1]，张冬冬[2]，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. 微软亚洲研究院，北京 100190)

摘 要：在不同的语言中，句法成分的相对位置往往不同，介词短语表现尤为明显，因此正确的对介词短语进行调序对提高翻译质量至关重要。层次短语模型借助于形式语法规则，具有较强的处理长距离调序的能力，但是其并不对短语的句法成分进行区分，这会导致规则的使用不当，从而引起翻译错误。该文在层次短语模型的基础上，针对介词短语进行处理。首先利用条件随机场模型识别出介词短语，然后抽取出带有介词短语的规则，构建一个新的同步上下文无关文法。解码的时候，在这个同步上下文无关文法定义的空间里搜索找到最优的译文。相对于层次短语模型，该方法在我们内部的英汉数据集上调高了 0.8 个 BLEU 百分点，在 NIST 2008 英汉翻译数据集上提高了 0.5 个 BLEU 百分点。

关键词：统计机器翻译；层次短语模型；介词短语调序；条件随机场

中图分类号：TP391 **文献标识码**：A

Prepositional Phrase Reordering for Hierarchical Phrase-Based Translation

FENG Yang[1], ZHANG Dongdong[2], LIU Qun[1]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. Microsoft Research Asia, Beijing 100190, China)

Abstract: In different languages, the relative order of syntactic constituents is usually different, especially for prepositional phrases. Therefore, to the proper treatment of the syntactic order differences is of vital importance to translation quality. Hierarchical phrase-based model learns formally syntax from a parallel corpus and is capable of dealing with long-distance reordering. But it fails in discriminating the syntactic constituents to select the correct translation rule. In this paper, we introduce linguistic information into hierarchical phrase-based model in the form of prepositional phrases so as to well capture the reordering of prepositional phrases. This method first identifies prepositional phrases via conditional random fields and extracts rules including prepositional phrases. Under the SCFG defined by these rules, it searches for the best derivation and produces translation simultaneously. Experiments show that, comparing to hierarchical phrase-based model, our method can get an absolute improvement of 0.8 BLEU point on our in-house English-Chinese test set and 0.5 BLEU point on the NIST 2008 English-Chinese test set.

Key words: statistical machine translation; hierarchical phrase-based translation; prepositional phrase reordering; conditional random field

1 引言

不同语言之间的异构性，导致了机器翻译中源语言和目标语言的词语顺序不同，词语调序成为机器翻译中的关键问题。一个句法成分覆盖的词语在经过翻译之后，其对应的译文通常也是相邻的[1]，这就是所谓的句法黏着性。于是，词语顺序的不同通常表现为在不同的语言中，句法成分的相对位置往往不同。在所有的句法成分中，介词短语表现地尤

收稿日期：2011-01-18 **定稿日期**：2011-04-25

作者简介：冯洋(1982—)，女，博士研究生，主要研究方向为机器翻译；张冬冬(1976—)，男，博士、副研究员，研究方向为自然语言处理，统计机器翻译技术；刘群(1966—)，男，博士、研究员、教授、博士生导师，研究方向为自然语言处理，机器翻译，信息提取。

为明显。以中英文为例，在中文中，介词短语通常在所修饰成分的前面，而在英文中，介词短语通常在所修饰成分的后面，例如，英文句子“Bush held a talk with Sharon”，介词短语“with Sharon”在“held a talk”的后面，而在其对应的中文句子“布什 与 沙龙 举行 了 会谈”中，介词短语“与沙龙”就在“举行 了 会谈”的前面。因此，要得到高质量的译文，不但要保证介词短语的译文不能被分开，还要保证译文被放在正确的位置上。

层次短语模型[2]可以从双语句对中自动地抽取形式语法，而不需要语言学上的标注和假设，所以使用方便，目前被广泛地应用于机器翻译中。其形式语法信息的载体为层次短语（规则），它不仅可以利用短语来捕捉一些局部翻译，而且还可以利用层次短语来捕捉子短语之间的调序，所以层次短语模型对长距离的调序具有一定的处理能力。但是形式语法也存在一定的问题，它没有对子短语覆盖的成分进行区分，这导致了子短语可以匹配任何的句法成分，这往往会带来翻译错误。例如，层次短语<held a talk X,X 举行 了 会谈>，既可以被应用在“held a talk with Sharon”，也可以被应用在“held a talk and reached an agreement”上，但是对于后者，调序发生了错误。所以，利用语言学信息来对非终结符覆盖的句法成分进行区分还是很有必要的。

本文我们在层次短语模型的基础上，以介词短语的形式引入语言学句法信息。一方面，我们利用层次短语来捕捉长距离调序，另一方面我们对介词短语进行重点处理，以保证介词短语被正确的调序，并满足句法黏着性。我们把介词短语的识别看成是一个序列标注问题，用条件随机场（Conditional RandomField，CRF）进行标注，然后基于已被识别的介词短语抽取规则，将抽取的包含介词短语的规则和层次短语模型的规则进行合并，得到一个大的规则表。然后在测试句子已被标注出介词短语的情况，用大的规则表对其进行匹配，得到最终的译文。

本文的组织结构如下：首先介绍添加介词短语的层次短语模型，再介绍如何利用条件随机场识别介词短语，接着介绍规则抽取以及如何解码，最后是实验结果和结论。

2 模型

我们在层次短语模型的基础上对介词短语进行重点处理，因为介词短语的位置在不同语言中可能会差异很大。相对于层次短语模型中，普通规则只采用一个非终结符 X，我们特地为介词短语引进一个非终结符 Y。于是，对于图 1 中所示例子，其包含两个介词短语“with Sharon”和“in US”，其可以匹配的规则包括：

r_1： $X \rightarrow \langle \text{held } X_{[1]}\ Y_{[2]}, Y_{[2]}\text{举行了 } X_{[1]} \rangle$

r_2： $X \rightarrow \langle \text{Bush}, \text{布什} \rangle$

r_3： $X \rightarrow \langle \text{a talk}, \text{会谈} \rangle$

r_4： $X \rightarrow \langle \text{with Sharon}, \text{与 沙龙} \rangle$

r_5： $X \rightarrow \langle X_{[1]}\ Y_{[2]}, Y_{[2]}\ X_{[1]} \rangle$

r_6： $X \rightarrow \langle \text{in US}, \text{在 美国} \rangle$

我们除了还需要层次短语模型的粘贴规则 r_7、r_8 之外，还需要一条规则 r_9，来实现非终结符 Y 到非终结符 X 的转变：

r_7： $S \rightarrow \langle S_{[1]}\ X_{[2]}, S_{[1]}\ X_{[2]} \rangle$

r_8： $S \rightarrow \langle X_{[1]}, X_{[1]} \rangle$

r_9： $Y \rightarrow \langle X_{[1]}, X_{[1]} \rangle$

于是，对于图 1 所示的例子，其对应的翻译过程如图 2 所示。

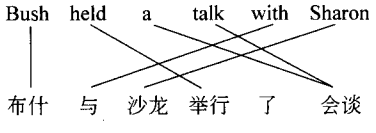

图 1 一个英汉句对

$\langle S_{[1]}, S_{[1]} \rangle$
$\Rightarrow \langle S_{[2]}\ X_{[3]}, S_{[2]}\ X_{[3]} \rangle$
$\Rightarrow \langle X_{[4]}\ X_{[3]}, X_{[4]}\ X_{[3]} \rangle$
$\Rightarrow \langle \text{Bush } X_{[3]}, \text{布什 } X_{[3]} \rangle$
$\Rightarrow \langle \text{Bush } X_{[5]}\ Y_{[6]}, \text{布什 } Y_{[6]}\ X_{[5]} \rangle$
$\Rightarrow \langle \text{Bush held } X_{[7]}\ Y_{[8]}\ Y_{[6]}, \text{布什 } Y_{[6]}\ Y_{[8]} \text{举行 了 } X_{[7]} \rangle$
$\Rightarrow \langle \text{Bush held a talk } Y_{[8]}\ Y_{[6]}, \text{布什 } Y_{[6]}\ Y_{[8]} \text{举行 了 会谈} \rangle$
$\Rightarrow \langle \text{Bush held a talk } X_{[9]}\ Y_{[6]}, \text{布什 } Y_{[6]}\ X_{[9]} \text{举行 了 会谈} \rangle$
$\Rightarrow \langle \text{Bush held a talk with Sharon } Y_{[6]}, \text{布什 } Y_{[6]} \text{与 沙龙 举行 了 会谈} \rangle$
$\Rightarrow \langle \text{Bush held a talk with Sharon } X_{[10]}, \text{布什 } X_{[10]} \text{与 沙龙 举行 了 会谈} \rangle$
$\Rightarrow \langle \text{Bush held a talk with Sharon in US}, \text{布什 在 美国 与 沙龙 举行 了 会谈} \rangle$

图 2 翻译过程

我们将模型形式化为对数线性模型，在规则上采用的特征以及整个推导上的特征与层次短语模型相同。其中，对于 r_7，其权重设为 1。

3　介词短语识别

标注介词短语，相当于为句子中的每个词打上一个标签，来表示该词是一个介词短语的开始词(标记为 B)，或者属于一个介词短语但不是其第一个词(标记为 I)，或者不属于任何一个介词短语(标记为 O)。因此，我们把识别介词短语看成是一个序列标注问题。

为了克服 label bias，并且采用尽可能多的统计信息，我们用条件随机场来进行序列标注。我们用 $x=x_1\cdots x_n$ 表示输入句子，$y=y_1\cdots y_n$ 表示相应的标记序列，一个条件随机场可以用一个局部特征变量 f 和一个特征权重向量 λ 来表示。每个局部特征可以是以下两类中的一类：状态特征 $s(y,x,i)$；转换特征 $t(y,y',x,i)$。所以，条件随机场的全局特征可以表示为：

$$F(y,x)=\sum_{i=1}^{n} f(y,x,i)$$

于是，条件随机场定义的条件概率可以表示为

$$p_\lambda(y\mid x)=\frac{\exp \lambda F(y,x)}{\sum_{y}\lambda F(y,x)}$$

一个句子可以对应多个标注序列，我们只选取其中概率最大的标注序列 $\hat{y}$ 作为最终的标注结果

$$\hat{y}=\arg\max_{y} p(y,x)$$

关于条件随机场的更多细节可以参照文献[3-5]。

识别介词短语的时候，对于每个位置 i，我们用 $l=l_1\cdots l_n$ 表示输入句子对应的词性标注序列，采用的特征如表 1 所示。

表 1　识别介词短语采用的特征

$t(y_{i-1},y_i)$	$s(x,i)$
$y_i=y,y_{i-1}=y'$	true
$y_i=y$	$x_i=x$
	$x_{i-1}=x$
	$x_{i+1}=x$
	$x_{i-1}=x',x_i=x$
	$x_i=x',x_{i+1}=x$
	$l_i=l$
	$l_{i-1}=l$
	$l_{i-2}=l$
	$l_{i+1}=l$
	$l_{i+2}=l$
	$l_{i-2}=l',l_{i-1}=l$
	$l_{i-1}=l',l_i=l$
	$l_i=l',l_{i+1}=l$
	$l_{i+1}=l',l_{i+2}=l$
	$l_{i-2}=l'',l_{i-1}=l',l_i=l$
	$l_{i-1}=l'',l_i=l',l_{i+1}=l$
	$l_i=l'',l_{i+1}=l',l_{i+2}=l$

由于采用了词性标记作为特征，所以在识别介词短语之前，需要得到每个词的词性。同样，我们也采用条件随机场来进行词性标注，采用的特征如表 2 所示。除此之外，我们还采用了一些词汇拼写方面的特征，包括：一个单词是否以数字或者大写字母开头，是否包括连字符，其后缀是否包括-ing，-ed，-ogy，-s，-ly，-ion，-tion，-ity，-ies。

表 2　词性标注采用的特征

$t(l_{i-1},l_i)$	$s(x,i)$
$l_{i-1}=l',l_i=l$	true
$l_i=l$	$x_i=x$
	$x_{i-1}=x$
	$x_{i-2}=x$
	$x_{i+1}=x$
	$x_{i+2}=x$
	$x_{i-2}=x',x_{i-1}=x$
	$x_{i-1}=x',x_i=x$
	$x_i=x',x_{i+1}=x$
	$x_{i+1}=x',x_{i+2}=x$
	$x_{i-2}=x'',x_{i-1}=x',x_i=x$
	$x_{i-1}=x'',x_i=x',x_{i+1}=x$
	$x_i=x'',x_{i+1}=x',x_{i+2}=x$

4　训练

引入介词短语的层次短语模型的训练过程与层

次短语模型相同，只是抽取规则的时候有所不同。我们的模型抽取规则的时候分为两步：第一步，不考虑介词短语，采用层次短语的方法抽取规则；第二步，抽取包含介词短语的规则。抽取包含介词短语的规则也分两步进行：首先抽取初始短语和介词短语，然后在此基础上抽取包含介词短语的层次化短语。

抽取初始短语的方法与层次短语模型相同。对于一个初始短语$<f_i^j, e_{i'}^{j'}>$，当且仅当满足以下条件，我们称之为介词短语：

- $c_i = B$；
- $c_j = I$；
- f_j 是源句子的最后一个词，或者 $c_{j+1} \neq I$。

句对$<f, e>$的包含介词短语的规则的抽取方法如下：

- 如果$<f_i^j, e_{i'}^{j'}>$是一个初始短语，则 $X \rightarrow <f_i^j, e_{i'}^{j'}>$是一条规则；
- 如果 $r = X \rightarrow <\alpha, \beta>$是一条规则，并且$<f_i^j, e_{i'}^{j'}>$是一条初始短语，$\alpha = \alpha_1 f_i^j \alpha_2$，$\beta = \beta_1 e_{i'}^{j'} \beta_2$，则 $X \rightarrow <\alpha_1 X_{\boxed{k}} \alpha_2, \beta_1 X_{\boxed{k}} \beta_2>$是一条规则，且下标 k 没在 r 中出现；
- 如果 $r = X \rightarrow <\alpha, \beta>$是一条规则，并且$<f_i^j, e_{i'}^{j'}>$是一条介词短语，$\alpha = \alpha_1 f_i^j \alpha_2$，$\beta = \beta_1 e_{i'}^{j'} \beta_2$，则 $X \rightarrow <\alpha_1 Y_{\boxed{k}} \alpha_2, \beta_1 Y_{\boxed{k}} \beta_2>$是一条包含介词短语的规则，且下标 k 没在 r 中出现。

同样，我们得到的规则数量很大，对于不包含介词短语的规则，按照层次短语模型的方法进行过滤，对于包含介词短语的规则，为了加快解码速度以及避免歧义性，我们添加了以下限制。

1）每个规则在源端和目标端的边界词均不能对齐到空；

2）初始短语所包含的源端词的个数不超过10，而层次短语在源端的符号数（包括非终结符和终结符）不超过5个；

3）每条规则在源端不能为空，且至少要包含一个终结符；

4）每条规则最多可以有两个非终结符；

5）每条规则在源端和目标端的词语之间至少要有一条对齐。

以上限制和层次短语的主要区别在于，我们允许包含介词短语的规则的两个非终结符相邻，这主要是因为句中的介词短语已经确定，两个非终结符相邻不会引起很多模棱两可组合的情况。

我们将包含介词短语的规则和层次短语模型的规则分开估计概率，概率的估计方法和层次短语模型相同。

5 解码

与层次短语模型相同，我们采用 CKY 算法来搜索概率最大的推导，并将其对应的译文作为最终的译文。我们采用柱搜索来减小搜索空间，采用的剪枝策略为：每个区间最多可以匹配的规则限制为 c 个；每个柱对应的栈中保留的译文的个数最多为 b 个；每个柱对应的栈中保留译文的分数必须大于栈中当前最好译文的分数的 β 倍。与此同时，采用 cube pruning[6] 来加快解码速度，并限制每个规则最多可以匹配的源端词语个数不超过10。于是，整个解码过程的时间复杂度为 $O(10ncb^2)$，与句子长度 n 成线性关系。

我们的模型在解码的过程中，只有规则的匹配方法与层次短语模型不同。对于每个测试句子，我们采用两部分规则：一部分是不考虑介词短语的规则，其匹配方法与层次短语模型相同；一部分是包含介词短语的规则。我们首先枚举出句中的所有包含介词短语的规则的源端部分，然后去规则表中查找相应的规则。

6 实验

我们首先测试采用 CRF 进行词性标注和介词短语识别的效果，因为介词短语识别的准确率直接影响到解码效果，然后我们测试一下引入介词短语的情况下解码的性能。

6.1 介词短语识别效果

我们将标准宾州树库的英语句法分析任务数据的1～22节的39 832个句子分成两部分，前面的38 832个句子作为训练集，后面的1 000个句子作为测试集。对于训练集语料的获得，我们采用后序遍历的方法来识别介词短语，对于标注为 PP 的节点覆盖的源语言串则标注为介词短语，且一旦一个节点被我们识别为介词短语，我们不再遍历其祖先节点，这样保证我们得到的介词短语均为最小的介词短语。为了保证与机器翻译语料的一致性，我们将宾州树库中的"和"用"来替换。CRF 采用 L_2 方法来训练。

我们采用序列标注问题中通用的标准——准确

率(P)，召回率(R)，F_1 值，来评估介词短语识别的结果。我们还采用正确率(A)来评估每个词的标注结果。对于介词短语标注，准确率 P 和正确率 A 是不同的。例如，下面的标注序列

参考序列：O O B I O O B I

标注序列：O O B I I O B I

其准确率 P 为 50%，正确率 A 为 87.5%。词性标注的结果如表 3 所示，介词短语识别的结果如表 4 所示。

从以上实验结果可以看出，词形标注的正确率比介词短语识别的正确率要高很多，这主要是因为，我们识别介词短语的时候，窗口的大小只有 3，而有的介词短语的长度超过 3，对于这一部分介词短语的识别会比较吃力。另外，由于识别介词短语的时候用词性作为特征，而词性识别的时候会引入一部分错误，这部分错误会累加到介词短语识别上来，导致最后的正确率降低。

表 3　词性标注结果

词语个数	正确率 A
23 861	95.77%

表 4　介词短语识别结果

介词短语个数	1 679
CRF 标注出的个数	1 667
CRF 正确识别个数	1 428
正确率 A	93.67%
准确率 P	85.66%
召回率 R	85.05%
F_1 值	85.36%

6.1　机器翻译性能

我们接下来比较引入介词短语的模型和层次短语模型的性能。我们采用的开发集为 NIST2008 英汉双语训练语料，除去其中的香港法律和香港会议记录部分，大约剩下 49 万句对。对于训练语料，我们先用 GIZAC++ 工具包[7]进行双向对齐，然后采用“final-and”策略将双向对齐合并成一个多到多对齐。采用的语言模型为在 GIGA 语料的新华部分上训练的一个五元语言模型，并采用 KN 方法进行平滑。我们的实验结果都进行了显著性测试[8]。

我们采用的开发集为微软亚洲研究院内部的英汉新闻测试集，包括 1 010 个句子，分别在两个测试集上比较两个解码器的性能：一个是 NIST 2008 英汉机器翻译测试集，包括 1 859 个句子，另一个是我们内部的另一个英汉新闻测试集，包括 966 个句子。翻译结果的评测标准采用基于字的 BLEU 值[9]，最高进行四元的 n-gram 匹配取。我们在开发集上采用最小错误率[10]来进行参数训练，训练的目标为使得开发集上的 BLEU 值最大。

表 5　机器翻译性能比较

系统	开发集	内部测试集	NIST 2008
层次短语模型	0.284	0.532	0.392
介词短语模型	0.292	0.540**	0.397**

表 5 给出了实验结果，“**”表示在显著性测试中 $\rho<0.01$。从实验结果可以看出，引入介词短语之后，在我们内部的测试集上，BLEU 值提高 0.8 个点，在 NIST 2008 上提高了 0.5 个点。性能提高的原因在于通过引入介词短语，可以针对介词短语选择更好的规则，从而减轻引言中提到的由于 X 可以匹配任何短语而导致规则使用不恰当的情况。

7　结论

层次短语模型在短语模型的基础上，引入在双语句对上自动学习得到的形式语法信息，这些形式句法信息不需要基于语言学的标注和假设，使得形式短语模型用起来很方便，所以现在层次短语模型使用很广泛。在层次短语模型中，形式句法信息是以层次短语为载体的。而层次短语由词和短语组成，所以层次短语模型一方面可以通过短语来学习局部翻译，一方面可以利用层次短语来掌握短语之间的调序，所以层次短语具有一定的捕捉长距离调序的能力。由于形式语法并不对每个短语的句法成分进行细化，这导致了层次短语在规则匹配的时候可能会被用在不恰当的地方，所以对层次短语的短语进行句法标注还是很有必要的。我们尝试在层次短语模型的基础上，以介词短语的形式来引入语言学句法信息，并对介词短语的调序进行重点处理。由于介词短语在不同语言中相对位置差异很大，如此可以以较小的代价来获得翻译性能的较大提高。

对于介词短语的识别，我们采用序列标注的方法，通过对宾州树库中的句法分析树进行处理来得到短语识别的训练语料，来训练得到一个条件随机场(CRF)。然后用训练得到的 CRF 在机器翻译任

务的训练集上识别介词短语，对于规则抽取，除了抽取层次短语模型的规则，还抽取一些包含介词短语的规则。在训练和解码的时候，也是先识别介词短语，然后一起应用两部分规则，一部分是层次短语模型的规则，一部分是包含介词短语的规则。实际上，抽取出来的介词短语并不多，对解码器的速度影响不大，却能取得显著的效果，在我们内部的英汉翻译数据集上可以提高 0.8 个 BLEU 值，在 NIST2008 英汉机器翻译测试集上可以提高 0.5 个点。这充分说明语言学句法信息对提高机器翻译性能还是很有帮助的。

参考文献

[1] Heidi Fox. Phrasal Cohesion and Statistical Machine Translation[C]//Proceedings of EMNLP, 2002: 304-311.

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

[3] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proceedings of ICML, 2001: 282-289.

[4] Ben Taskar, Pieter Abbeel, Daphne Koller. Discriminative Probabilistic Models for Relational Data[C]//Proceedings of Eighteenth Conference on Uncertainty in Artificial Intelligence, 2002.

[5] Fei Sha, Fernando Pereira. Shallow Parsing with Conditional Random Fields [C]//Proceedings of HLT-NAACL, 2003: 134-141.

[6] Liang Huang and David Chiang. Better k-best parsing [C]//Proceeding of IWPT, 2005: 53-64.

[7] Franz Josef Och, Hermann Ney. Improved Statistical Alignment Models[C]//Proceedings of the 38th ACL, 2000.

[8] Michael Collins, Philipp Koehn, Ivona Kucerova. Clause restructuring for statistical machine translation [C]//Proceeding of ACL, 2005: 531-540.

[9] Kishore Papineni, Salim Roukos, Todd Ward, et al.. Bleu: a Method for Automatic Evaluation of Machine Translation [C]//Proceedings of the 40th ACL, 2002: 311-318.

[10] Frans J. Och. Minimum error rate training in statistical machine translation [C]//Proceeding of ACL, 2003: 160-167.

（上接第 8 页）

[2] 吴云芳，王淼，金澎，等. 多分类器集成的汉语词义消歧研究[J]. 计算机研究与发展，2008，45(8)：1354-1361.

[3] 全昌勤，何婷婷，姬东鸿，等. 基于多分类器决策的词义消歧方法[J]. 计算机研究与发展，2006，43(5)：933-939.

[4] Latinne P, Debeir O, Decaestecker C. Combining Different Methods and Numbers of Weak Decision Trees [J]. Pattern Analysis & Applications, 2002, 5(2): 201-209.

[5] 张仰森，郭江. 四种统计词义消歧模型的分析与比较. 北京信息科技大学学报，2011，26(2)：13-18.

[6] Kilgarriff A, Rosenzweig J. Framework and results for English SenSeval[J]. Computers and the Humanities 34: 15-48, 2000.

[7] Xiaojie Wang, Yuji Matsumoto. Trajectory based word sense disambiguation [C/OL]//COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. http://aclweb.org/anthology/C/C04/C04-1130.pdf.