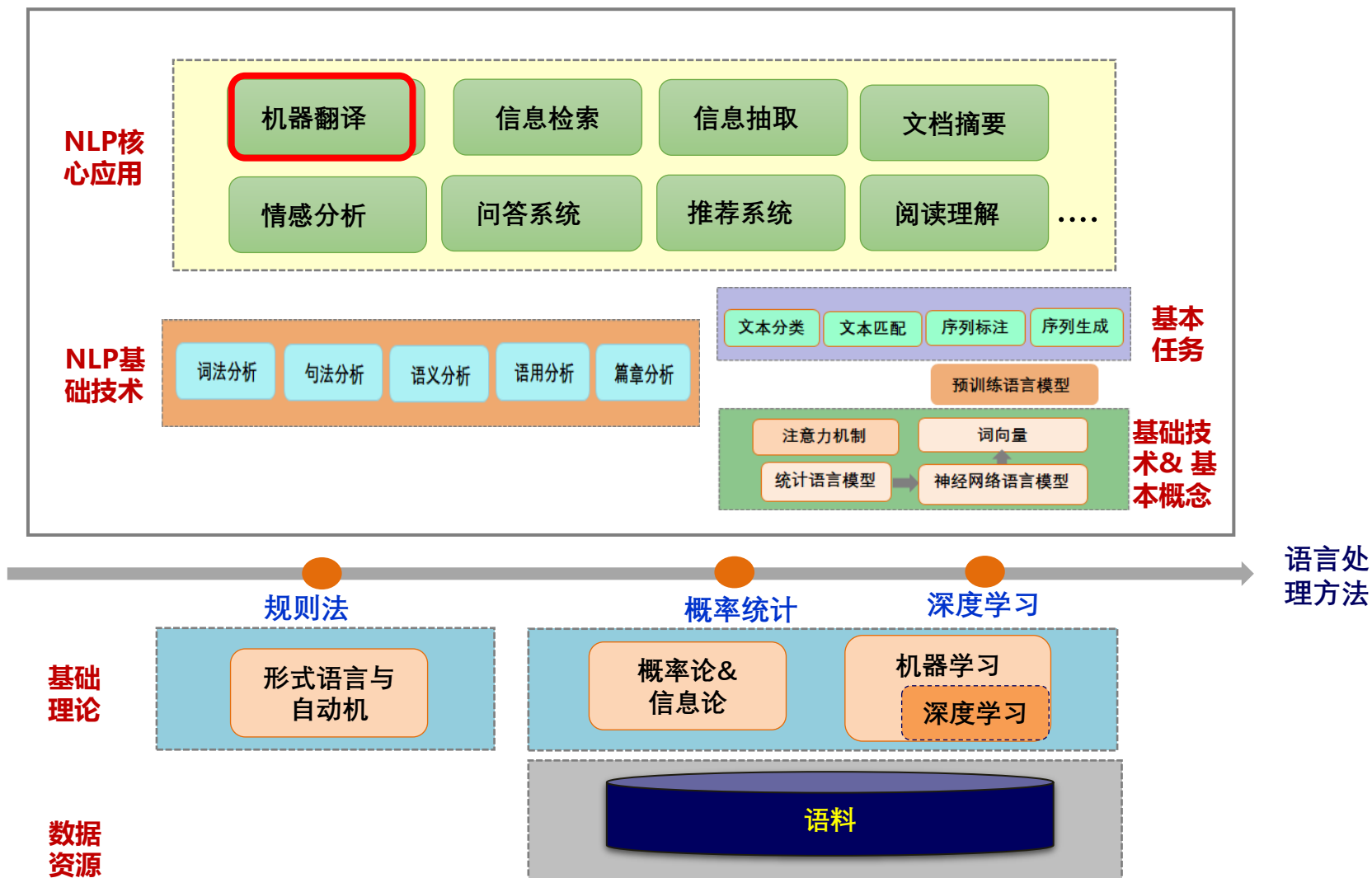


第 17 章 附录：统计机器翻译

授课教师：胡玥

授课时间：2020.12

自然语言处理体系架构



统计机器翻译概述

统计机器翻译的诞生

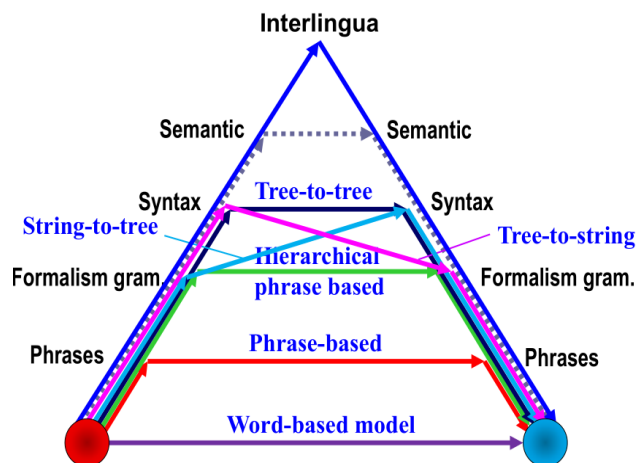
1946年 美国洛克菲勒基金会 (Rockefeller Foundation) 副总裁 W. Weaver提出机器翻译的想法； 1949年，韦弗发表了一份以《翻译》为题的备忘录，提出：“当我阅读一篇用汉语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，在阅读时，我是在进行解码”。韦弗的卓越思想成为了而后统计机器翻译 (Statistic Machine Translation,简称SMT) 的理论基础。



1990 年IBM 的Peter F. Brown 等人在Computational Linguistics 上发表论文 “统计机器翻译方法”； 1993 年 他们发表在该 杂志发表论文 “统计机器翻译的数学：参数估计” 两篇文章奠定了统计机器翻译的理论基础。

统计机器翻译概述

统计机器翻译的方法大体可以划分为三类



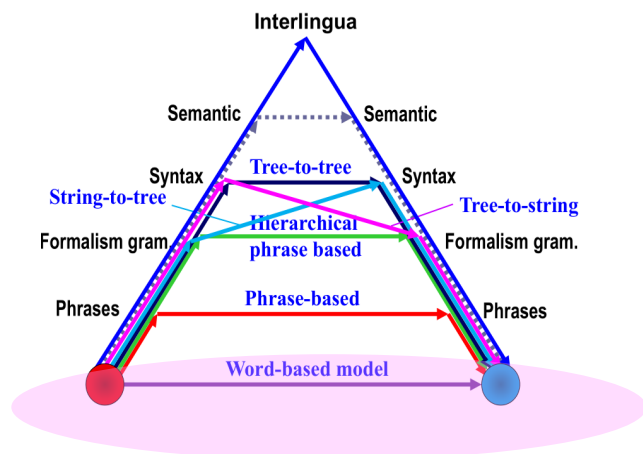
- 基于词的方法
- 基于短语的方法
- 基于句法方法
 - 基于层次化短语方法（形式句法）
 - 基于树的方法（语义句法）

内 容 提 要

- 附录. 1 基于词的统计机器翻译方法
- 附录. 2 基于短语的统计机器翻译方法
- 附录. 3 基于层次化短语的统计机器翻译方法
- 附录. 4 基于树的统计机器翻译方法
- 附录. 5 译文评估方法

附录.1 基于词的统计机器翻译方法

基于词的统计翻译模型:



翻译单位： 词

翻译模型： 噪声信道模型

代表模型： IBM 统计翻译模型 (IBM 1-5)

最早模型，性能已经被其他方法所超越，仍然是现有其他是各种方法的基础，特别是其中建立在各种词对齐思想上的词语对齐工具例如(GIZA++), 仍然被所有的其他统计机器翻译方法所采用。

附录.1 基于词的统计机器翻译方法

□ 噪声信道模型

核心思想：

一种语言 T 由于经过一个噪声信道而发生变形，从而在信道的另一端呈现为另一种语言 S (信道意义上的输出，翻译意义上的源语言)。翻译问题实际上就是如何根据观察到的 S ，恢复最为可能的 T 问题。这种观点认为，任何一种语言的任何一个句子都有可能是另外一种语言中的某个句子的译文，只是可能有大有小[Brown *et. al*, 1990]。



$$t^* = \operatorname{argmax}_t \Pr(t|s)$$

附录.1 基于词的统计机器翻译方法

问题： 源语言句子： $S=s_1^m=s_1s_2\cdots s_m$

目标语言句子： $T=t_1^l=t_1t_2\cdots t_l$

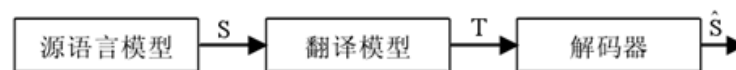
目标： 求 $t^* = \operatorname{argmax}_t \Pr(t|s)$

解： 贝叶斯公式： $P(T|S) = \frac{P(T) \times P(S|T)}{P(S)}$

$$T' = \operatorname{argmax}_T \underbrace{P(T)}_{\text{语言模型}} \times \underbrace{P(S|T)}_{\text{翻译模型}}$$

统计翻译中的三个关键问题：

1. 估计语言模型概率 $p(T)$;
2. 估计翻译概率 $p(S|T)$;
3. 快速有效地搜索 T 使得 $p(T) \times p(S | T)$ 最大。



统计机器翻译的信源信道模型

附录.1 基于词的统计机器翻译方法

1. 估计语言模型概率 $p(T)$

$$T' = \operatorname{argmax}_T P(T) \times P(S|T)$$

语言模型

□ n -gram

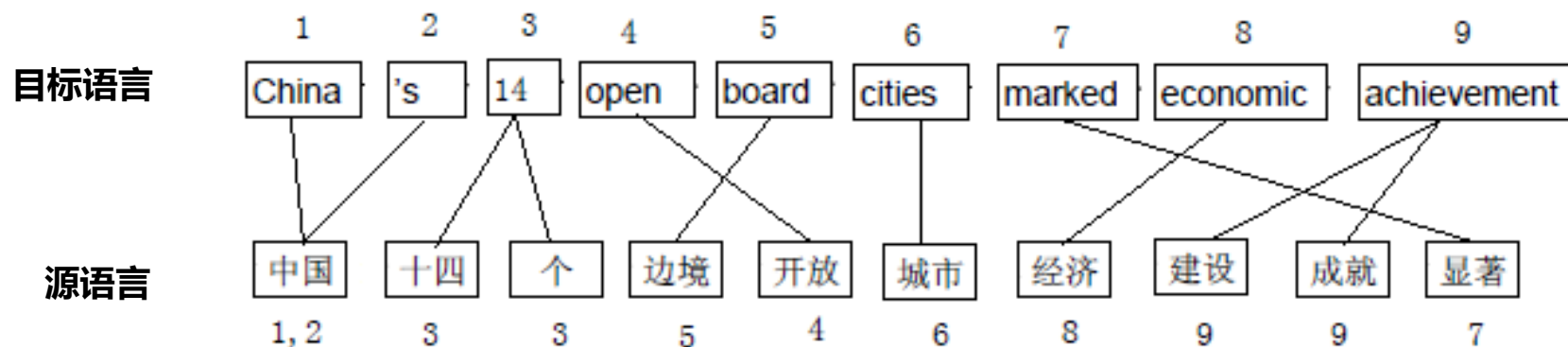
附录.1 基于词的统计机器翻译方法

2. 估计翻译概率 $p(S|T)$

$$T' = \operatorname{argmax}_T P(T) \times P(S|T)$$

翻译模型

翻译时词之间的对应情况



目标语言→源语言 问题:

- 一对多问题
- 源语言序列词序问题

如何定义词对齐关系?

附录.1 基于词的统计机器翻译方法

IBM基于词的统计翻译模型 (IBM1-5)

模型	假设	参数训练	简评
IBM1	翻译模型仅与单词间的直译概率有关，句长概率和对齐概率都是均匀分布。	应用EM算法，从双语语料库中训练获得，可以得到全局最优参数，与初始值无关。	模型简单、易于实现，但仅考虑了单词的影响，没有考虑词序的影响。
IBM2	翻译模型和句长模型同IBM1，对位概率为0阶对齐。	应用EM算法，从双语语料库中训练获得，只能收敛到局部最优。	模型简单、易于实现，同时考虑了单词和词序的影响。
IBM3	翻译模型依赖于繁衍率模型和单词间的直译概率，对齐概率取0阶词对齐。	需要首先应用模型IBM1或IBM2对双语语料进行单词级对位，然后训练繁衍概率参数。	引入了描述单词间一对多情况的繁衍概率，参数较多，实现过程较复杂。
IBM4	翻译模型依赖于单词间的直译概率、繁衍概率、词类、语言片断中心位置和语言片断内相对位置等因素，对齐概率取1阶词对齐。	需要首先应用模型IBM1~IBM3对双语语料进行单词级对位和语言片断划分，然后训练两种位置概率参数。	不仅考虑了一对多的情况，还将语言片断作为一个整体进行考虑。参数较多、不易实现。
IBM5	翻译模型依赖于直译概率、繁衍概率、语言片断中心位置、语言片断内相对位置和对位的历史等因素。	需要在模型IBM1~ IBM4参数训练的基础上获得参数。	对IBM4进行了修正，同时考虑了当前对位信息和对位历史。模型的表现力最强，但过于复杂，实用性不强。
HMM	句长模型和翻译模型同IBM1，对齐模型为1阶对齐。	应用EM算法，从双语对照语料中训练获得。	模型简单，易于实现，考虑了词序的影响。

Och 基于这些模型开发并发布了目前被广泛使用的无指导词对齐开源工具GIZA++

附录.1 基于词的统计机器翻译方法

3. 快速有效地搜索T 使得 $p(T) \times p(S | T)$ 最大

给定S, 求T, 使得 $P(T) * P(S|T)$ 最大

经典的算法:


- 贪婪算法
- 堆栈搜索
-

附录.1 基于词的统计机器翻译方法

基于词翻译例：

源语： 澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

翻译过程： 澳洲 是 与 北韩 有 邦交 的 少数 国家 之一



Australia is one of the few countries that have diplomatic relations with North Korea

目标语： Australia is one of the few countries that have diplomatic relations with North Korea

内 容 提 要

- 附录. 1 基于词的统计机器翻译方法
- 附录. 2 基于短语的统计机器翻译方法
- 附录. 3 基于层次化短语的统计机器翻译方法
- 附录. 4 基于树的统计机器翻译方法
- 附录. 5 译文评估方法

附录.2 基于短语的统计机器翻译方法

基于词的翻译模型存在的问题

基于词的翻译模型只刻画了词到词的翻译概率，词翻译的时候没有考虑上下文，在词语调序方面能力很差。

- 难以刻画一些固定搭配、习惯用法的翻译；
- 很难处理词义消歧问题
- 很难处理一对多、多对一和多对多的翻译问题

很多研究者想到了在短语层面进行建模，以改进局部词语调序的效果。最成功的工作是 Och、Zens、Koehn 等人的工作。

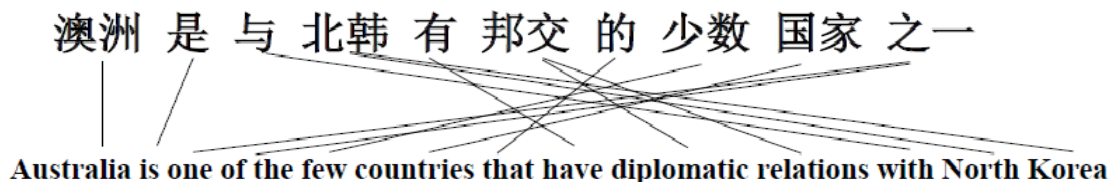
[Koehn, 2003] 提出基于短语的对数线性模型翻译模型

附录.2 基于短语的统计机器翻译方法

如： **基于词翻译**

源语： 澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

目标语： Australia is one of the few countries that have diplomatic relations with North Korea



基于短语翻译

源语：

短语划分

澳洲	是	与	北韩	有	邦交	的	少数	国家	之一
----	---	---	----	---	----	---	----	----	----

短语翻译

Australia is	with North Korea	have diplomatic relations	one of the few countries that
--------------	------------------	---------------------------	-------------------------------

目标语：

短语调序

Australia is	one of the few countries that	have diplomatic relations	with North Korea
--------------	-------------------------------	---------------------------	------------------

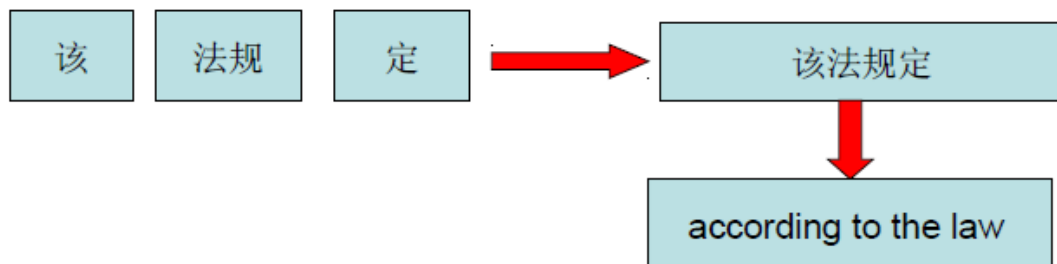
对齐的短语：

(澳洲 是, Australia is)
(与 北韩, with North Korea)
(有 邦交, have the diplomatic relations)
(的 少数 国家 之一, one of the few countries that)

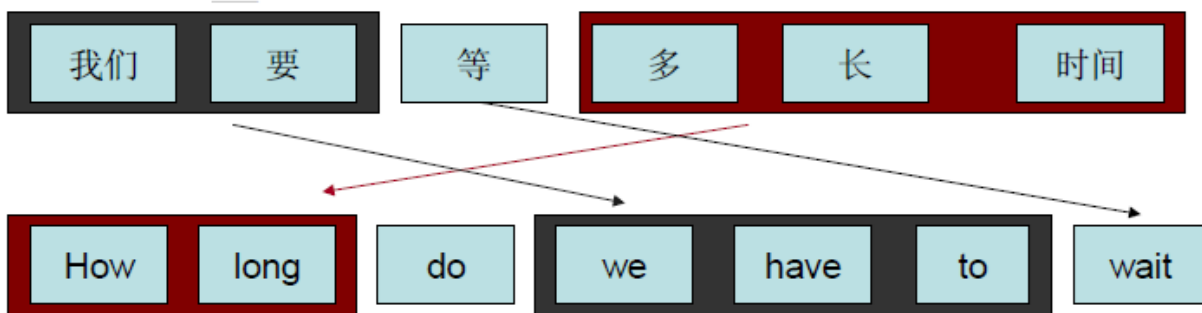
附录.2 基于短语的统计机器翻译方法

基于短语建立翻译模型的优点

分词错误

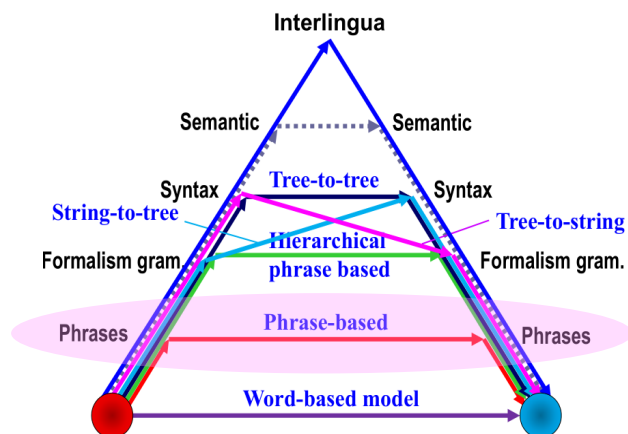


词序调整



附录.2 基于短语的统计机器翻译方法

基于短语的统计翻译模型:



翻译单位： 短语

翻译模型： 对数线性模型

代表模型：

基于短语的对数线性翻译模型[Koehn, 2003]

关于模型问题，原始论文的提法是“最大熵”模型，现在通常使用“对数线性（Log-Linear）模型”这个概念。“对数线性模型”的含义比“最大熵模型”更宽泛，而且现在这个模型通常都不再使用最大熵的方法进行参数训练，因此“对数线性”模型的提法更为准确。

附录.2 基于短语的统计机器翻译方法

□ 短语:

指一个连续的词串(*n-gram*), 不一定是语言学中定义的短语(*phrase*)

如: 我想预订一个单人间。

I would like to reserve a single room.

在基于短语的模型中, 直接将繁衍率信息、上下文信息以及局部对位调序信息记录在短语翻译规则中

附录.2 基于短语的统计机器翻译方法

□ 对数线性模型

核心思想:

假设 T 、 S 是机器翻译的目标语言和源语言句子, $h_1(T, S), \dots, h_M(T, S)$ 分别是 T, S 上的 M 个特征, $\lambda_1, \dots, \lambda_M$ 是与这些特征分别对应的 M 个参数 (权值), 利用对数线性模型, 给定源语言句子 S , 其最佳译文 T 可以用以下公式求得:

$$\begin{aligned} T' = \operatorname{argmax}_T P(T|S) &= \operatorname{argmax}_T \frac{\exp\{\sum_1^M \lambda_m h_m(T, S)\}}{\sum_{T^*} \exp\{\sum_1^M \lambda_m h_m(T^*, S)\}} \\ &= \operatorname{argmax}_T \left\{ \sum_1^M \lambda_m h_m(T, S) \right\} \end{aligned}$$

对数线性模型

附录.2 基于短语的统计机器翻译方法

※ 目标语言T和源语言S 上的特征: $h_1(T, S), \dots, h_M(T, S)$

基于短语的统计翻译过程

原文: 他 将 于 4 月 10日 访问 美国

短语划分: 他 将 于 4月 10日 访问 美国 **短语划分模型** $P(S_1^K | S)$

翻译: He will on April 10 visit America **短语翻译模型** $P(T_1^K | S_1^K, S)$

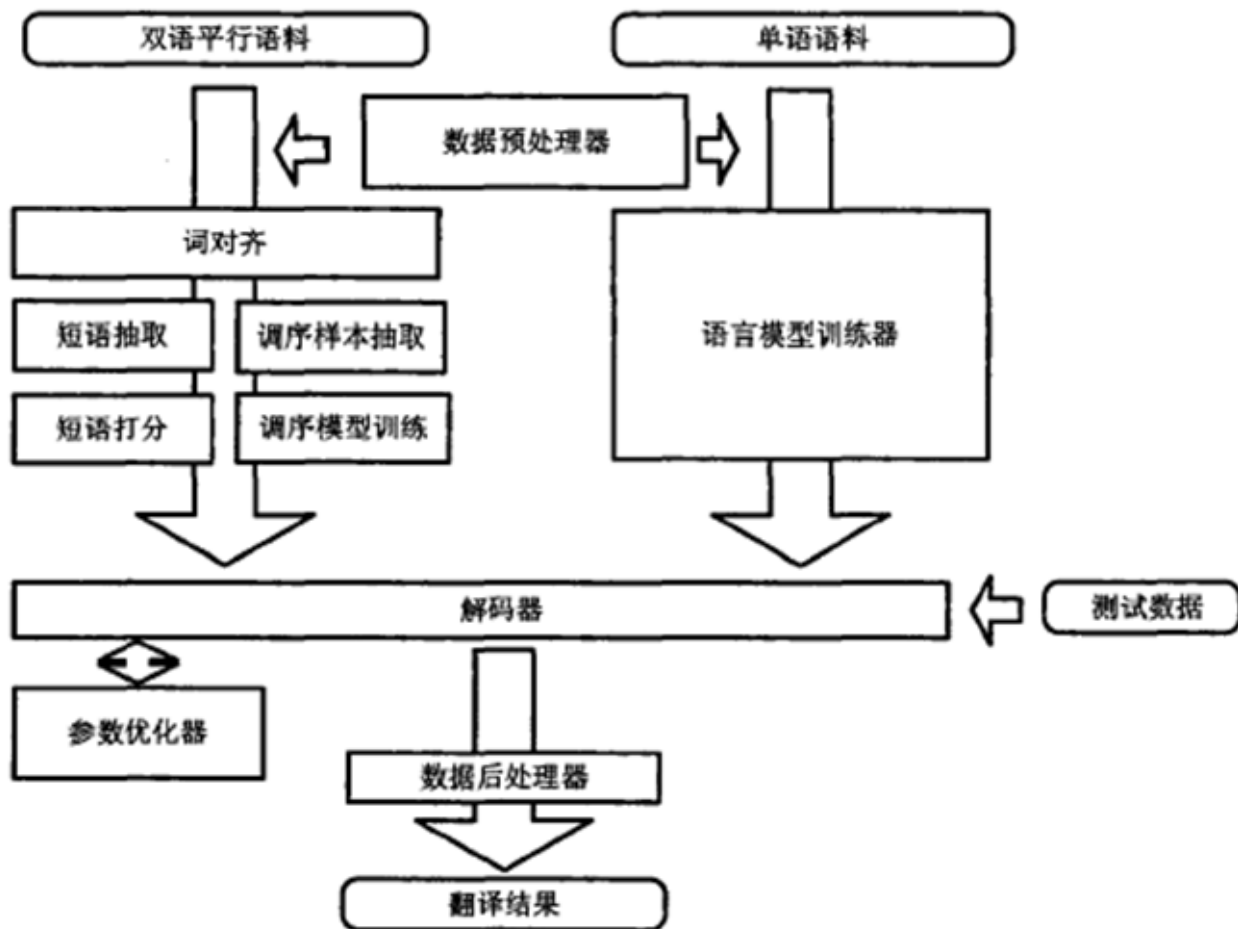
短语调序: He will visit America on April 10 **短语调序模型** $P(T_1^{K'} | T_1^K, S_1^K, S)$

目标语言模型 $P(T | T_1^{K'}, T_1^K, S_1^K, S)$

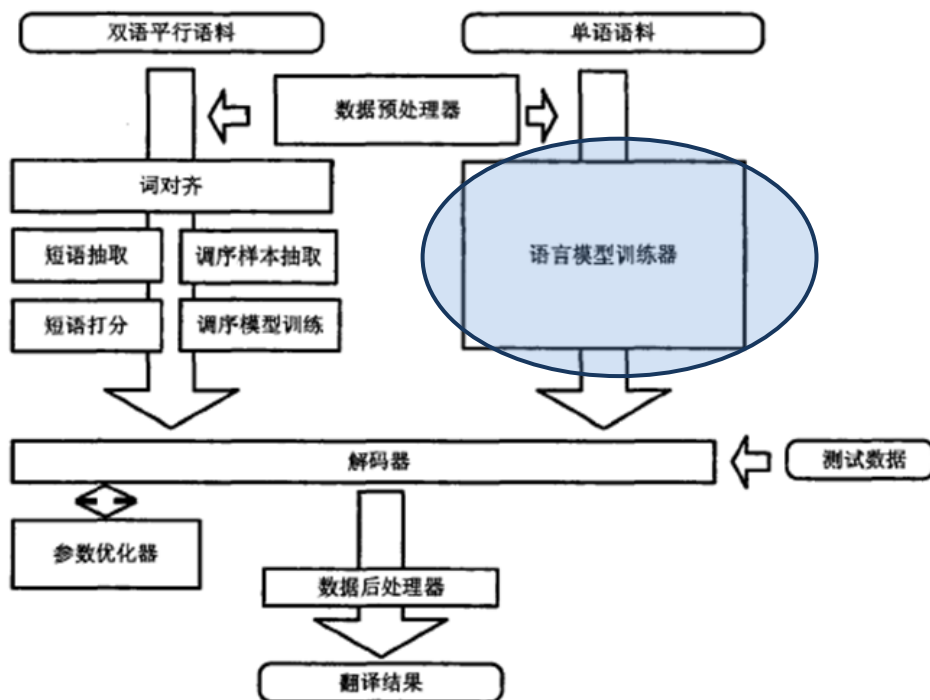
$$T' = \operatorname{argmax}_T P(T | S) = \operatorname{argmax}_T \left\{ \sum_1^M \lambda_m h_m(T, S) \right\} \quad \text{线性对数模型}$$

附录.2 基于短语的统计机器翻译方法

基于短语的对数线性模型翻译系统架构



附录.2 基于短语的统计机器翻译方法



基于短语的对数线性模型翻译系统架构

1. 目标语言模型

给定句子: $T = t_1^l = t_1 t_2 \dots t_l$

句子概率: $P(T) =$

$P(t_1) P(t_2 | t_1) \dots P(t_l | t_1 t_2 \dots t_{l-1})$

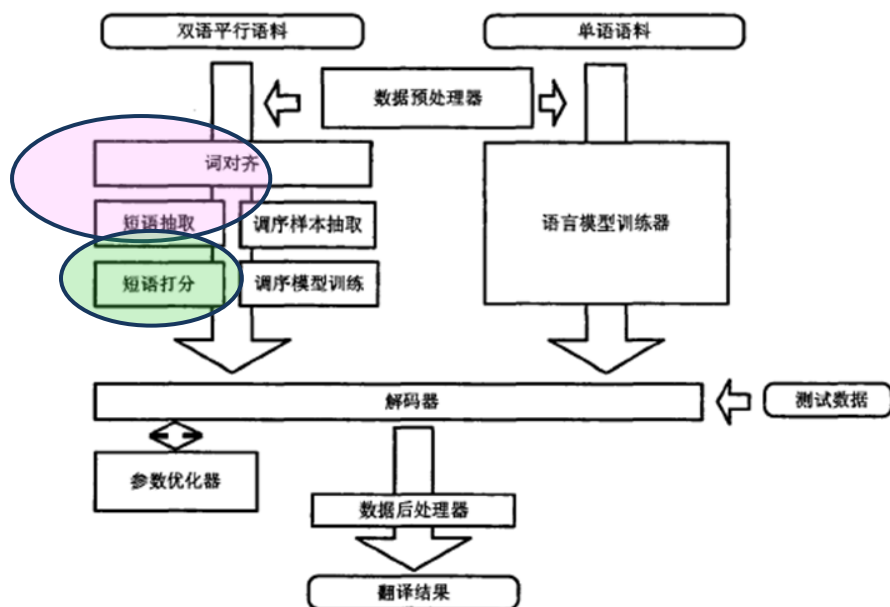
目标语言模型学习问题

n -gram

神经网络语言模型

.....

附录.2 基于短语的统计机器翻译方法



基于短语的对数线性模型翻译系统架构

2. 短语划分&翻译模型

以短语为基本翻译单元 把训练语料库中所有对齐的短语及其翻译概率存储起来, 作为一部带概率的短语词典, 翻译时选择最好的短语进行翻译

- 中国 经济 发展 十分 迅速

```
<src phrase=" 中国" beg="0" end="0">
<trans lex="China">-0.282683 | -0.242107 | -0.995656 | -0.391329 | -1 | </trans>
<trans lex="China 's">-0.331812 | -0.69002 | -1.76154 | -1.87886 | -2 | </trans>
</src>
<src phrase=" 中国经济" beg="0" end="1">
<trans lex="China 's economic">-0.344841 | -0.916413 | -0.904456 | -2.43623 | -3 | </trans>
<trans lex="China 's economy">-0.241162 | -1.04799 | -1.33977 | -2.98078 | -3 | </trans>
</src>
<src phrase=" 中国经济发展" beg="0" end="2">
<trans lex="China 's economic development">0 | -1.34465 | -0.606135 | -2.8362 | -4 | </trans>
<trans lex="China 's economic development .">0 | -1.34465 | -1.70475 | -6.01209 | -5 | </trans>
</src>
<src phrase=" 经济" beg="1" end="1">
<trans lex="economic">-0.121522 | -0.226393 | -0.943174 | -0.557362 | -1 | </trans>
<trans lex="economy">-0.0626278 | -0.357969 | -1.45799 | -1.10192 | -1 | </trans>
</src>
.....
```

语翻译表构造步骤:

- (1) 词对齐
- (2) 短语抽取
- (3) 短语打分

附录.2 基于短语的统计机器翻译方法

(1) 词对齐

词对齐大多数统计机器翻译系统中必不可少的一个模块。许多规则抽取模块都依赖于词对齐信息。

输入：是双语平行句对

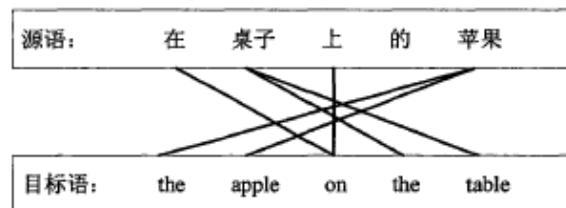
如：

源语：	在	桌子	上	的	苹果
-----	---	----	---	---	----

目标语：	the	apple	on	the	table
------	-----	-------	----	-----	-------

双语平行句对

输出：是双语平行句对中的词对齐信息。



包含词对齐信息的双语平行句对

附录.2 基于短语的统计机器翻译方法

词对齐步骤:

一般词对齐模块只能得到一端到另一端的一个映射关系，做对齐时需要从源语到目标语和从目标语到源语做两次词对齐，并将两者结合得到**最终的词对齐信息**。

如:

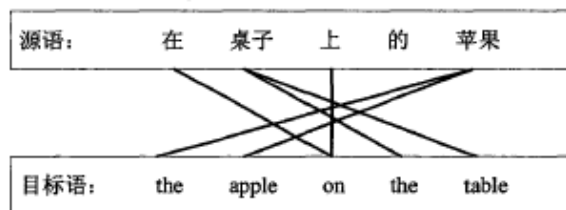
输入

源语: 在 桌子 上 的 苹果

目标语: the apple on the table

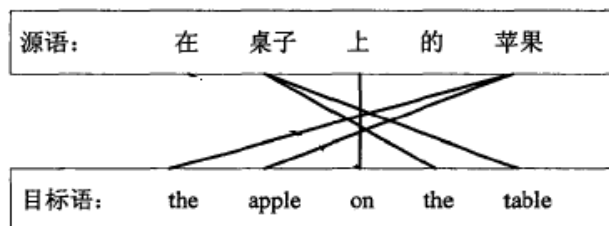
双语平行句对

对齐结果

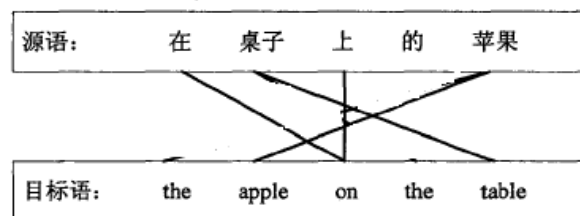


包含词对齐信息的双语平行句对

目标→源语



源语→目标



对齐工具 GIZA++

附录.2 基于短语的统计机器翻译方法

(2) 短语抽取

从包含词对齐信息的双语平行语料中抽取短语 (Koehn等人) :
任意长度的双语短语翻译对只要满足词对齐一致性原则, 就是可抽取的短语翻译对。

词对齐一致性原则:

假定短语翻译对记为 (\bar{s}, \bar{t}) , 词对齐记为A, 如果短语翻译对 (\bar{s}, \bar{t}) 与词对齐A一致, 当且仅当它们满足如下关系式:

$$\begin{aligned} & \forall s_i \in \bar{s} : (s_i, t_j) \in A \rightarrow t_j \in \bar{t} \\ \text{AND } & \forall t_j \in \bar{t} : (s_i, t_j) \in A \rightarrow s_i \in \bar{s} \\ \text{AND } & \exists s_i \in \bar{s}, t_j \in \bar{t} : (s_i, t_j) \in A \end{aligned}$$

其中 s_i 和 t_j , 和分别是源语言短语和目标语言短语中的一个词。

附录.2 基于短语的统计机器翻译方法

短语抽取算法:

短语抽取算法

输入: 源语言句子 $s = s_1 \dots s_J$, 目标语句子 $t = t_1 \dots t_I$, 词对齐关系 A

输出: 所有与词对齐关系满足一致性原则的短语翻译对

```
1:  Function ExtractAllPhrases( $s, t, A$ )
2:    for  $j_1 = 1$  to  $J$                                 ▷ 源语言短语开始位置
3:      for  $j_2 = j_1$  to  $j_1 + l_{s_{\max}} - 1$           ▷ 源语言短语结束位置
4:        for  $i_1 = 1$  to  $I$                             ▷ 目标语短语开始位置
5:          for  $i_2 = i_1$  to  $i_1 + l_{t_{\max}} - 1$         ▷ 目标语短语结束位置
6:            if IsValid( $j_1, j_2, i_1, i_2, A$ ) then
7:              add Phrase( $j_1, j_2, i_1, i_2$ ) into PhraseList
8:    return PhraseList
9:  Function IsValid( $j_1, j_2, i_1, i_2, A$ )
10:    for  $j = j_1$  to  $j_2$ 
11:      if  $\exists i' \notin (i_1, i_2): A[j, i'] = 1$  then    ▷ 如果一个源语言词对到目标语短语外
12:        return false
13:    for  $i = i_1$  to  $i_2$ 
14:      if  $\exists j' \notin (j_1, j_2): A[j', i] = 1$  then    ▷ 如果一个目标语词对到源语言短语外
15:        return false
16:    return true
```

其中, $l_{s_{\max}}$ 和 $l_{t_{\max}}$ 和分别是源语言短语和目标语言短语的最大长度限制。理论上分别可以是源语言句子和目标语言句子的长度, 但在实际汉英翻译任务中一般设成 3和5

附录.2 基于短语的统计机器翻译方法

例1:

	beijing	housing	prices	continued	to	rise
北京	●					
房价		●	●			
持续				●		
上涨						●

一致性短语

北京			beijing
持续			continued
上涨			rise
房价			housing prices
持续			continued to
上涨			to rise
北京 房价			beijing housing prices
房价 持续			housing prices continued
持续 上涨			continued to rise
北京 房价 持续			beijing housing prices continued
房价 持续			housing prices continued to
北京 房价 持续			beijing housing prices continued to
房价 持续 上涨			housing prices continued to rise
北京 房价 持续 上涨			beijing housing prices continued to rise

附录.2 基于短语的统计机器翻译方法

(3) 短语打分

在使抽取算法获得短语翻译对集合后，还需要根据特征函数对集合中的短语翻译对作概率估计，得到相应的特征函数值。

常用的特征函数。这包括**正向、反向短语翻译概率**，**正向、反向词汇化短语翻译概率**，**目标短语词数特征**以及**目标短语数特征**。

附录.2 基于短语的统计机器翻译方法

➤ 正向短语翻译概率

给定某个源语言短语翻译到某个目标语短语的概率。这个概率可以使用如下公式计算得到。

$$\Pr(\bar{t} | \bar{s}) = \text{count}(\bar{s}, \bar{t}) / \sum_{\bar{t}_j} \text{count}(\bar{s}, \bar{t}_j)$$

其中, $\text{count}(\bar{s}, \bar{t})$, 统计了源语言短语 \bar{s} 与目标语言短语 \bar{t} 在语料中互为短语翻译对时出现的次数, $\sum_{\bar{t}_j} \text{count}(\bar{s}, \bar{t}_j)$ 统计了源语言短语与所有可与之互为短语翻译对的目标语言短语的总计数 (最大似然估计)

➤ 反向短语翻译概率

概率的估计过程类似正向短语翻译概率的估计过程

计算公式:

$$\Pr(\bar{s} | \bar{t}) = \text{count}(\bar{s}, \bar{t}) / \sum_{\bar{s}_i} \text{count}(\bar{s}_i, \bar{t})$$

分母针对源语言端作一个计数统计。

如:

桌子	↔	the table
的	↔	<NULL>
苹果	↔	the apple
的 苹果	↔	the apple
在 桌子 上	↔	on the table
在 桌子 上 的	↔	on the table
在 桌子 上 的 苹果	↔	the apple on the table

附录.2 基于短语的统计机器翻译方法

➤ 正向词汇化翻译概率

本质上是一个对短语翻译概率的平滑方法

计算公式：

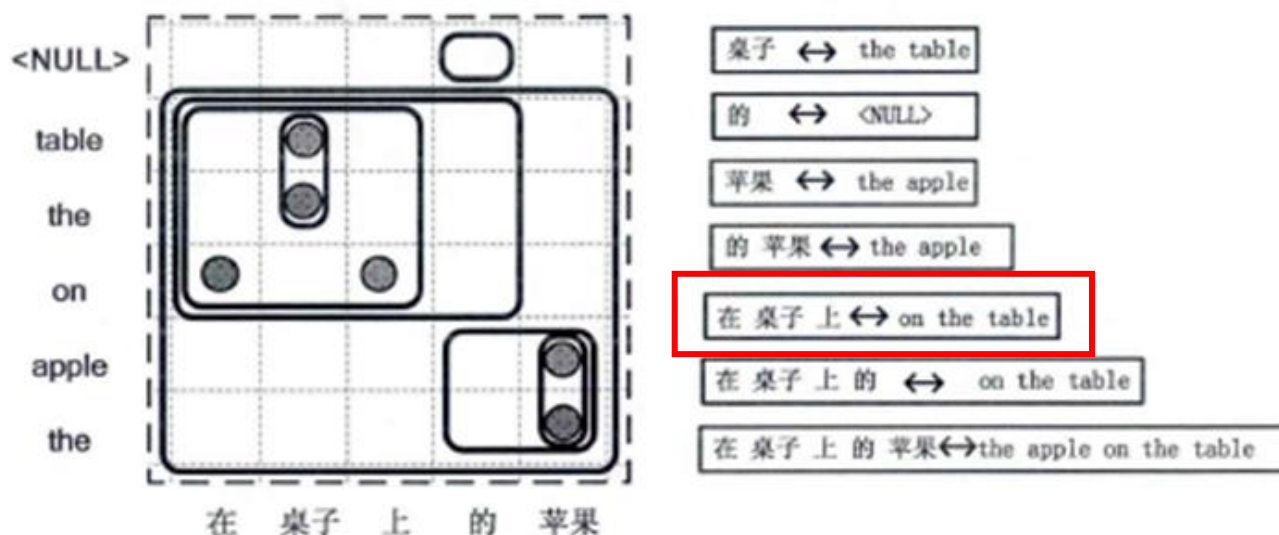
$$lex(\bar{t} | \bar{s}, A) = \prod_{i=1}^{length(\bar{t})} \frac{1}{|\{j | (i, j) \in A\}|} \sum_{(i,j) \in A} w(t_i | s_j)$$

其中， \bar{s} 和 \bar{t} 分别表示源语言短语与目标语言短语， A 表示 \bar{s} 和 \bar{t} 的词对齐关系， s_j 和 t_i 分别表示 \bar{s} 和 \bar{t} 中的一个词。

计算公式含义：对每个目标语言短语中的词计算得到任意一个源语言短语中的词翻译到它的概率的算术平均值，并对这些算术平均值累乘

附录.2 基于短语的统计机器翻译方法

如：



$$lex(\bar{t}|\bar{s}, A) = \frac{1}{2} \{w(on|在) + w(on|上)\} \times w(the|桌子) \times w(table|桌子)$$

如果目标语言短语中的某个词是空对词，那么源语言短语中的词可以包含译空符号 "<NULL>"

附录.2 基于短语的统计机器翻译方法

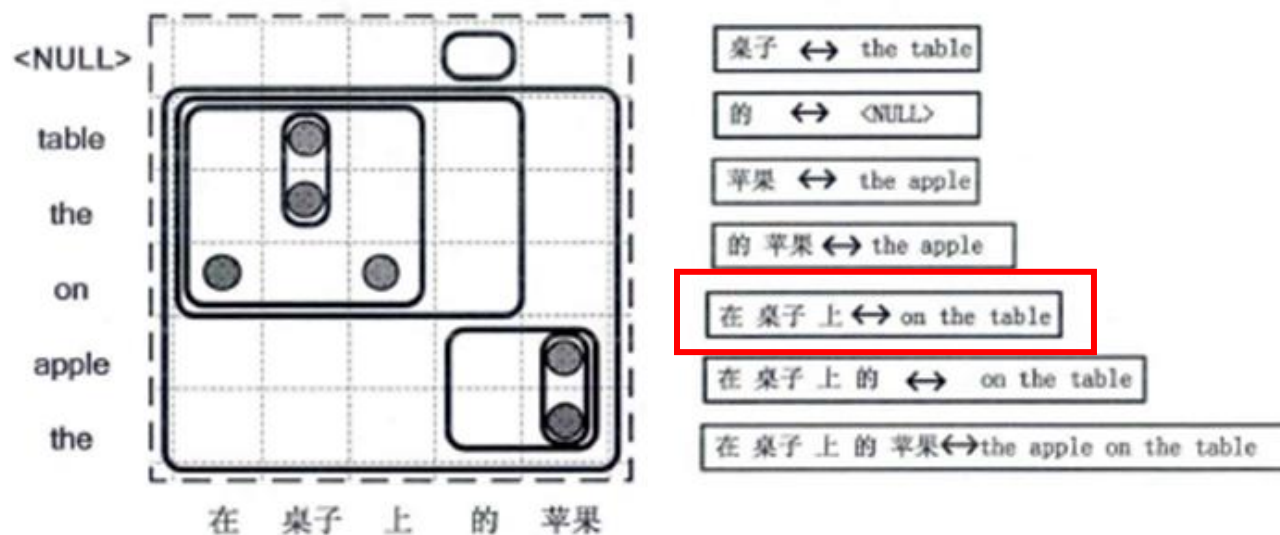
➤ 反向词汇化翻译概率

与正向词汇化翻译概率的估计方法类似

计算公式:

$$\text{lex}(\bar{s}|\bar{t}, A) = \prod_{j=1}^{\text{length}(\bar{s})} \frac{1}{|\{i | (i, j) \in A\}|} \sum_{(i, j) \in A} w(s_j | t_i)$$

如:



$$\text{lex}(\bar{s}|\bar{t}, A) = w(\text{在}|\text{on}) \times \frac{1}{2} \{ w(\text{桌子}|\text{the}) + w(\text{桌子}|\text{table}) \} \times w(\text{上}|\text{on})$$

附录.2 基于短语的统计机器翻译方法

构造的短语翻译表结果:

- 中国 经济 发展 十分 迅速

```
<src phrase=" 中国" beg="0" end="0">
```

```
<trans lex="China">-0.282683 | -0.242107 | -0.995656 | -0.391329 | -1 | </trans>
```

```
<trans lex="China 's">-0.331812 | -0.69002 | -1.76154 | -1.87886 | -2 | </trans>
```

```
</src>
```

```
<src phrase=" 中国经济" beg="0" end="1">
```

```
<trans lex="China 's economic">-0.344841 | -0.916413 | -0.904456 | -2.43623 | -3 | </trans>
```

```
<trans lex="China 's economy">-0.241162 | -1.04799 | -1.33977 | -2.98078 | -3 | </trans>
```

```
</src>
```

```
<src phrase=" 中国经济发展" beg="0" end="2">
```

```
<trans lex="China 's economic development">0 | -1.34465 | -0.606135 | -2.8362 | -4 | </trans>
```

```
<trans lex="China 's economic development .">0 | -1.34465 | -1.70475 | -6.01209 | -5 | </trans>
```

```
</src>
```

```
<src phrase=" 经济" beg="1" end="1">
```

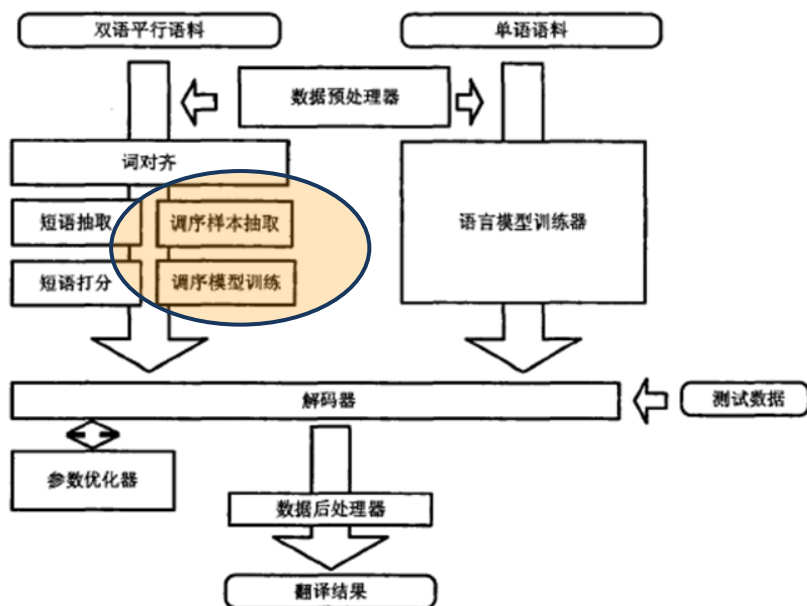
```
<trans lex="economic">-0.121522 | -0.226393 | -0.943174 | -0.557362 | -1 | </trans>
```

```
<trans lex="economy">-0.0626278 | -0.357969 | -1.45799 | -1.10192 | -1 | </trans>
```

```
</src>
```

```
.....
```

附录.2 基于短语的统计机器翻译方法



基于短语的对数线性模型翻译系统架构

3. 短语调序模型

调序模型是基于短语的统计机器翻译模型中非常重要的一个模块,用于刻画如何组合源语言短语译文,以得到一个合理的、通顺的翻译结果的问题。

基于短语的机器翻译方法中最常用的调序模型:

- (1) 距离跳转模型
- (2) 分类模型 (MSD模型)

附录.2 基于短语的统计机器翻译方法

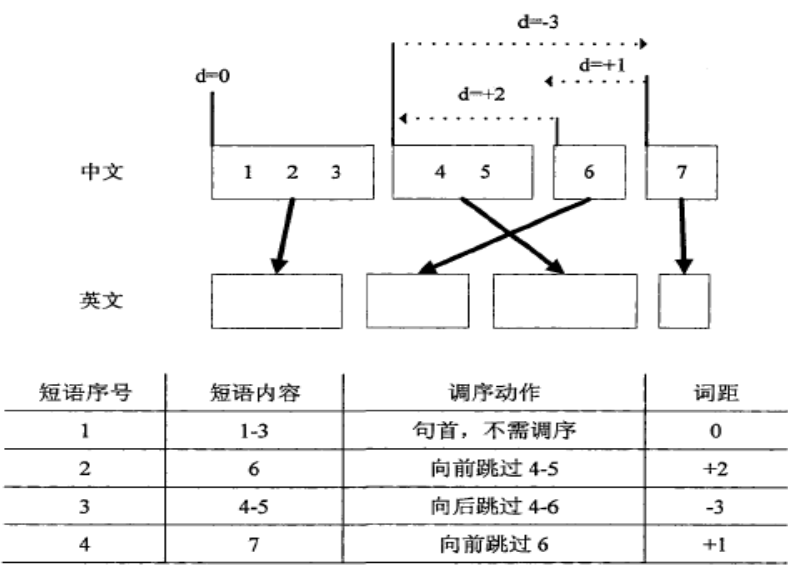
(1) 距离跳转模型

主要考虑在翻译过程中在源语言端所要跳过的词的个数这一代价, 该代价定义为**调序词距**。

调序词距 = $start_i - end_{i-1} - 1$

其中, $start_i$ 为翻译到第*i*个目标短语的源语言短语的第一个词的位置, end_i 为源语最后一个词的位置。

如:



调序概率 $d(x) = \alpha^{|x|}$

其中, α 是 0~1 常数 (人工设或训练得)

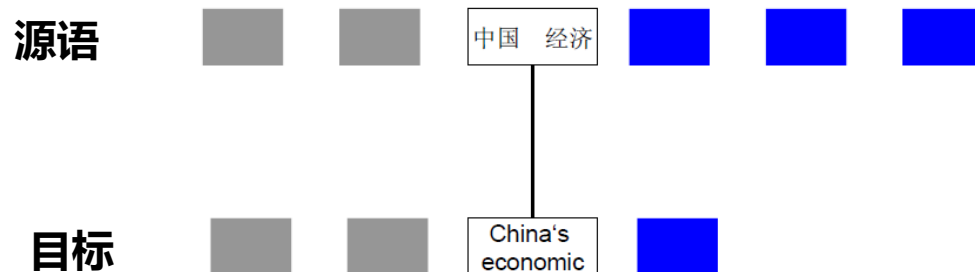
该分布函数的意义: 使得长距离调序的代价比较高, 不支持翻译系统作过多调序动作。该模型常适用于语序差异不大的语言对之间的翻译任务.

附录.2 基于短语的统计机器翻译方法

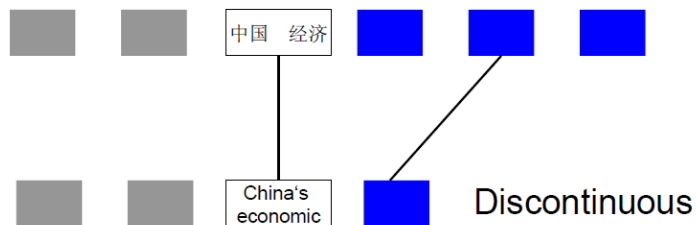
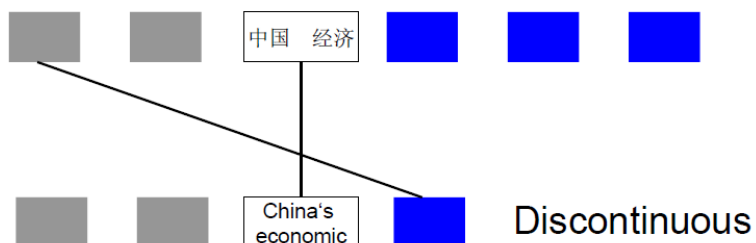
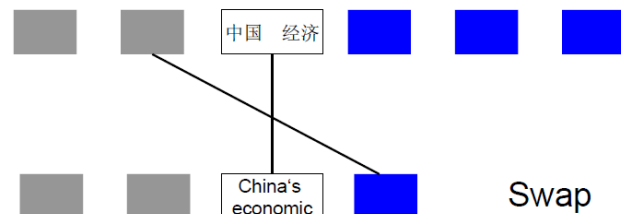
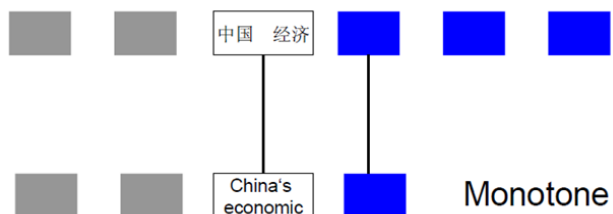
(2) 分类模型 (MSD模型)

对于任意一个源语言短语与其对应的目标语言短语所构成的短语对，根据它与前一个短语对的关系，MSD调序模型定义了三种调序类型，分别为：顺序 (M)、交换 (S) 和非连续 (D)。

如：统计目标端



附录.2 基于短语的统计机器翻译方法



对每一对双语短语，在训练语料库中统计目标端相邻的下一个短语在源语言端是Monotone、Swap、和Discontinuous的概率

附录.2 基于短语的统计机器翻译方法

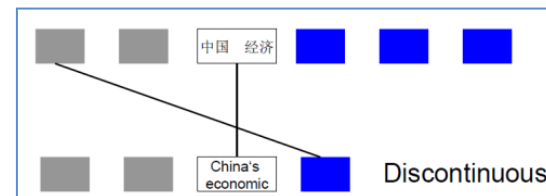
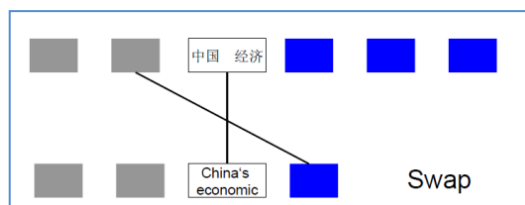
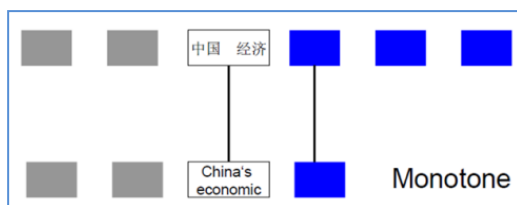
计算MSD调序模型的概率值 (以目标语端连续这为例)

假设由源语言短语序列构成的源语言句子为 $s = \bar{s}_1 \dots \bar{s}_K$ ，由目标语言短语序列构成的目标语言句子为 $t = \bar{t}_1 \dots \bar{t}_K$ ，并且 $a = \bar{a}_1 \dots \bar{a}_K$ 是 s 和 t 的短语对齐关系，表示 t 中的第 i 个短语 \bar{t}_i 与 s 中的第 \bar{a}_i 个短语 $\bar{s}_{\bar{a}_i}$ 对应，MSD调序模型的概率如下公式定义

$$\Pr(o \mid s, t, a) = \prod_{i=1}^K \Pr(o_i \mid \bar{s}_{\bar{a}_i}, \bar{t}_i, \bar{a}_{i-1}, \bar{a}_i)$$

其中 o_i 是集合 $O = \{M, S, D\}$ 中的一个值，并由条件 \bar{a}_{i-1} 和 \bar{a}_i 确定。

如果 $\bar{a}_i - \bar{a}_{i-1} = 1$ ，那么 $o_i = M$ ；如果 $\bar{a}_i - \bar{a}_{i-1} = -1$ ，那么 $o_i = S$ ；否则 $o_i = D$ 。



概率 $\Pr(o_i \mid \bar{s}_{\bar{a}_i}, \bar{t}_i, \bar{a}_{i-1}, \bar{a}_i)$ 是从训练语料中自动估计出来的

附录.2 基于短语的统计机器翻译方法

将概率 $\Pr(o_i | \bar{s}_{\bar{a}_i}, \bar{t}_i, \bar{a}_{i-1}, \bar{a}_i)$ 简记为 $\Pr(o | \bar{s}, \bar{t})$

$$\Pr(o | \bar{s}, \bar{t}) = \frac{\text{count}(o, \bar{s}, \bar{t})}{\sum_o \text{count}(o, \bar{s}, \bar{t})}$$

对应每种调序类型

$$f_{M-pre}(d) = \prod_{i=1}^K \Pr(o_i = M | \bar{s}_{\bar{a}_i}, \bar{t}_i, \bar{a}_{i-1}, \bar{a}_i)$$

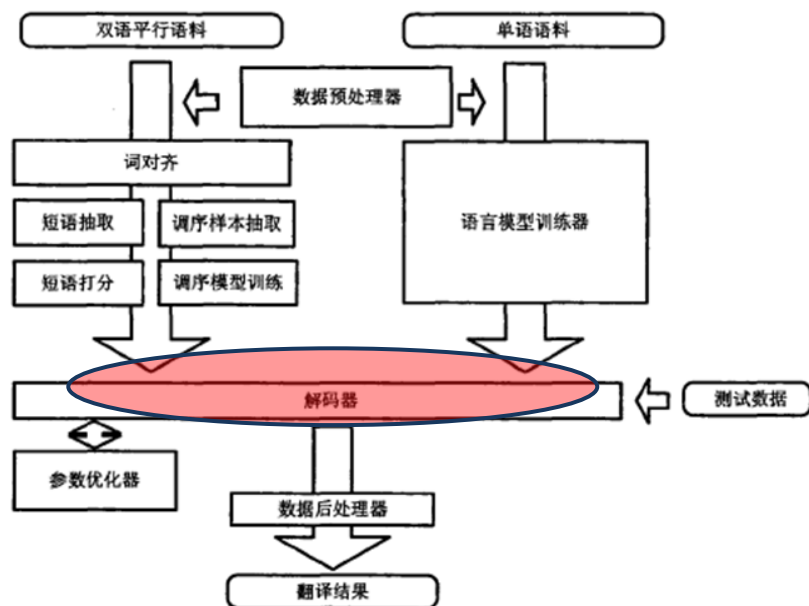
$$f_{S-pre}(d) = \prod_{i=1}^K \Pr(o_i = S | \bar{s}_{\bar{a}_i}, \bar{t}_i, \bar{a}_{i-1}, \bar{a}_i)$$

$$f_{D-pre}(d) = \prod_{i=1}^K \Pr(o_i = D | \bar{s}_{\bar{a}_i}, \bar{t}_i, \bar{a}_{i-1}, \bar{a}_i)$$

如，计算结果：

Source	Target	Monotone	Swap	Discontinuous
中国 经济	China's Economic	0.5	0.1	0.4
.....
.....

附录.2 基于短语的统计机器翻译方法



基于短语的对数线性模型翻译系统架构

4. 解码器

翻译系统核心模块，其任务：

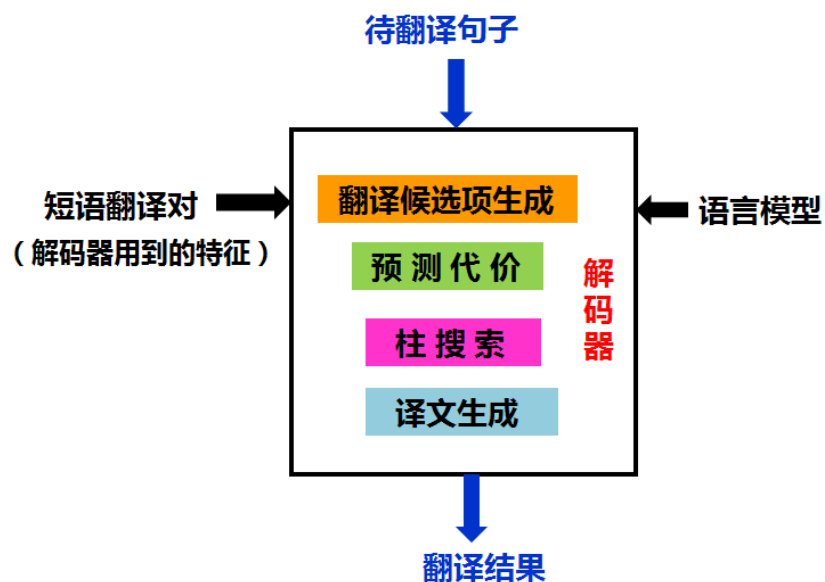
在给定源语言句子与训练好的模型资源（短语翻译表、调序模型、语言模型等）的情况下从所有可能的翻译选项中找到最好的翻译结果。

$$t^* = \arg \max_t \sum_{d \in D(s, t)} \Pr(t, d | s)$$

其中， s 是源语序列， t 是目标语序列； $D(s, t)$ 是可以从源语言句子生成翻译结果的所有翻译路径的集合， d 是翻译路径空间 $D(s, t)$ 中一条翻译路径。

附录.2 基于短语的统计机器翻译方法

4. 解码器



解码器对于统计机器翻译系统而言，可以说是最核心也是最关键的模块。真正的翻译过程就是通过解码器完成的。统计机器翻译发展初期到现在，学者们提出了许多解码算法。

附录.2 基于短语的统计机器翻译方法

◆ 翻译候选项生成

在解码过程开始前，系统可以依据短语翻译表为源语言句子的每个短语加载其对应的目标语翻译选项以及一些必要的信息，以便计算后继模型分数，对译文的好坏作出评估。

这些翻译选项需要存储起来，形成解码过程中的翻译中间结果。这些中间结果会随着解码过程继续扩展、更新直至不能被再扩展。

附录.2 基于短语的统计机器翻译方法

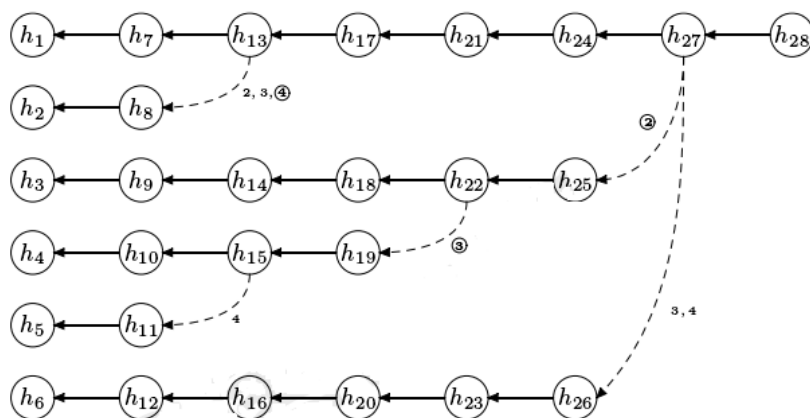
Beam Search 算法

- (1) 初始化假设栈 **HypoStack[0 .. nf]**;
- (2) 创建初始假设 **hyp_init**; 将假设压入栈 **HypoStack[0]**
- (3) 从 $i=0$ 到 $i = nf-1$ 执行如下循环:
 - 对于栈 **HypoStack[i]** 中的每个假设 **hx** 执行如下循环{
 - 对于 **hx** 可生成的每个新假设 **new_hx** 执行操作 {
 - nf[new_hx]** = 被 **new_hx** 覆盖的源语言单词个数;
 - 把新假设 **new_hx** 压栈 **hypoStack[nf[new_hx]]**;
 - 对栈 **HypoStack[nf[new_hx]]** 进行剪枝; } }
- (4) 从栈 **HypoStack[nf]** 中找到最好的假设 **best_hyp**;
- (5) 输出产生 **best_hyp** 的最佳路径。

附录.2 基于短语的统计机器翻译方法

◆ 译文生成

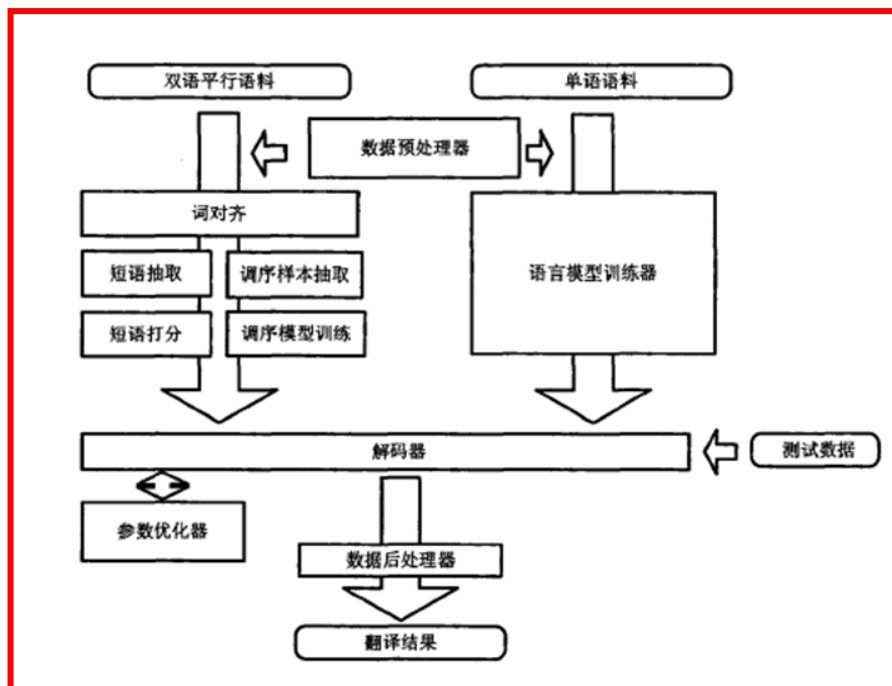
基于 beam search 的解码算，由于将翻译假设压入栈的时候对符合限制的翻译假设进行了合并，最后的n-best 翻译结果不一定都在最后一个栈中。



Beam search 解码算法产生 n-best 翻译结果

从最后一个栈开始从后向前回溯产生路径找到概率最大的 n 个 Hypothesis
根据其指向父亲节点的指针向前回溯。可以产生 n-best 翻译结果。

附录.2 基于短语的统计机器翻译方法



基于短语的对数线性模型翻译系统架构

短语-线性对数模型的SMT系统实现

- ◆ 语料准备及预处理
- ◆ 词语对齐
- ◆ 短语抽取和概率计算
- ◆ 调序模型训练
- ◆ 目标语的语言模型训练
- ◆ 解码器实现
- ◆ 输出后处理
- ◆ 数优化
- ◆ 系统测试 (性能评价)

附录.2 基于短语的统计机器翻译方法

5. 短语-线性对数模型的SMT系统实现

1. 语料准备及预处理

- 准备大规模**双语句子级对齐语料**和**目标语言句子语料**

- 对语料预处理

 - 汉语：分词、全角转半角等

 - 英语：词汇化、大小写转换、标点分离等

- 双语平行语料库分成三个部分：

 - 训练数据集**：统计机器翻译系统从学习翻译知识库（训练短语翻译、调序模型）。一般占到整个双语平行语料库的 80%

 - 开发数据集**：用于调试系统模型参数。一般占 10%

 - 测试数据集**：用于测试系统的翻译性能。一般占 10%

 - 三个数据集相互之间没有交集**

- 目标语的单语语料

 - 训练目标语的语言模型，来衡量翻译译文的流畅度

附录.2 基于短语的统计机器翻译方法

2. 词语对齐

利用词对齐工具：GIZA++

3. 短语抽取和概率计算

4 个常用的概率：

(1) 英汉短语翻译概率
$$p(\tilde{c} | \tilde{e}) = \frac{N(\tilde{c}, \tilde{e})}{\sum_{\tilde{c}'} N(\tilde{c}', \tilde{e})}$$

(2) 英汉词汇化的翻译概率

$$lex(c_1^J | e_1^I, a) = \prod_{j=1}^J \frac{1}{|\{i | (j, i) \in a\}|} \sum_{\forall (j, i) \in a} p(c_j | e_i)$$

(3) 汉英短语翻译概率 $p(\tilde{e} | \tilde{c})$

(4) 汉英词汇化的翻译概率 $lex(\tilde{e} | \tilde{c})$

附录.2 基于短语的统计机器翻译方法

4. 调序模型训练

选择适当的调序模型

5. 目标语的语言模型训练

可采用 SRILM 等工具

6. 解码器实现

7. 输出后处理

- ☐ 未登录词的处理
- ☐ 去分词化
- ☐ 大小写信息恢复
- ☐ 标点符号的特殊匹配处理等。

8. 数优化

- ☐ 参数：模型特征权重
- ☐ 训练数据：开发数据集
- ☐ 训练算法：最小错误率训练算法

9. 系统测试（性能评价）

附录.2 基于短语的统计机器翻译方法

系统实现可用工具

(1) 语料预处理工具

工具: EGYPT- TokenizeE.perl.tmpl

<http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/>

汉语分词工具 :

<http://ictclas.nlpir.org/downloads>

<http://www.nlpr.ia.ac.cn/cip/software.htm>

(2)语言模型:

SRI LM:

<http://www.speech.sri.com/projects/srilm/>

CMU-Cambridge LM

<http://mi.eng.cam.ac.uk/~prc14/toolkit.html>

附录.2 基于短语的统计机器翻译方法

(3) 词对齐工具: GIZA++

<http://www.fjoch.com/GIZA++.html>

(4) 解码器:

- 法老 Pharaoh

<http://www.isi.edu/licensed-sw/pharaoh/>

- 摩西 Moses

<http://www.statmt.org/moses/>

- ReWrite

<http://www.isi.edu/licensed-sw/rewrite-decoder/>

(5) EM 最小错误率参数训练工具

<http://www.cs.cmu.edu/ashish/mer.html>

附录.2 基于短语的统计机器翻译方法

短语翻译模型的问题

基于短语的翻译模型有益于实现局部词义消歧和局部语序调整，但

1. 当短语长度扩展到3个以上的单词时，翻译系统的性能提高很少，短语长度增大以后，数据稀疏问题变得非常严重。因此短语翻译很难解决长距离调序问题。
2. 由于短语是形式上的连续字串，无法处理非连续语言短语翻译现象，例如（在 ... 时，when ...）

本质上讲，这些基于短语的模型所存在的问题都和语言本身的结构性有关。基于此，机器翻译领域的研究者相继开展了基于句法翻译方法的研究并已提出了一些模型。这些句法模型可以分为两类：

附录.2 基于短语的统计机器翻译方法

■ 形式上基于句法的模型

- 不使用任何语言学知识，语法规则中的非终结符并不是语言学意义上的句法标记，而仅仅是一些抽象的变量符号。
- 所有句法结构直接从未标注的语料库中自动学习得到

例如： **层次化短语翻译**

■ 语言学上基于句法的模型

- 使用语言学知识
- 语言通常要从句法树库训练得到

例如：

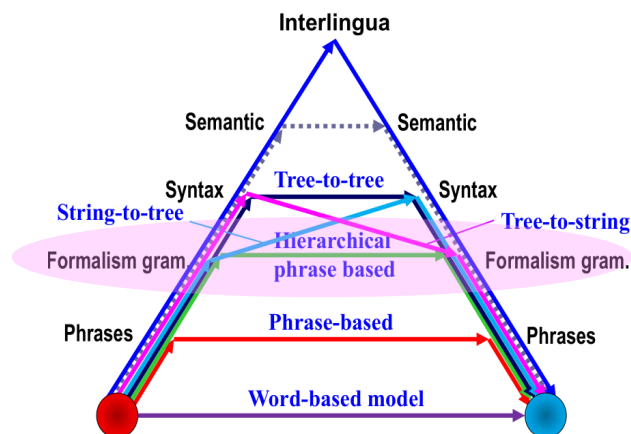
- **树到串模型：只在源语言端使用语言知识**
- **串到树模型：只在目标语言端使用语言知识**
- **树到树模型：在源语言端和目标语言端都使用语言知识**

内 容 提 要

- 附录. 1 基于词的统计机器翻译方法
- 附录. 2 基于短语的统计机器翻译方法
- 附录. 3 基于层次化短语的统计机器翻译方法
- 附录. 4 基于树的统计机器翻译方法
- 附录. 5 译文评估方法

附录.3 基于层次化短语的统计机器翻译方法

基于层次短语的统计翻译模型：



翻译单位： 层级短语模板

翻译模型： 句法分析方法

代表模型：

基于层次短语的翻译模型
[David Chiang, 2005]

核心是引入了**嵌套层次短语**的思想，并采用**括号转录语法**（同步上下文无关语法的特例）作为形式化方法，在完成源语言句法分析的同时，目标语言就生成了，因此可利用各种成熟的句法分析算法进行机器翻译，而无需另外设计专门的翻译算法。

优点：易于实现，解码过程复杂度相对较低，效果比传统短语模型有很大提高

附录.3 基于层次化短语的统计机器翻译方法

基于层次短语的翻译模型

特点:

- 采用 括号转录语法的形式句法
- 在完成源语言句法分析的同时, 生成目标语言
- 所有句法结构规则 (短语模板) 不使用任何语言学知识直接从平行语料库中自动学习得到

涉及内容:

1. 层次短语模型文法
2. 层次短语模型翻译过程
3. 层次短语模型句法结构规则学习

1. 层次短语模型文法

附录.3 基于层次化短语的统计机器翻译方法

◆ 同步上下文无关文法 (SCFG)

SCFG是CFG的扩展，是上下文无关文法针对两个输出符号串的泛化。

CFG (Σ, N, P, S) 包括终结符集合 Σ ，非终结符集合 N ，和产生式集合 $\{P \rightarrow \{N^* \times N^*\}\}$ ，而在**同步上下文无关文法**中，文法指定每个产生式包含两个输出。这些产生式通过共标的非终结符在两个输出字符串间建立联系。

如

$NP \rightarrow DT_1 NPB_2 / \text{DT}_1 NPB_2$

$NPB \rightarrow NPB_1 AJ_2 / \text{AJ}_2 NPB_1$

$NPB \rightarrow JJ_1 NN_2 / \text{JJ}_1 NN_2$

$DT \rightarrow \text{the} / \epsilon$

$AJ \rightarrow \text{strong} / \text{呼啸}$

$JJ \rightarrow \text{north} / \text{北}$

$NN \rightarrow \text{wind} / \text{风}$

同步上下文无关文法生成同构的源语和目标语一对树，树上对应的非终结符对齐。其中一个树可以通过旋转非终结符节点转换为另一个树。

附录.3 基于层次化短语的统计机器翻译方法

同步上下文无关文法

$NP \rightarrow DT_1NPB_2 / \underline{DT_1NPB_2}$

$NPB \rightarrow NPB_1AJ_2 / \underline{AJ_2NPB_1}$

$NPB \rightarrow \underline{JJ_1NN_2} / \underline{JJ_1NN_2}$

$DT \rightarrow \text{the} / \epsilon$

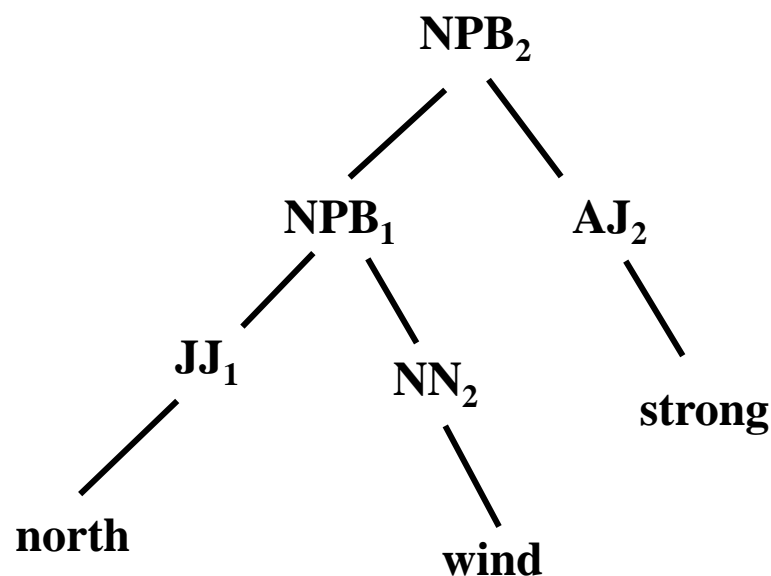
$AJ \rightarrow \underline{\text{strong} / \text{呼啸}}$

$JJ \rightarrow \underline{\text{north} / \text{北}}$

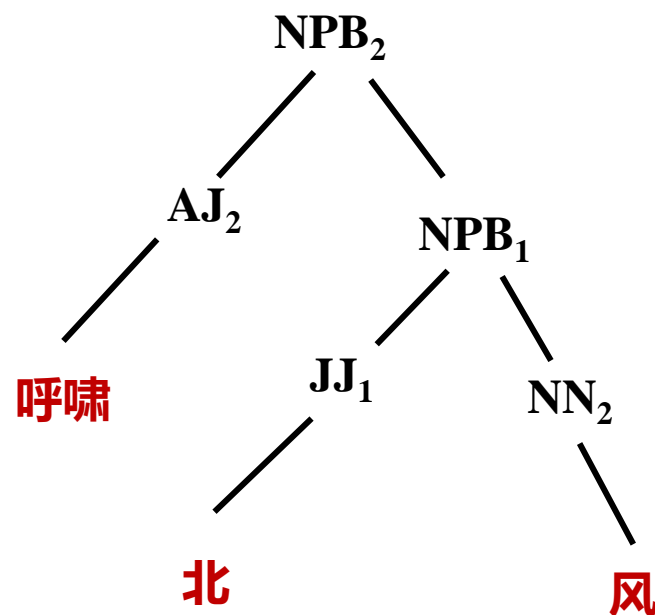
$NN \rightarrow \underline{\text{wind} / \text{风}}$

如： 源语言： north wind strong

用同步上下文无关文法对源语言进行分析



源语言： north wind strong



目标语： 呼啸北风

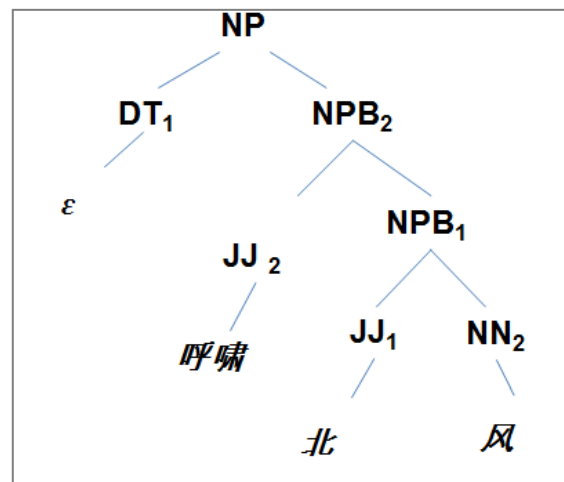
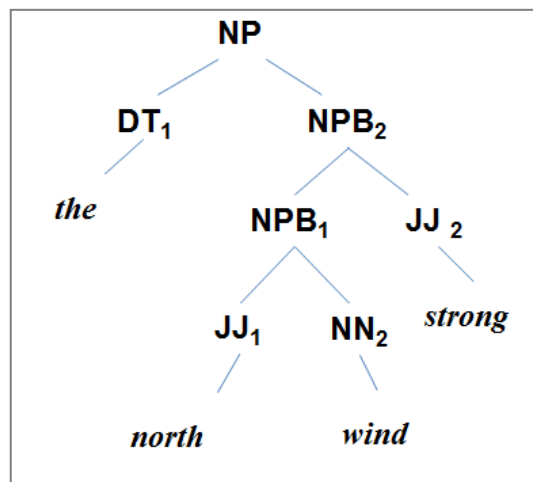
附录.3 基于层次化短语的统计机器翻译方法

如： 源语言: the north wind strong

用同步上下文无关文法对源语言进行分析

同步上下文无关文法

$NP \rightarrow DT_1 NPB_2 / DT_1 NPB_2$
 $NPB \rightarrow JJ_1 NN_2 / JJ_1 NN_2$
 $NPB \rightarrow NPB_1 JJ_2 / JJ_2 NPB_1$
 $DT \rightarrow the / \epsilon$
 $JJ \rightarrow strong / 呼啸$
 $JJ \rightarrow north / 北$
 $NN \rightarrow wind / 风$



源语言: the north wind strong

目标语: 呼啸北风

附录.3 基于层次化短语的统计机器翻译方法

◆ 括号转录语法

括号转录语法 (Bracketing Transduction Grammar: BTG) 是 ITG 的一个特例

- 只有唯一的一个非终结符 X
- 可以理解为：BTG 仅仅给出了两种语言的句子结构结构之间的对应关系，没有任何句法标记信息（如 NP、VP 等等）

括号文法使用单独一种非终结符和如下三个规则：

$$X \rightarrow X_1 X_2 / X_1 X_2 \quad (\text{规则 1})$$

$$X \rightarrow X_1 X_2 / X_2 X_1 \quad (\text{规则 2})$$

$$X \rightarrow e / f \quad (\text{规则 3})$$

附录.3 基于层次化短语的统计机器翻译方法

括号文法:

$X \rightarrow X_1 X_2 / X_1 X_2$ (规则 1)

$X \rightarrow X_1 X_2 / X_2 X_1$ (规则 2)

$X \rightarrow e / f$ (规则 3)

规则3是所谓的词汇化规则，用于将双语的源语言部分和目标语部分联系起来。而规则1和规则2主要是针对调序的，使得文法只可以支持正向调序和反向调序这两种调序。

步同上下文无关文法

$S \rightarrow NP_1 VP_2 / NP_1 VP_2$

$NP \rightarrow e_1 / f_1$

$VP \rightarrow N_1 V_2 / V_2 N_1$

$V \rightarrow e_2 / f_2$

$N \rightarrow e_3 / f_3$

反向转录语法

$S \rightarrow [NP VP]$

$NP \rightarrow e_1 / f_1$

$VP \rightarrow \langle N V \rangle$

$V \rightarrow e_2 / f_2$

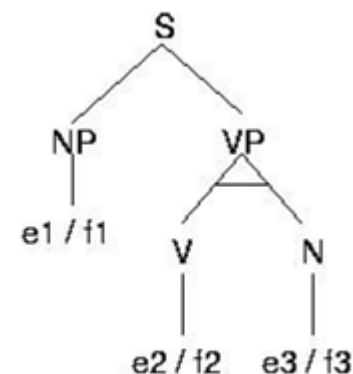
$N \rightarrow e_3 / f_3$

括号文法

$X \rightarrow X_1 X_2 / X_1 X_2$

$X \rightarrow X_1 X_2 / X_2 X_1$

$X \rightarrow e / f$



附录.3 基于层次化短语的统计机器翻译方法

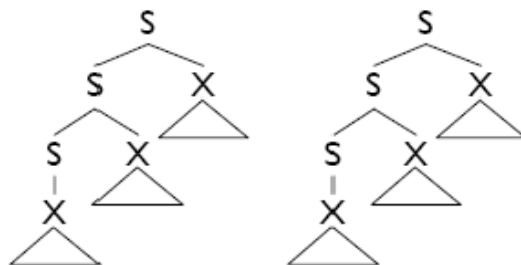
括号文法的粘合规则 (Glue Rules)

找不到可用的规则时，引入粘合规则

$$(S \rightarrow S_1 X_2, S_1 X_2)$$

$$(S \rightarrow X_1, X_1)$$

粘合规则的作用：将短语的译文从左到右依次顺序“粘合”成完整的译文：



括号文法

$$X \rightarrow X_1 X_2 / X_1 X_2 \quad (\text{规则 1})$$

$$X \rightarrow X_1 X_2 / X_2 X_1 \quad (\text{规则 2})$$

$$X \rightarrow e / f \quad (\text{规则 3})$$

粘合规则

$$(S \rightarrow S_1 X_2, S_1 X_2)$$

$$(S \rightarrow X_1, X_1)$$

附录.3 基于层次化短语的统计机器翻译方法

例如： 层次短语翻译模型括号文法

- (1) $X \rightarrow \langle \text{与 } X_1 \text{ 有 } X_2, \text{ have } X_2 \text{ with } X_1 \rangle$
- (2) $X \rightarrow \langle X_1 \text{ 的 } X_2, \text{ the } X_2 \text{ that } X_1 \rangle$
- (3) $X \rightarrow \langle X_1 \text{ 之一, one of } X_1 \rangle$
- (4) $X \rightarrow \langle \text{澳洲, Australia} \rangle$
- (5) $X \rightarrow \langle \text{邦交, diplomatic relations} \rangle$
- (6) $X \rightarrow \langle \text{少数国家, few countries} \rangle$

2. 层次短语模型翻译过程

附录.3 基于层次化短语的统计机器翻译方法

层次短语模型翻译

层次短语语法规则采用同步语法，使得在对源语端进行形式语法分析时可以自动生成译文。

例： 语法规则：

- (1) $X \rightarrow \langle \text{与 } X_1 \text{ 有 } X_2, \text{ have } X_2 \text{ with } X_1 \rangle$
- (2) $X \rightarrow \langle X_1 \text{ 的 } X_2, \text{ the } X_2 \text{ that } X_1 \rangle$
- (3) $X \rightarrow \langle X_1 \text{ 之一, one of } X_1 \rangle$
- (4) $X \rightarrow \langle \text{澳洲, Australia} \rangle$
- (5) $X \rightarrow \langle \text{邦交, diplomatic relations} \rangle$
- (6) $X \rightarrow \langle \text{少数国家, few countries} \rangle$
- (7) $X \rightarrow \langle \text{北韩, North Korea} \rangle$
- (8) $(S \rightarrow S_1 X_2, S_1 X_2)$
- (9) $(S \rightarrow X_1, X_1)$

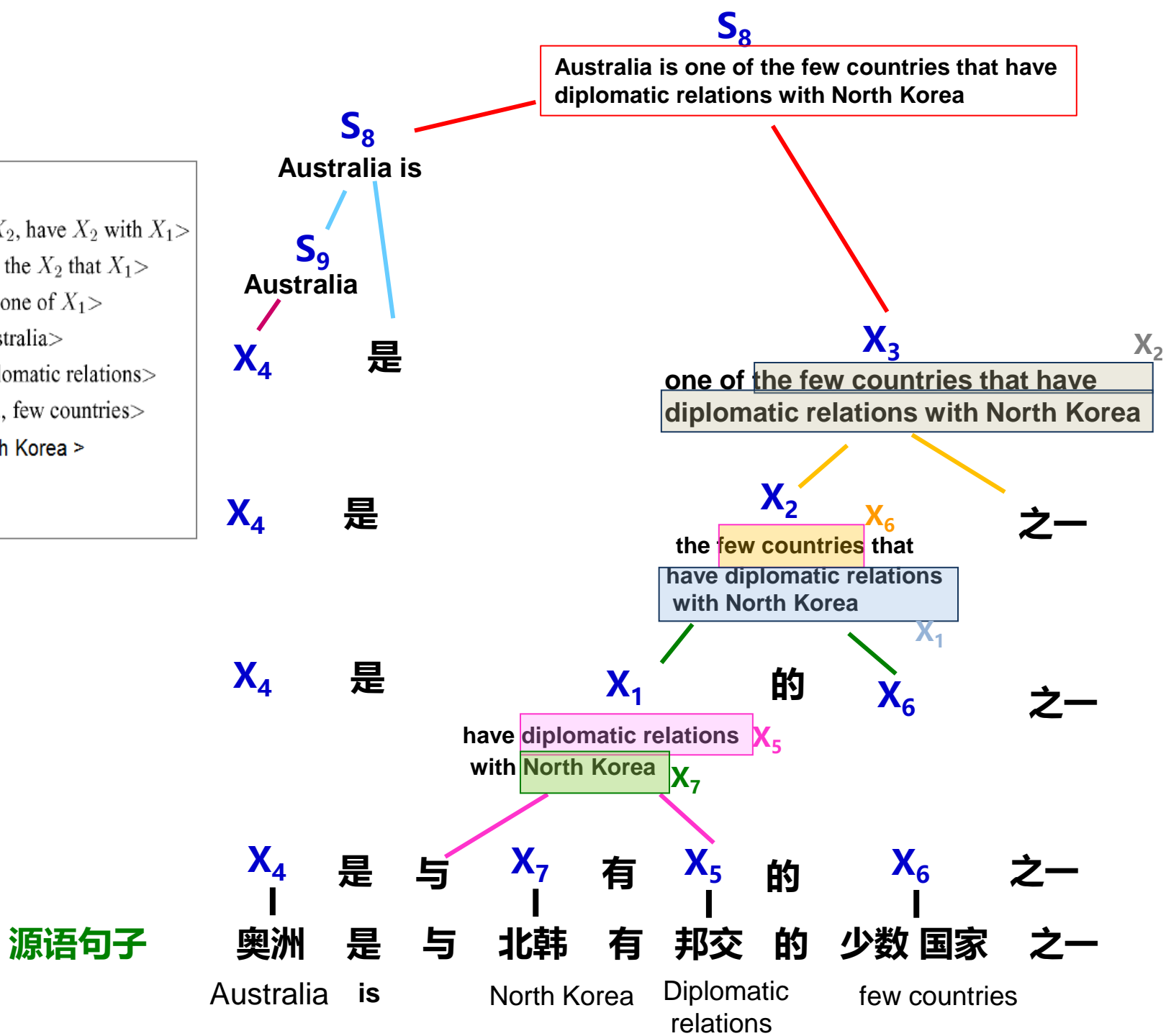
输入：源语句子 **澳洲 是 与 北韩 有 邦交 的 少数 国家 之一**

翻译过程： 用语法规则对源语句子按层进行分析，**翻译过程类似句法分析**，在分析源语言同时会生成目标语的句法分析树。

翻译过程

文法规则：

- (1) $X \rightarrow \langle \text{与 } X_1 \text{ 有 } X_2, \text{ have } X_2 \text{ with } X_1 \rangle$
- (2) $X \rightarrow \langle X_1 \text{ 的 } X_2, \text{ the } X_2 \text{ that } X_1 \rangle$
- (3) $X \rightarrow \langle X_1 \text{ 之一, one of } X_1 \rangle$
- (4) $X \rightarrow \langle \text{澳洲, Australia} \rangle$
- (5) $X \rightarrow \langle \text{邦交, diplomatic relations} \rangle$
- (6) $X \rightarrow \langle \text{少数国家, few countries} \rangle$
- (7) $X \rightarrow \langle \text{北韩, North Korea} \rangle$
- (8) $(S \rightarrow S_1 X_2, S_1 X_2)$
- (9) $(S \rightarrow X_i, X_i)$



3. 层次短语模型句法结构规则学习

附录.3 基于层次化短语的统计机器翻译方法

基于层次短语的翻译模型

特点:

- 形式上基于句法的模型
- 不使用任何语言学知识
- 所有句法结构规则直接从平行语料库中自动学习得到

如何从平行语料句对中抽取层次短语规则?

附录.3 基于层次化短语的统计机器翻译方法

问题：给定平行句对：

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

Australia is one of the few countries that have diplomatic relations with North Korea

如何抽取翻译文法规则？

- 如：
- (1) $X \rightarrow \langle \text{与 } X_1 \text{ 有 } X_2, \text{ have } X_2 \text{ with } X_1 \rangle$
 - (2) $X \rightarrow \langle X_1 \text{ 的 } X_2, \text{ the } X_2 \text{ that } X_1 \rangle$
 - (3) $X \rightarrow \langle X_1 \text{ 之一, one of } X_1 \rangle$
 - (4) $X \rightarrow \langle \text{澳洲, Australia} \rangle$
 - (5) $X \rightarrow \langle \text{邦交, diplomatic relations} \rangle$
 - (6) $X \rightarrow \langle \text{少数国家, few countries} \rangle$
 - (7) $X \rightarrow \langle \text{北韩, North Korea} \rangle$

附录.3 基于层次化短语的统计机器翻译方法

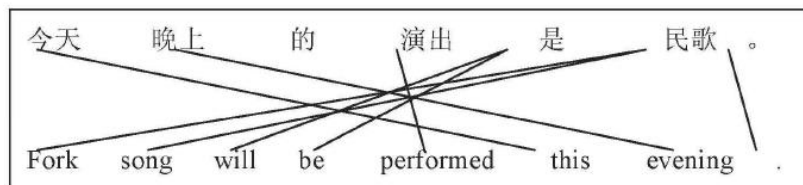
层次短语规则抽取

- 步骤:**
- 1.**首先使用GIZA++获得双语语料双向的词对齐, 然后取双向词对齐的并集作为最终的词对齐。
 - 2.**使用短语模型的方法抽取出**初始短语对**, 一个初始短语对要满足至少有一个词对齐, 并且两个短语之间的所有词对齐不能超过两个短语的范围。
 - 3.**从短语获得**规则**: 找到那些包含其他初始短语的短语, 将被包含的子短语用非终结符好代替。

附录.3 基于层次化短语的统计机器翻译方法

如:

句对齐的训练语料



抽取出的层次短语规则

< ||| 。 > <Folk song ||| 民歌> <[X,1] . ||| [X,1] 。 > <[X,1] will be ||| 是 [X,1]>
<evening ||| 晚上> <performed ||| 演出> <will be [X,1] ||| [X,1] 是>
<this ||| 今天> <this evening ||| 今天 晚上>
<Folk song [X,1] . ||| [X,1] 民歌 。 > <will be ||| 是>
<Folk song [X,1] performed [X,2] ||| [X,2] 的 演出 [X,1] 民歌>
<Folk song [X,1] performed ||| 演出 [X,1] 民歌>
<Folk song [X,1] this evening ||| 今天 晚上 的 [X,1] 民歌>
<Folk song [X,1] ||| [X,1] 民歌>
<Folk song will be [X,1] ||| [X,1] 是 民歌>
<Folk song will be performed ||| 演出 是 民歌>
<Folk song will be ||| 是 民歌>
<[X,1] performed [X,2] . ||| [X,2] 的 演出 [X,1] 。 >
<[X,1] performed [X,2] ||| [X,2] 的 演出 [X,1]>
<[X,1] performed this evening . ||| 今天 晚上 的 演出 [X,1] 。 >
<[X,1] performed this evening ||| 今天 晚上 的 演出 [X,1]>
<[X,1] performed ||| 演出 [X,1]>
<[X,1] this evening . ||| 今天 晚上 的 [X,1] 。 >
<[X,1] this evening ||| 今天 晚上 的 [X,1]>
<[X,1] will be [X,2] . ||| [X,2] 是 [X,1] 。 >
<[X,1] will be [X,2] ||| [X,2] 是 [X,1]>
<[X,1] will be performed [X,2] ||| [X,2] 的 演出 是 [X,1]>
<[X,1] will be performed ||| 演出 是 [X,1]>
<performed [X,1] ||| [X,1] 的 演出>
<performed this evening ||| 今天 晚上 的 演出>
<will be performed [X,1] ||| [X,1] 的 演出 是>
<will be performed this evening ||| 今天 晚上 的 演出 是>
<will be performed ||| 演出 是>

附录.3 基于层次化短语的统计机器翻译方法

抽出来的短语模板比短语模型的多很多。这就使得训练的过程和解码过程的速度变得很慢;同时还会产生大量的歧义, 因此在抽取时作一些 限制。

如, 限制条件:

- (1) 如果有不同的初始短语对包含相同的对齐集, 只保留最小初始短语对。即, 不保留那些在边界上有对空的词初始短语对。
- (2) 初始短语对的两侧都限制在10个词以内。
- (3) 规则的源语言端最多只能有5个符号(包括非终结符和终结符)。
- (4) 规则最多有两个非终结符。这样一方面可以简化解码器的实现, 另一方面使得层次短语翻译模型使用的文法弱等价于Inversion transduction grammar(ITG)文法。
- (5) 非终结符在规则的源语言端不能相邻。
- (6) 规则中至少存在一个词对齐。

附录.3 基于层次化短语的统计机器翻译方法

层次短语规则打分

抽取出来的层次短语需要进行短语打分. 如, 估计规则的短语翻译概率和词汇化翻译概率, 具体参见 短语 打分部分。

层次短语规则

```
< ||| 。 > <Fork song ||| 民歌> <[X,1] . ||| [X,1] 。 > <[X,1] will be ||| 是 [X,1]> score  
<evening ||| 晚上> <performed ||| 演出> <will be [X,1] ||| [X,1] 是> score  
<this ||| 今天> <this evening ||| 今天 晚上> score  
<Fork song [X,1] . ||| [X,1] 民歌 。 > <will be ||| 是> score  
<Fork song [X,1] performed [X,2] ||| [X,2] 的 演出 [X,1] 民歌> score  
<Fork song [X,1] performed ||| 演出 [X,1] 民歌> score  
<Fork song [X,1] this evening ||| 今天 晚上 的 [X,1] 民歌> score  
<Fork song [X,1] ||| [X,1] 民歌> score  
<Fork song will be [X,1] ||| [X,1] 是 民歌> score  
<Fork song will be performed ||| 演出 是 民歌> score  
<Fork song will be ||| 是 民歌> score  
<[X,1] performed [X,2] . ||| [X,2] 的 演出 [X,1] 。 > score
```

.....

附录.3 基于层次化短语的统计机器翻译方法

层次化短语翻译系统:

SAMT 系统

<http://www.cs.cmu.edu/~zollmann/samt/>

SAMT (Syntax Augmented Machine Translation) 系统由美国卡内基 - 梅隆大学 (CMU) 实现。

该系统同时融合了以下三种短语:

- (1) 基于Moses 系统抽取出来的一般短语;
- (2) 句法驱动的泛化短语;
- (3) 分层短语

附录.3 基于层次化短语的统计机器翻译方法

Joshua 系统

http://www.clsp.jhu.edu/wiki2/JosHUa_-_JHU_Open_Source_Architecture

Joshua 是2008年6月JHU暑期研讨班上开发的一个开源的基于层次短语的统计机器翻译系统。该系统实现了上下文无关语法所需的所有算法，并采用了基于后缀数组的文法规则抽取算法。。

内 容 提 要

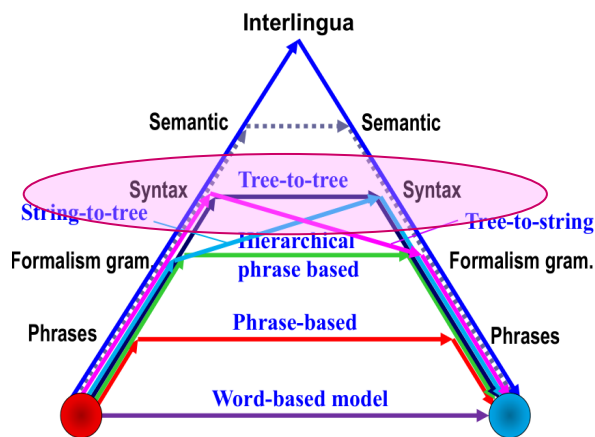
- 附录. 1 基于词的统计机器翻译方法
- 附录. 2 基于短语的统计机器翻译方法
- 附录. 3 基于层次化短语的统计机器翻译方法
- 附录. 4 基于树的统计机器翻译方法
 - (1) 树到树模型
 - (2) 树到串模型
 - (3) 串到树模型
- 附录. 5 译文评估方法

附录.4 基于树的统计机器翻译方法

基于树的统计机器翻译模型

特点:

- 树翻译模型属于语言学上基于句法的模型
 - 使用语言学知识
 - 语言知识规则通常从句法树库训练得到
- **树到树模型**: 在源语言端和目标语言端都使用语言知识
 - **树到串模型**: 只在源语言端使用语言知识
 - **串到树模型**: 只在目标语言端使用语言知识



(1) 树到树模型

树到树的翻译模型

Zhang et al.(2007, 2008) 提出了**树到树的翻译模型**

特点:

- 句法分析：将源语言句子分析为一棵句法结构树（短语结构树）
- 树到树的转换：递归地将源语言句子的句法结构树转换为目标语言句子的句法结构树，拼接叶结点得到译文。

涉及内容

1. 树到树模型文法
2. 树到树模型翻译过程
3. 树到树模型规则学习

1. 树到树模型文法

(1) 树到树模型

◆ 同步树替换文法 (STSG)

STSG是一个七元组: $G = \{\Sigma_s, \Sigma_t, N_s, N_t, P, S_s, S_t\}$

其中:

- (1) Σ_s 和 Σ_t 分别是源语言端和目的语言端的终结符(词语, 单词)字符集
- (2) N_s 和 N_t 分别是源语言端和目的语言端的非终结符(词性, 句法标记等)字符集
- (3) S_s 属于 N_s 以及 S_t 属于 N_t 是源语言端和目的语言端的起始符号 (相对应于源语言, 目的语言的句法树根节点)
- (4) P 是一个产生式规则集合。

注: Shieber在文献中给出定义同步树替换文法是五元组 $G = \{\Sigma_{in}, \Sigma_{out}, P, S_{in}, S_{out}\}$

(1) 树到树模型

树(tree): 假设 Σ 是一个终结符表(对应语言中的词表), N 是一个非终结符表(对应句法标记集), 那么一个三元组 $T = \langle V, E, V_t \rangle$ 是一个树, 如果:

- (1) V 是节点集, E 是边集;
- (2) $V_l \subset V$ 是叶子节点集, 并且有 $\forall u \in V_l, u \in \Sigma, \forall v \in (V \setminus V_l), v \in N$;
- (3) $|V| = |E| + 1$, 且 T 是一个连通图。

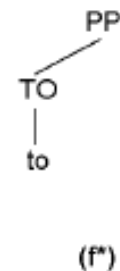
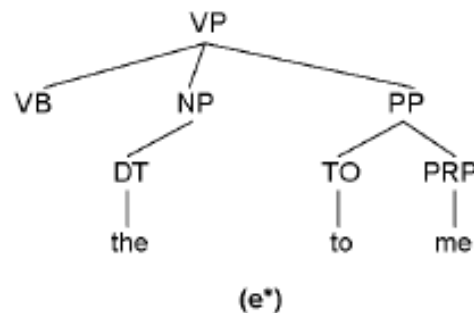
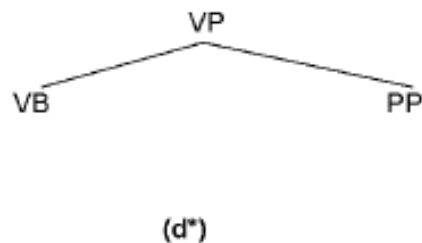
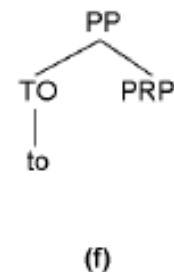
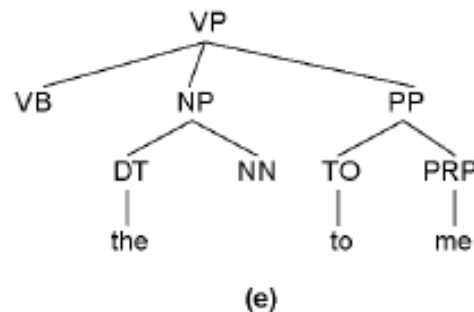
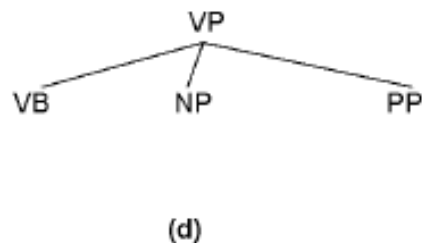
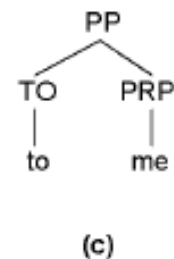
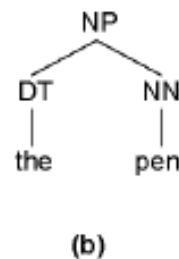
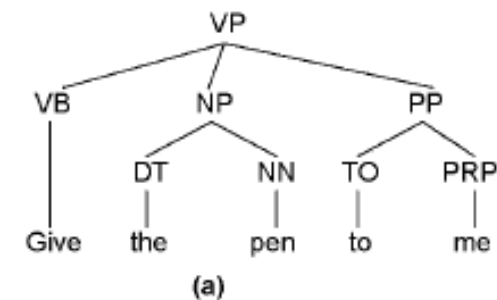
元树(elementary tree): 假设 $T = \langle V, E, V_t \rangle$ 是一个树, 那么元组 $\xi = \langle V', E', V'_t \rangle$ 是 T 的一个元树, 如果:

- (1) V' 是节点集, E' 是边集, $V'_l \subset V'$ 是叶子节点集;
- (2) $\forall u \in V'_l, u \in (\Sigma \cup N), V' \subseteq V, E' \subseteq E$;
- (3) $|V'| = |E'| + 1$, 且 ξ 是一个连通图,
- (4) $\forall w \in (V' \setminus V'_l), \forall v \in \{n | n \text{ 是 } w \text{ 的直接儿子}\} \Rightarrow v \in V'$ 。

一般的子树的叶子节点都必须是终结符, 而元树的叶子节点也可以是非终结符。元树是一个比子树更为广泛的概念, 子树一定是一个元树, 但反之则不然。元树具有局部完整性, 即对于某个非叶子节点, 它的所有儿子节点都必须保留。

(1) 树到树模型

如:



对英文句法树(a), 其中(b),(c)都是一般的子树; (d), (e), (f)均则是元树而不是子树;而(d*), (e*), (f*) 不是元树和子树。

(1) 树到树模型

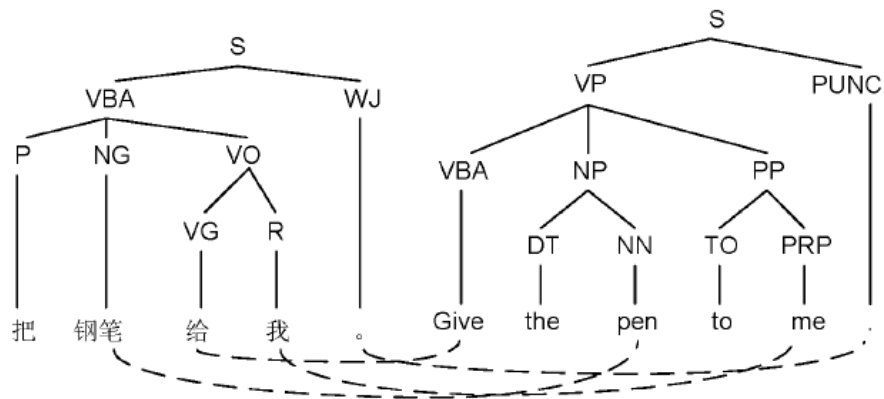
STSG中产生式规则一般形式为:

$$< \alpha, \beta, A >$$

其中, α 为源语言端元树, β 为目的语言端元树, A 为两者节点间对应关系。

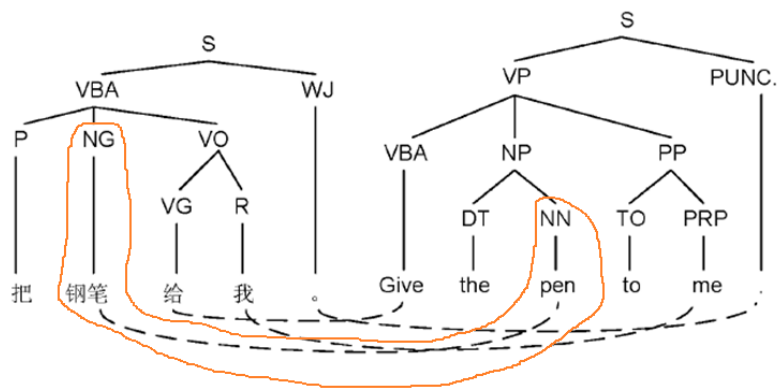
(1) 树到树模型

如:



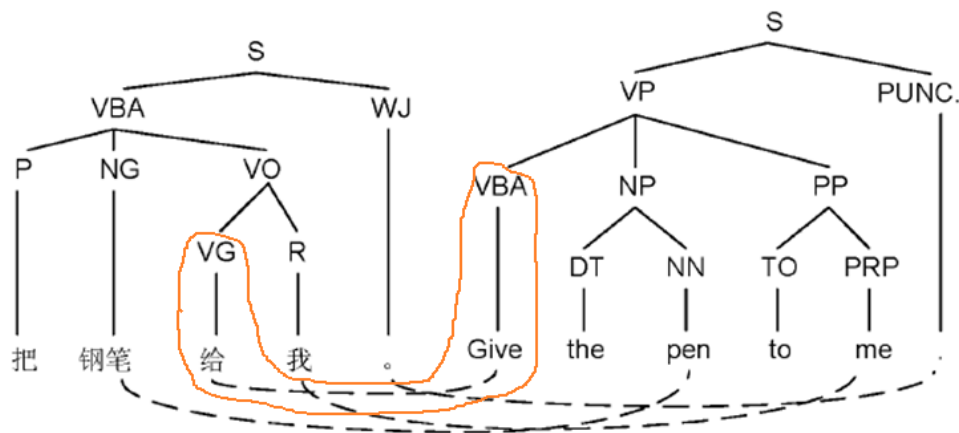
带词对齐关系的句法树对

句法规则:

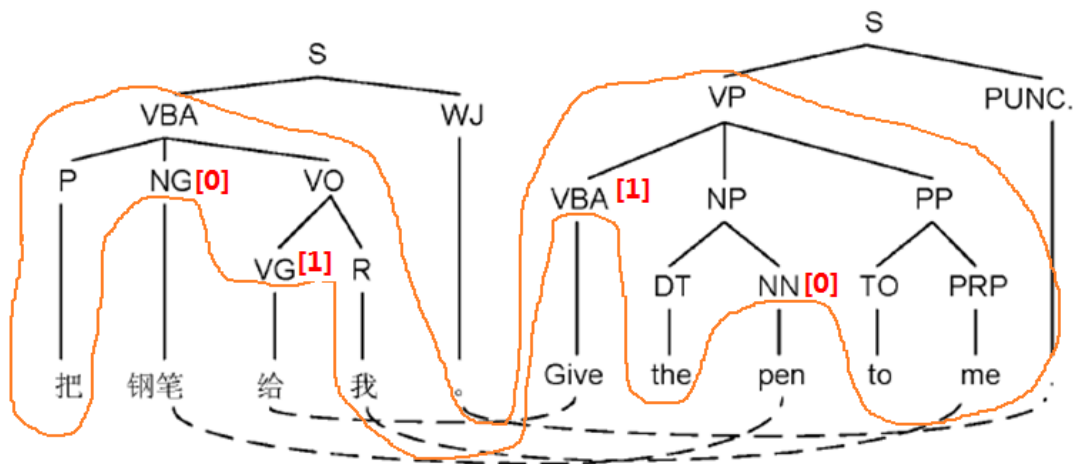


NG(钢笔) ; NN(pen)

(1) 树到树模型



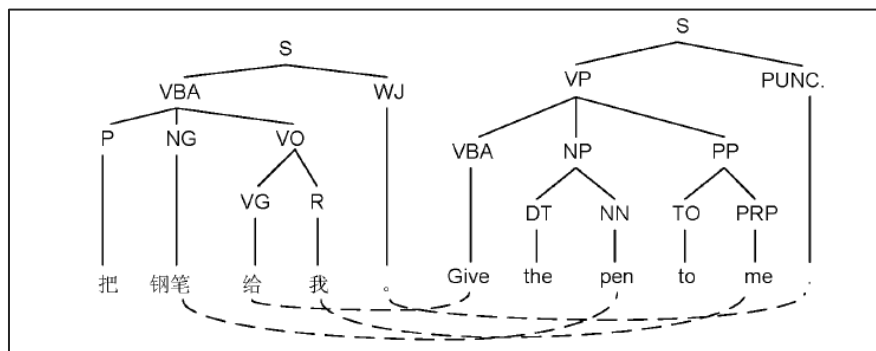
VG(给) ; VBP(Give)



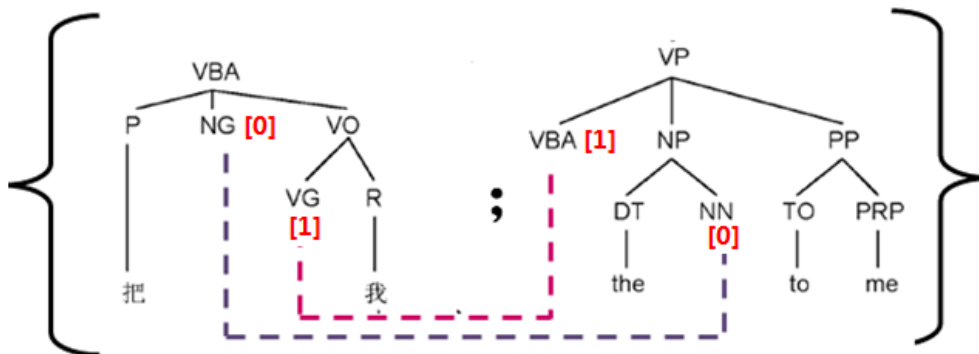
VBA(P(把),NG[0],VO(VG[1],R(我))) ; VP(VBP[1],NP(DT(tne),NN[0]),PP(TO(to),PRP(me)))

(1) 树到树模型

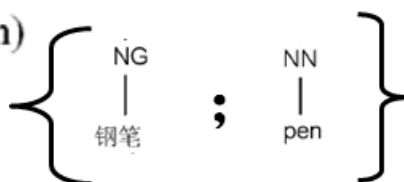
STSG文法的解决上图非兄弟节点节点“NG(钢笔)”和节点“VG(给)”之间的调序“VG(给)”之间的调序问题



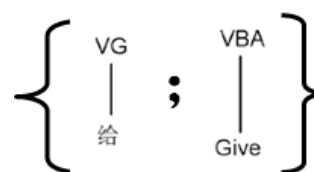
R8
$$\begin{aligned} & \text{VBA(P(把),NG[0],VO(VG[1],R(我)))} \\ \Leftrightarrow & \text{VP(VBP[1],NP(DT(the),NN[0]),PP(TO(to),PRP(me)))} \end{aligned}$$



R1 $\text{NG(钢笔)} \Leftrightarrow \text{NN(pen)}$



R2 $\text{VG(给)} \Leftrightarrow \text{VBP(Give)}$



(1) 树到树模型

树到树翻译模型的规则 (同步树替换文法)

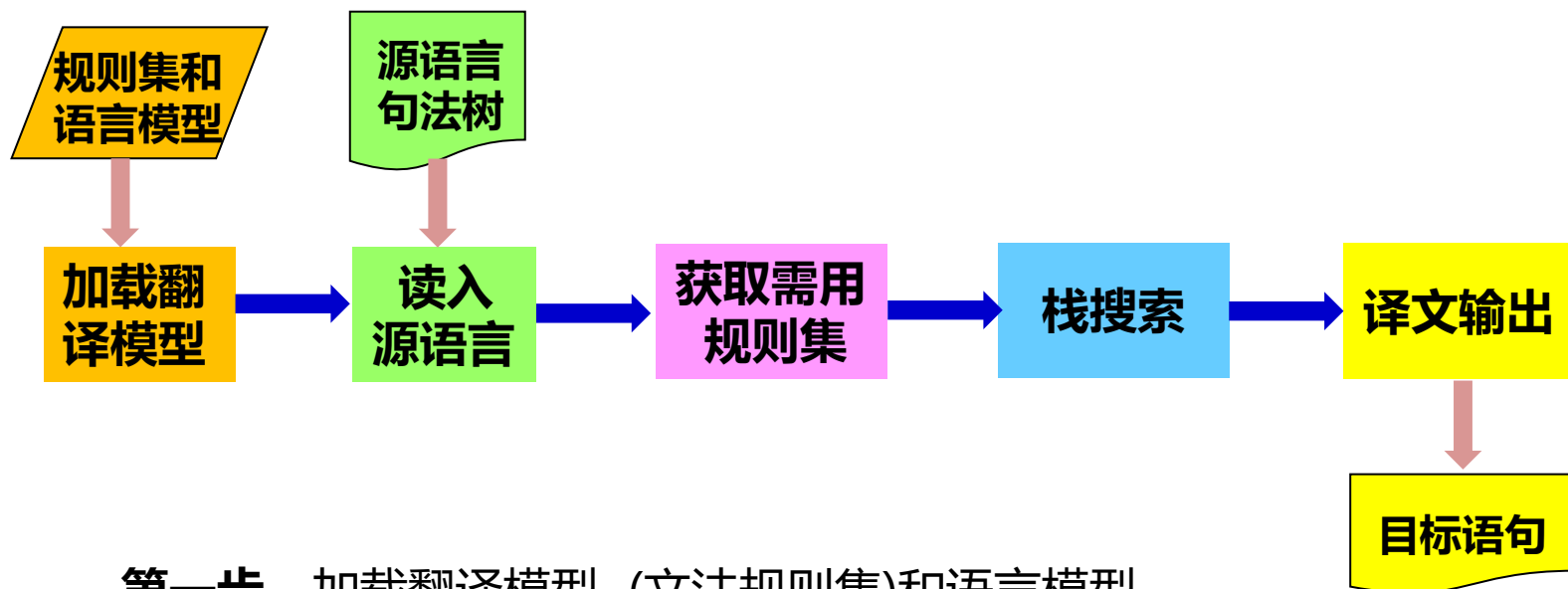
ID	规则(用“ \Leftrightarrow ”作为两端分隔符)
R1	NG(钢笔) \Leftrightarrow NN(pen)
R2	VG(给) \Leftrightarrow VBP(Give)
R3	R(我) \Leftrightarrow PRP(me)
R4	WJ(。) \Leftrightarrow PUNC.(.)
R5	VBA(P(把) NG(钢笔) VO(VG(给) R(我))) \Leftrightarrow VP(VBA(Give) NP(DT(the) NN(pen)) PP(TO(to) PRP(me)))
R6	VBA(P(把),NG[0],VO(VG(给),R(我))) \Leftrightarrow P(VBP(give),NP(DT(the),NN[0]),PP(TO(to),PRP(me)))
R7	VBA(P(把),NG(钢笔),VO(VG[0],R(我))) \Leftrightarrow P(VBP[0],NP(DT(the),NN(pen)),PP(TO(to),PRP(me)))
R8	VBA(P(把),NG[0],VO(VG[1],R(我))) \Leftrightarrow VP(VBP[1],NP(DT(the),NN[0]),PP(TO(to),PRP(me)))
R9	S(VBA[0],WJ[1]) \Leftrightarrow S(VP[0],PUNC.[1])

同步树替换文法中的规则中每条规则的两端(源端和目的端)可以是多层的树结构，所以容易地对处在不同层次中的节点之间的重排序进行模拟

2. 树到树模型翻译过程

(1) 树到树模型

树到树翻译模型翻译过程（解码）



第一步，加载翻译模型（文法规则集）和语言模型

第二步，读入源语言句法树 $T(f)$ ，并对每个树节点进行后续编号。

第三步，获取可用的规则集

第四步，进行从底向上树到树转化（过程称为栈搜索）

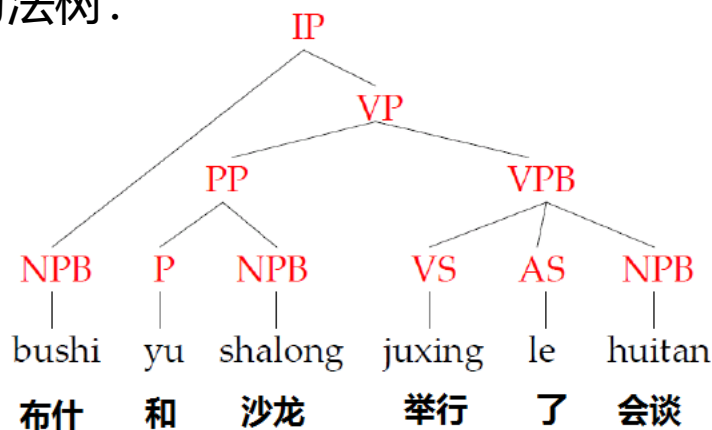
第五步，将最优译文输出

(1) 树到树模型

例:

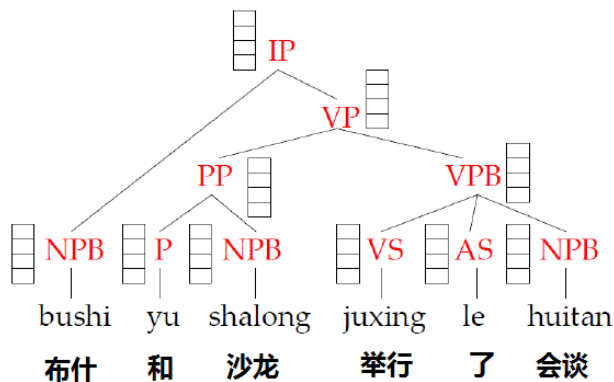
源语句子: 布什 和 沙龙 举行 了 会谈

输入句法树:



规则集和
语言模型

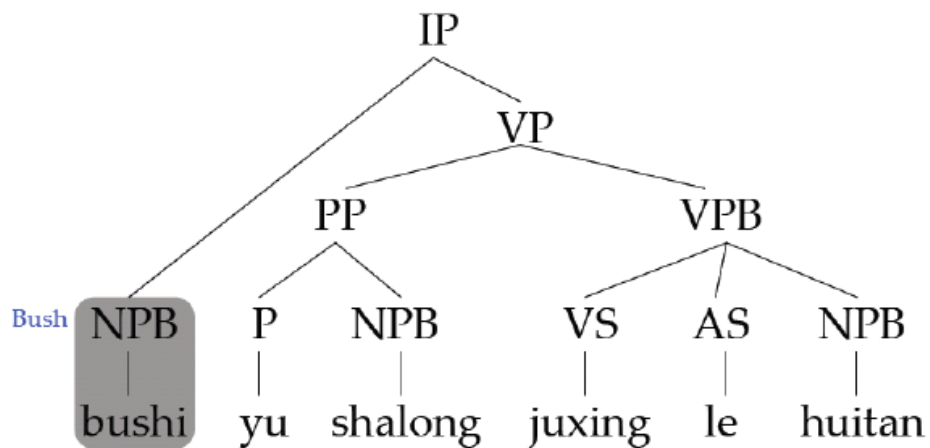
栈搜索



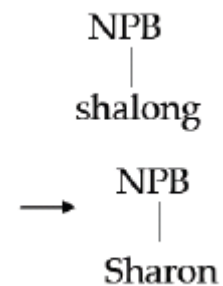
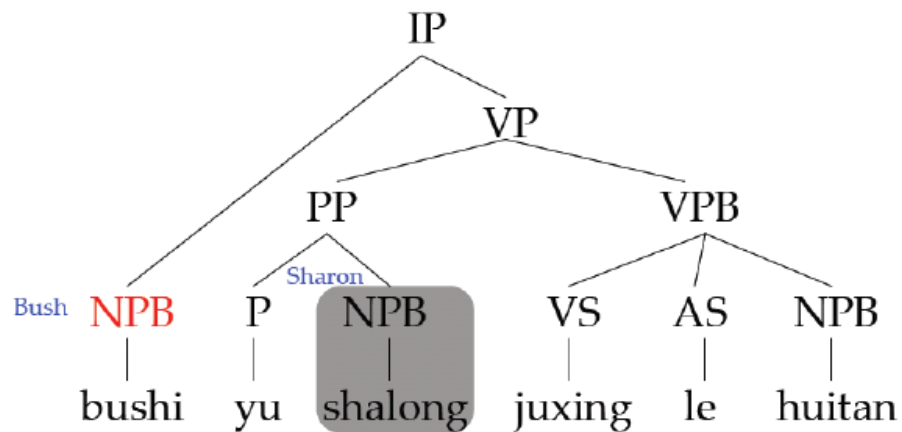
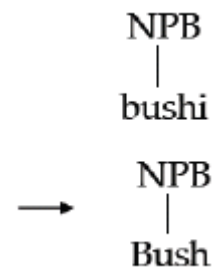
自底向上

- 柱搜索 (Beam Search)
- 对于每一棵子树, 找到所有与其根节点匹配的规则, 计算其候选译文 (Candidate)

(1) 树到树模型

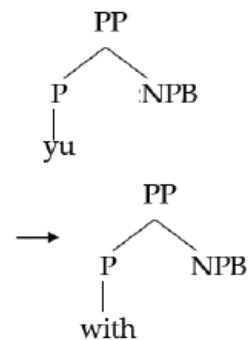
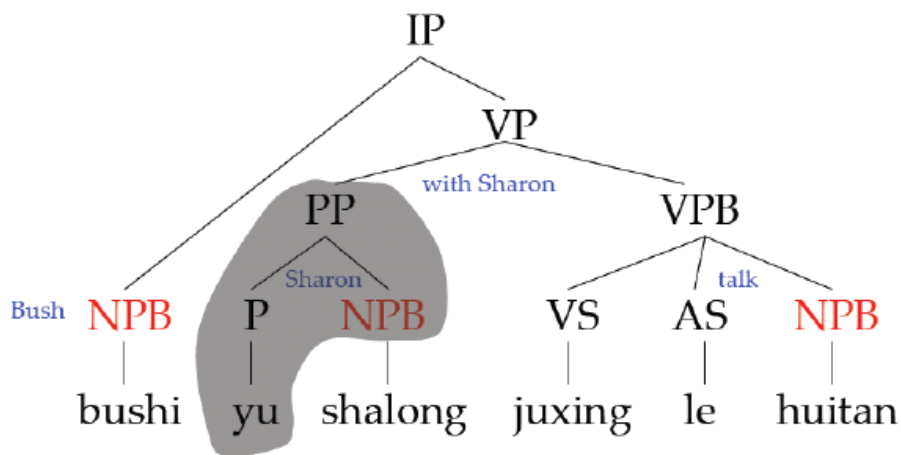
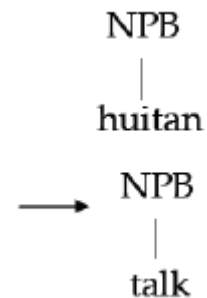
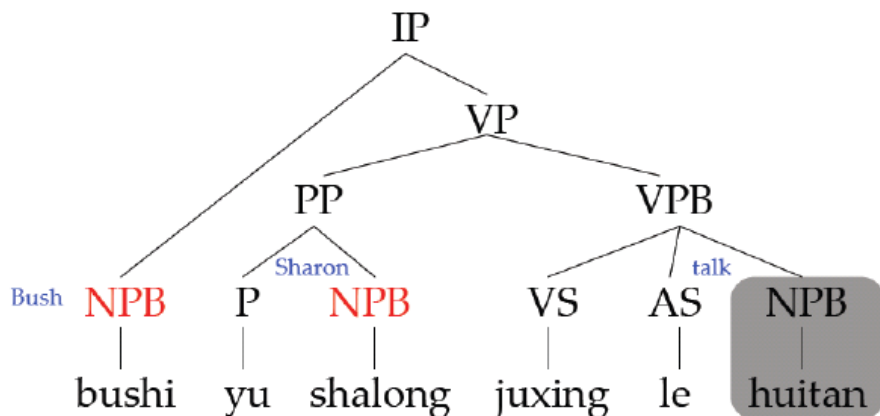


规则集和
语言模型



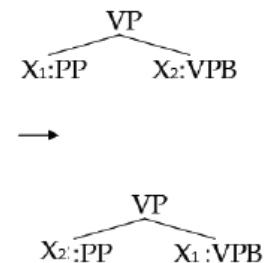
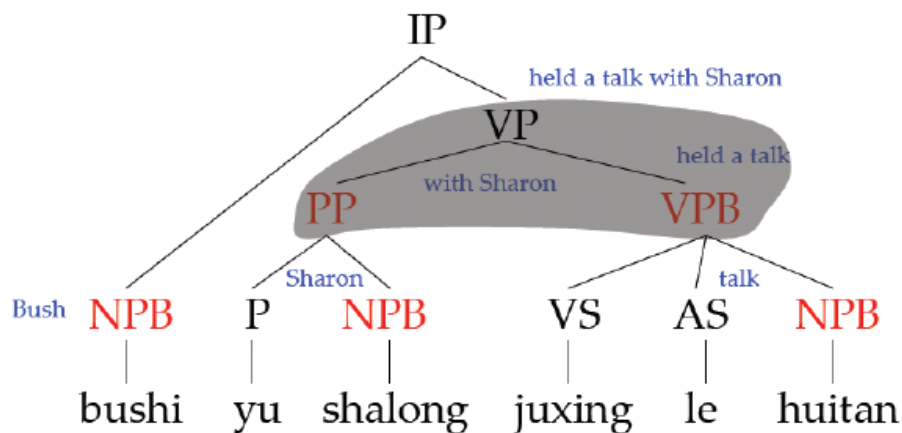
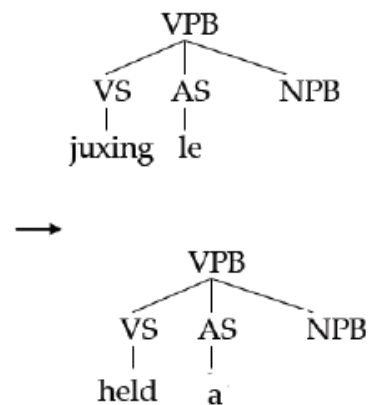
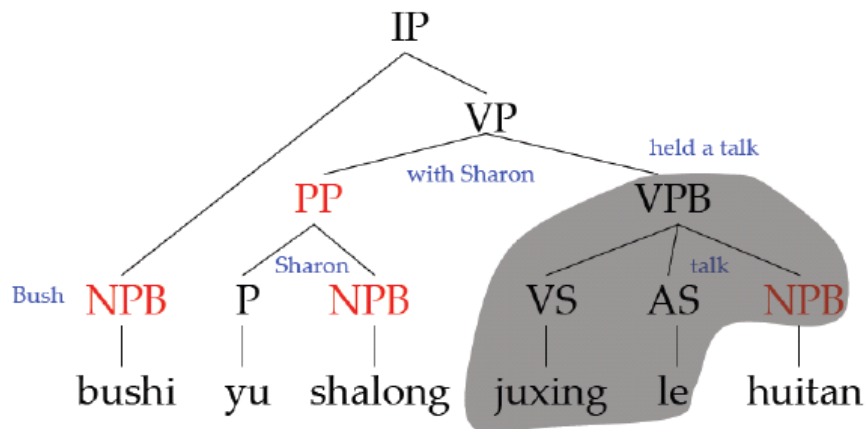
(1) 树到树模型

规则集和
语言模型



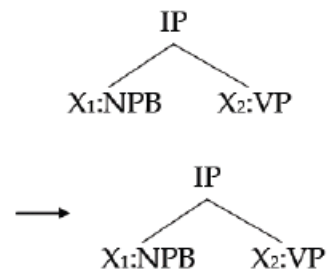
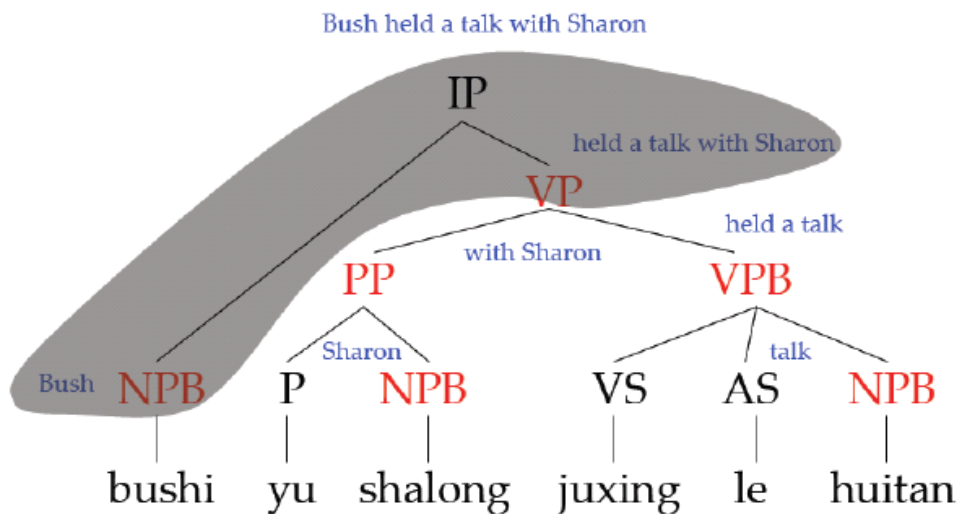
(1) 树到树模型

规则集和语言模型



(1) 树到树模型

规则集和
语言模型

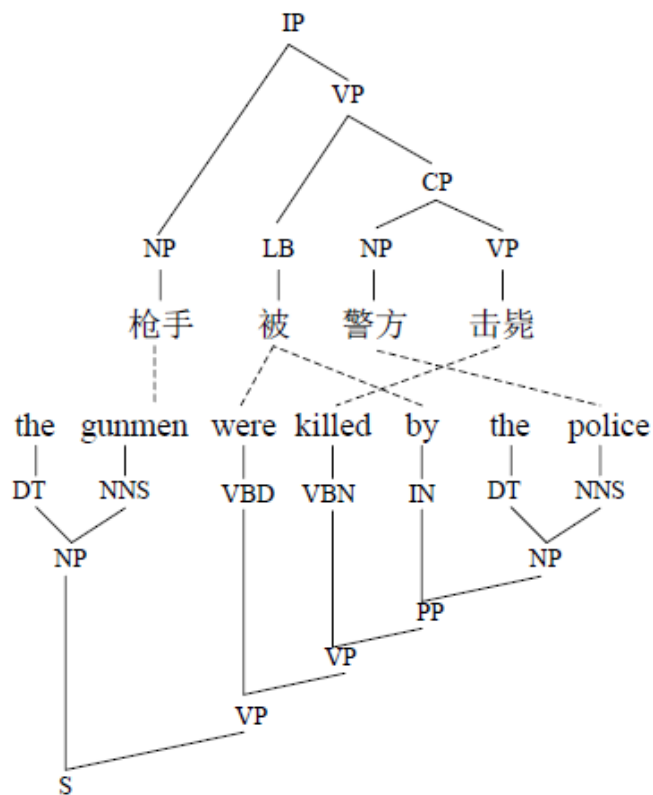


3. 树到树模型规则学习

(1) 树到树模型

树到树翻译模型的规则的学习

规则自动抽取模块的输入是一个具有词对齐信息的双语并行句法树对集合。输出是抽取有用的树到树翻译规则。



(1) 树到树模型

抽取的规则需满足词对齐约束和句法限制

(1) 词对齐约束: 满足 $\exists(i, j) \in A, i_1 \leq i \leq i_2, j_1 \leq j \leq j_2$ 并且满足 $\forall(i, j) \in A: i_1 \leq i \leq i_2 \Leftrightarrow j_1 \leq j \leq j_2$

(2) 句法限制: $T(f_{j_1}^{j_2})$ 是 $T(f)$ 的一棵元树, $T(e_{i_1}^{i_2})$ 是 $T(e)$ 的一棵元树

按照是否具有泛化能力, 可以将产生式规则分为两类:

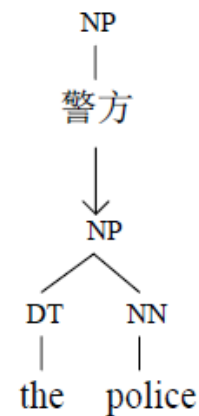
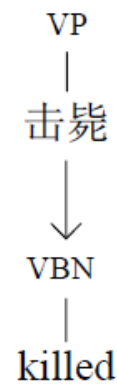
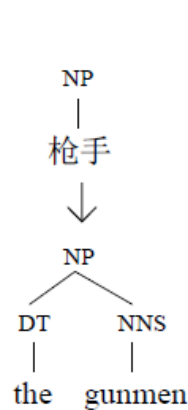
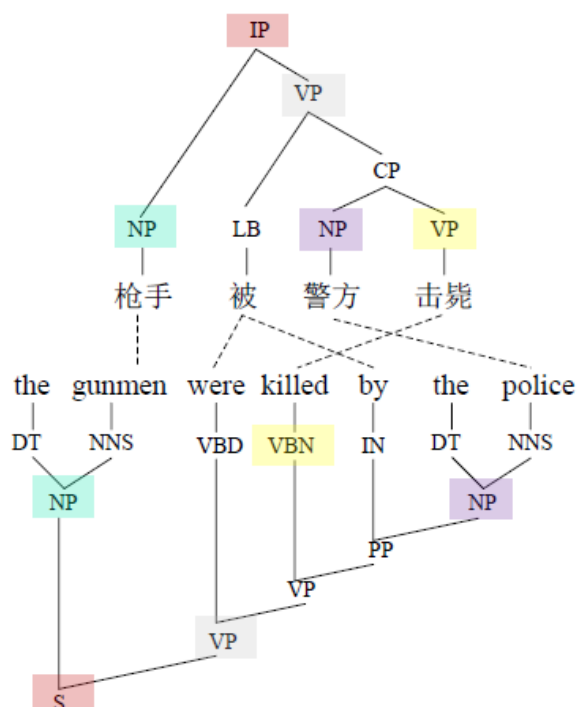
(1) 基本规则(Initial rule): 元树的叶子节点都是终结符(单词)的规则不具有泛化能力。

(2) 泛化规则(Abstract rule): 元树存在非终结符的叶子节点的规则具有泛化能力。

**在抽取过程中, 先对基本规则进行抽取,
然后基于基本规则再进一步生成泛化规则**

(1) 树到树模型

如：



引自：宗成庆：《自然语言理解》讲义，第 11 章

(1) 树到树模型

□ 树到树模型的优势

- 搜索空间小、解码效率高

□ 树到树模型的不足

- 强烈依赖于源语言和目标语言句法分析的质量
- 利用两端句法结构精确匹配，数据稀疏非常严重
- 翻译质量差

内 容 提 要

- 附录. 1 基于词的统计机器翻译方法
- 附录. 2 基于短语的统计机器翻译方法
- 附录. 3 基于层次化短语的统计机器翻译方法
- 附录. 4 基于树的统计机器翻译方法
 - (1) 树到树模型
 - (2) 树到串模型
 - (3) 串到树模型
- 附录. 5 译文评估方法

(2) 树到串模型

树到串的翻译模型

Yang Liu (ACL2006) 提出了树到串的翻译模型

特点:

- 在源语言端进行句法分析
- 在目标语言端不进行句法分析
- 从源语言端句法分析和词语对齐的语料库中抽取翻译规则
- 递归地将源语言句子的句法结构树转换为目标语言句子
(树到串转换)

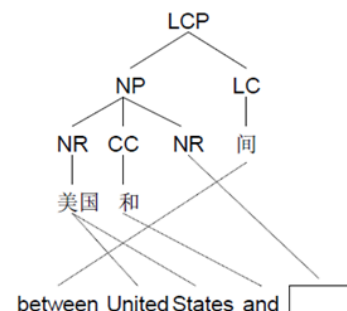
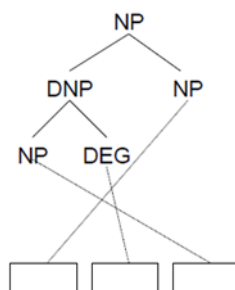
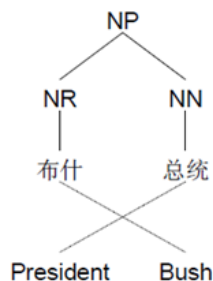
涉及内容

1. 树到串翻译模板 (规则)
2. 树到串模型翻译过程 (解码)
3. 树到串翻译规则 的学习

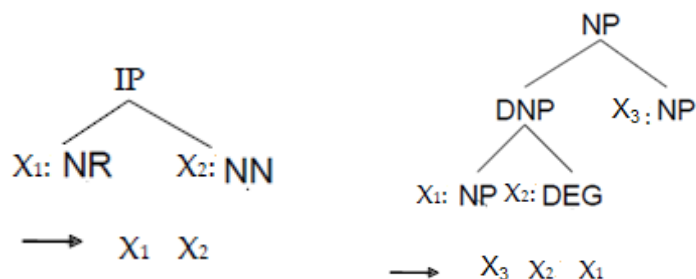
1. 树到串翻译模板（规则）

(2) 树到串模型

树到串翻译模板 (规则)



带词对齐关系的 树-串 句对



树-串翻译模板

树到串对齐模板 (简称TAT) 既可以生成终结符也可以生成非终结符, 既可以执行局部重排序也可以执行全局重排序; 从经过词语对齐和源语言句法分析的双语语料库上自底向上自动抽取

2. 树到串模型翻译过程（解码）

(2) 树到串模型

第一步，加载翻译模型（文法规则集）和语言模型

第二步，读入源语言句法树 $T(f)$

第三步，获取可用的规则集

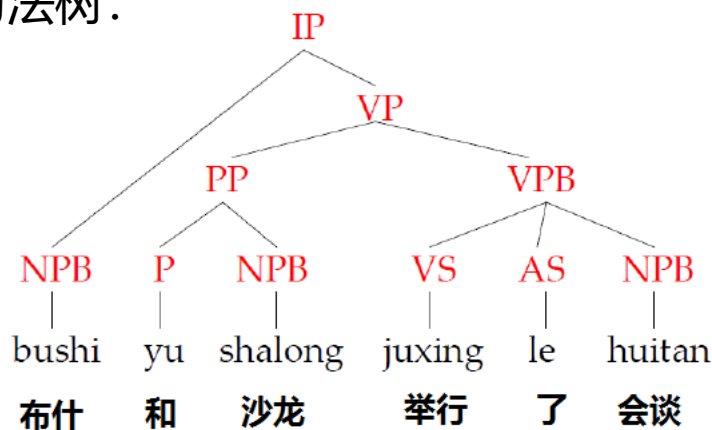
第四步，进行从顶向下树到串转化（过程称为栈搜索）

第五步，将最优译文输出

(2) 树到串模型

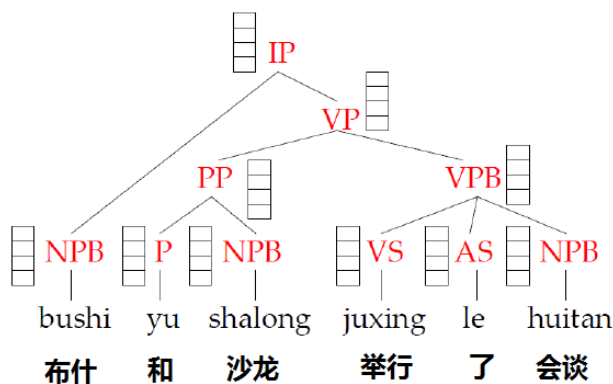
例： 源语句子： 布什 和 沙龙 举行 了 会谈

输入句法树：



规则集和
语言模型

栈搜索



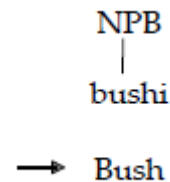
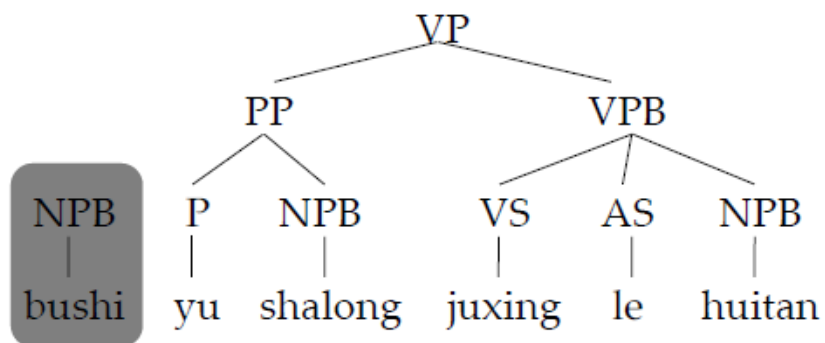
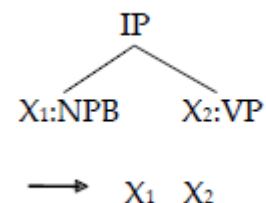
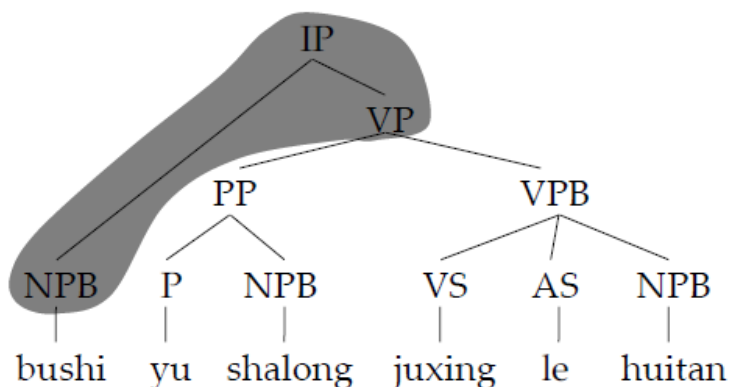
自顶向下

- 柱搜索 (Beam Search)
- 对于每一棵子树，找到所有与其根节点匹配的TAT，计算其候选译文 (Candidate)

(2) 树到串模型

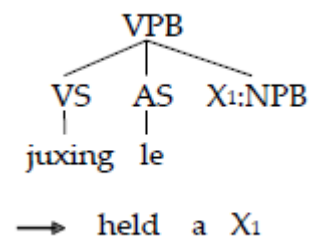
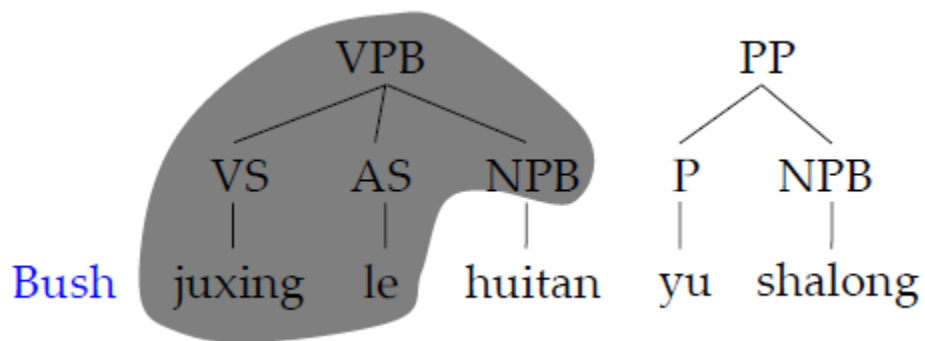
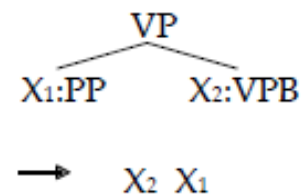
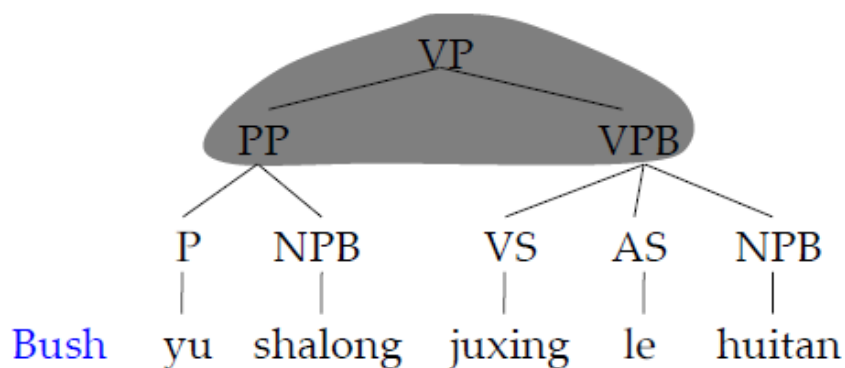
搜索过程:

规则集和
语言模型



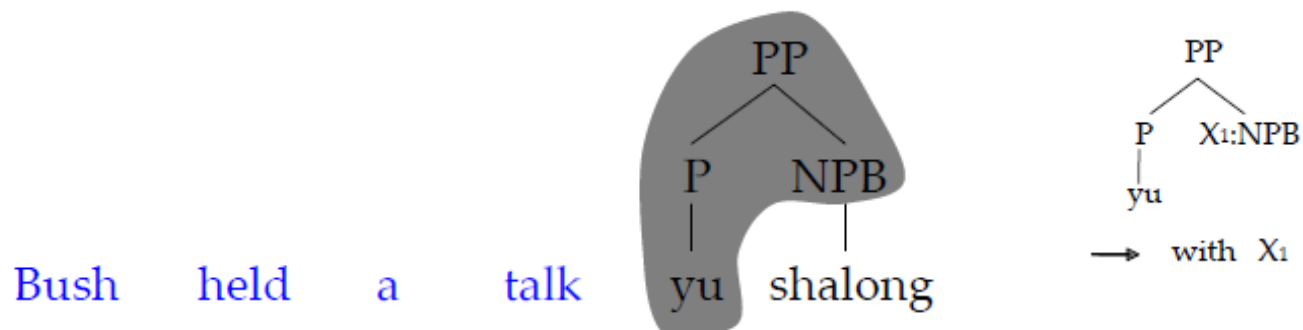
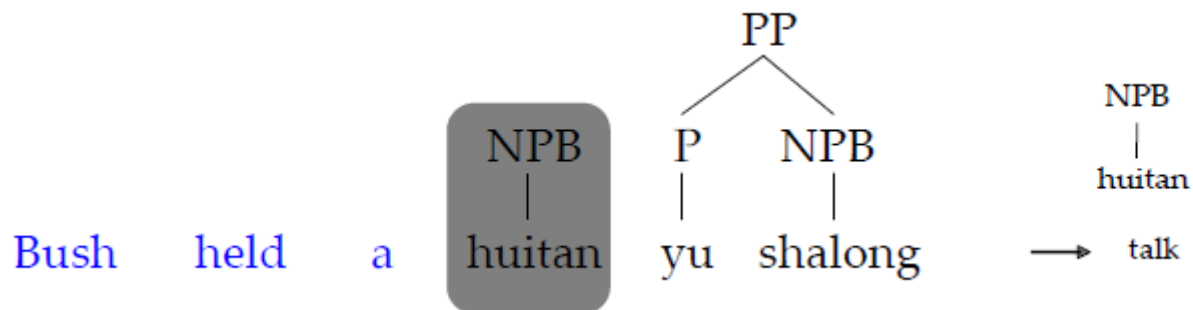
(2) 树到串模型

规则集和
语言模型



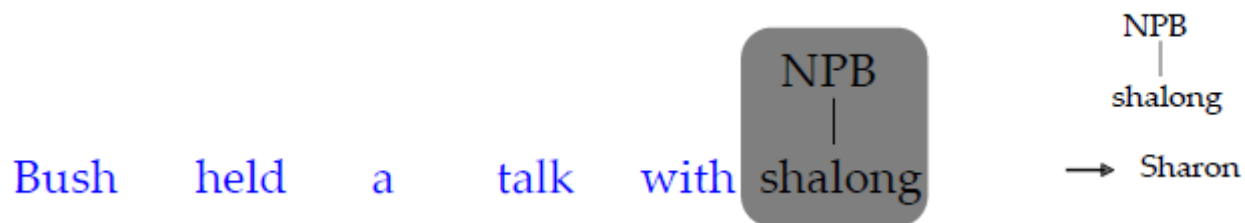
(2) 树到串模型

规则集和
语言模型



(2) 树到串模型

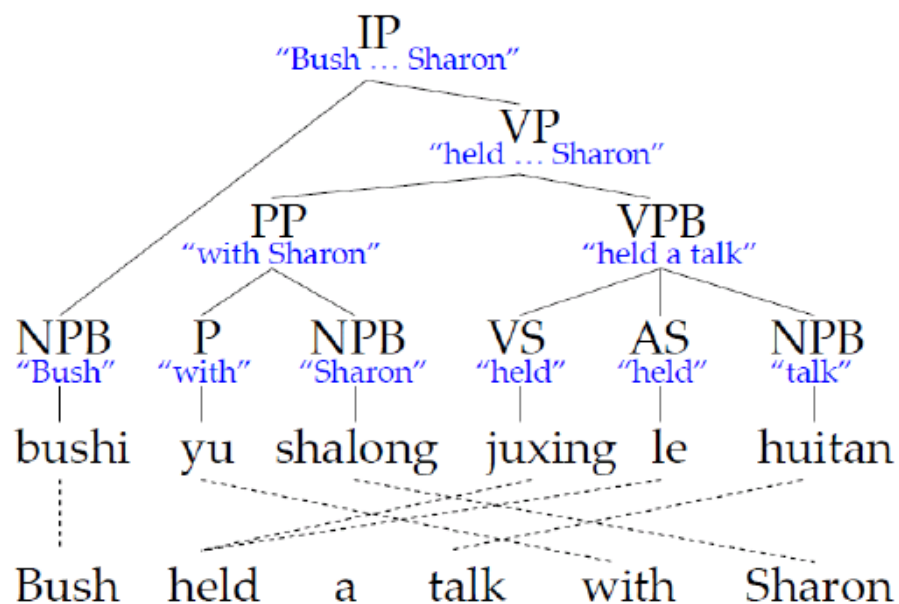
规则集和
语言模型



Bush held a talk with Sharon

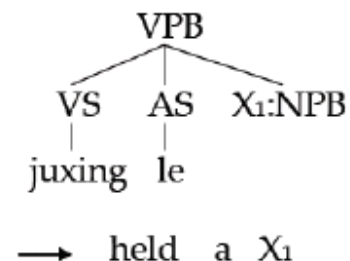
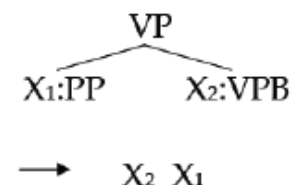
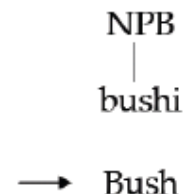
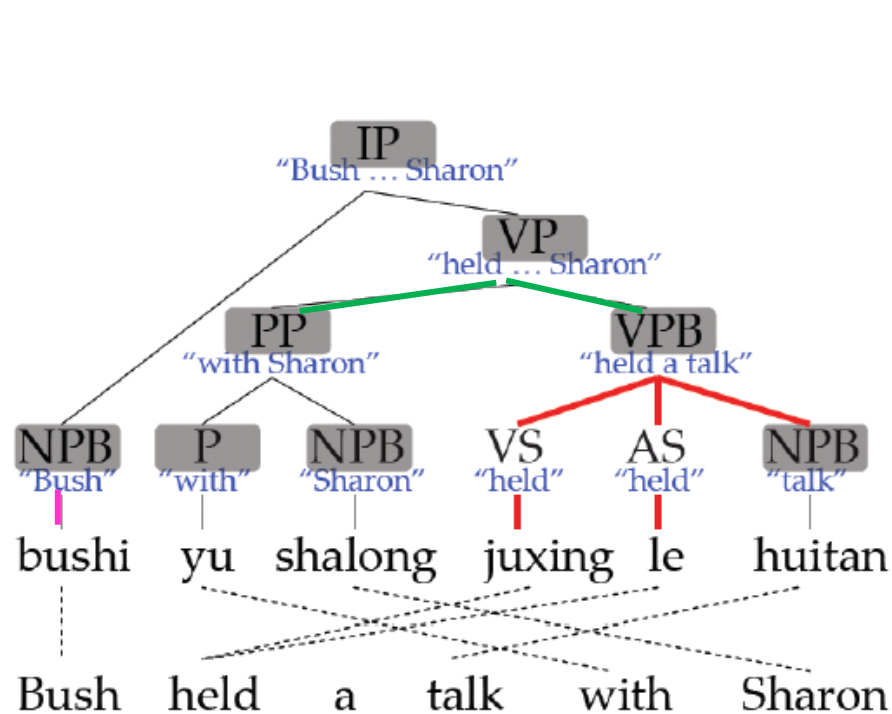
3. 树到串翻译规则的学习

(2) 树到串模型



树-串 对齐句对

(2) 树到串模型



(2) 树到串模型

□ 树到串模型的优势

- 搜索空间小、解码效率高
- 句法分析质量较高的前提下，翻译效果不错

□ 树到串模型的不足

- 强烈依赖于源语言句法分析的质量
- 利用源语言端句法结构精确匹配，数据稀疏严重
- 没有使用任何目标语言句法知识，无法保证目标译文符合文法

内 容 提 要

- 附录. 1 基于词的统计机器翻译方法
- 附录. 2 基于短语的统计机器翻译方法
- 附录. 3 基于层次化短语的统计机器翻译方法
- 附录. 4 基于树的统计机器翻译方法
 - (1) 树到树模型
 - (2) 树到串模型
 - (3) 串到树模型
- 附录. 5 译文评估方法

(3) 串到树模型

串到树的翻译模型

Galley et al.(2004, 2006) , 提出了串到树的翻译模型

特点:

- ❑ 在源语言端进行不句法分析
- ❑ 在目标语言端进行句法分析
- ❑ 从目标语言端句法分析和词语对齐的语料库中抽取翻译规则并构造翻译模型
- ❑ 利用串到树转换规则, 将源语言句子分析为一棵目标语言句法结构树, 拼接叶结点得到译文

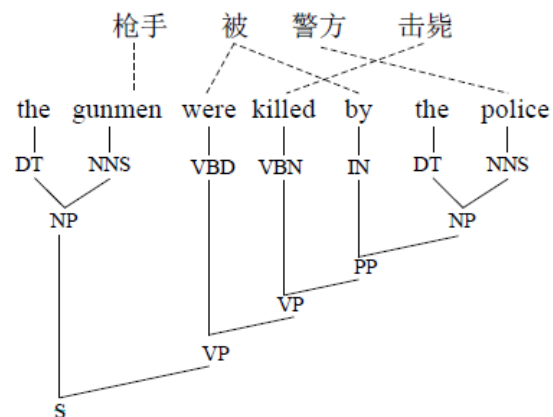
涉及内容

1. 串到树翻译模板 (规则)
2. 串到树模型翻译过程 (解码)
3. 串到树翻译规则 的学习

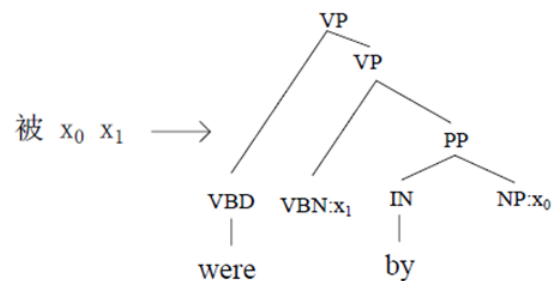
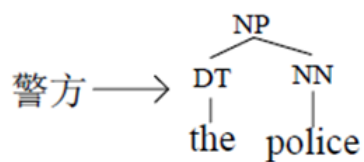
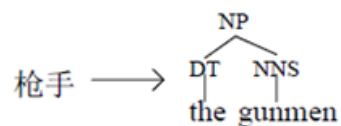
1. 串到树翻译模板（规则）

(3) 串到树模型

串到树翻译模板 (规则)



带词对齐关系的 串-树句对



串-树翻译模板

2.串到树模型翻译过程（解码）

(3) 串到树模型

第一步，加载翻译模型（文法规则集）和语言模型

第二步，读入源语言句子词串

第三步，获取可用的规则集

第四步，进行从底向上串到树转化

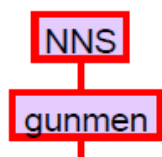
第五步，将最优译文输出

(3) 串到树模型

例： 源语句子： 枪手 被 警方 击毙 。

规则集和
语言模型

栈搜索



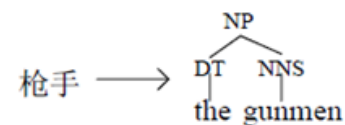
枪手

被

警方

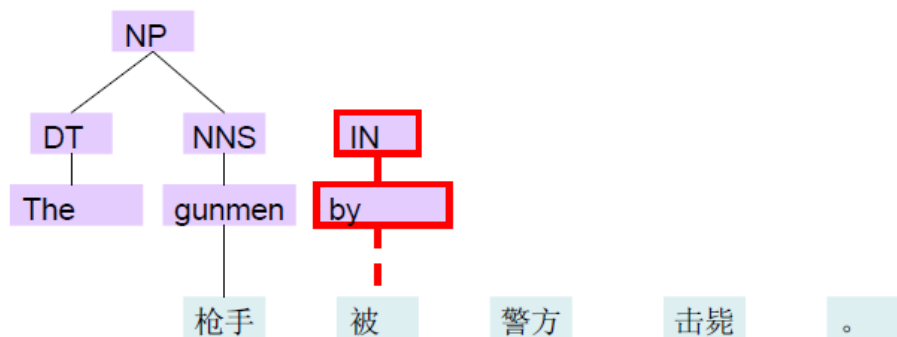
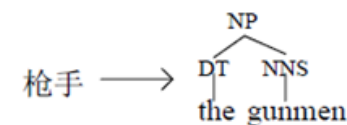
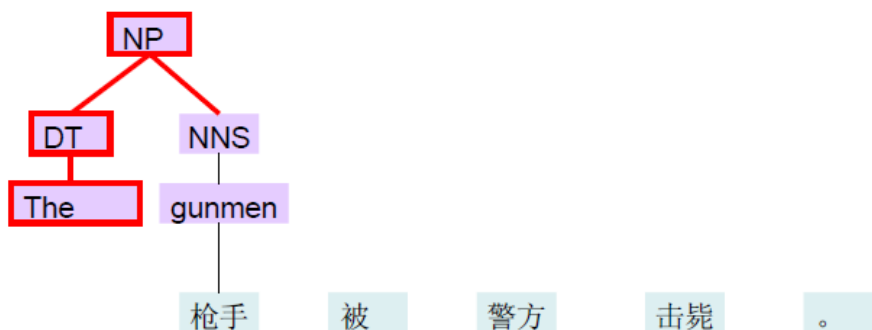
击毙

。



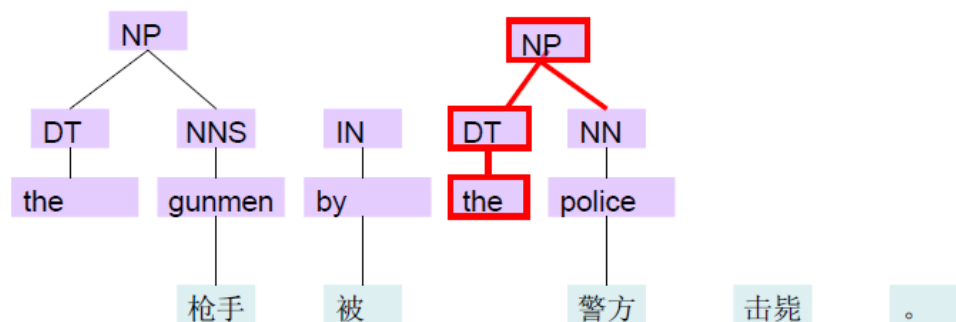
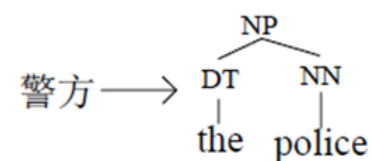
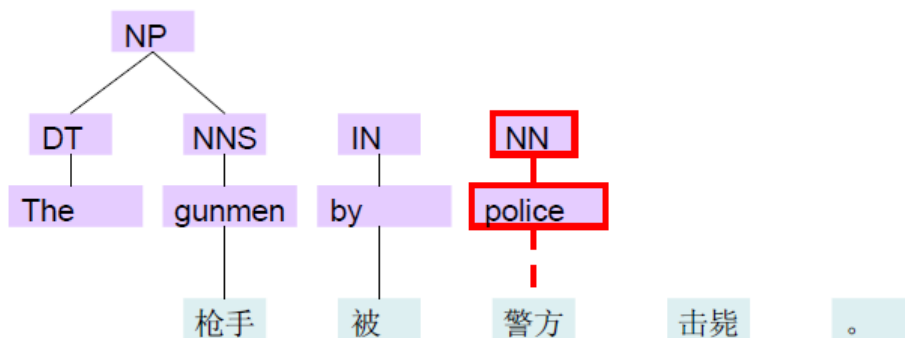
(3) 串到树模型

规则集和
语言模型



(3) 串到树模型

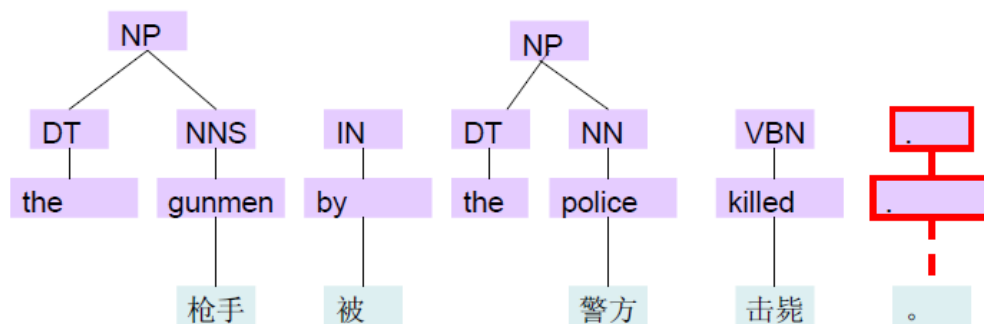
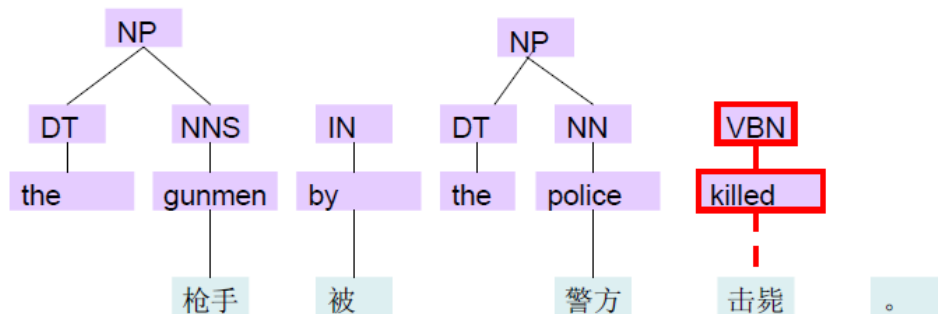
规则集和
语言模型



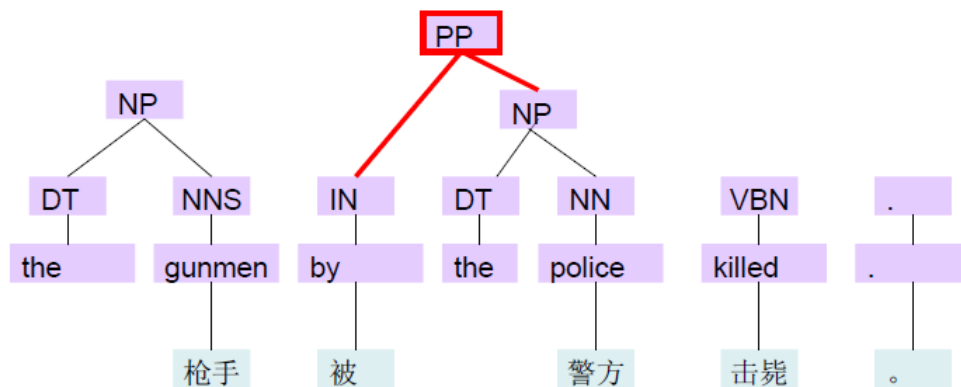
引自：刘群：机器翻译原理与方法(07) 基于句法的机器翻译方法

(3) 串到树模型

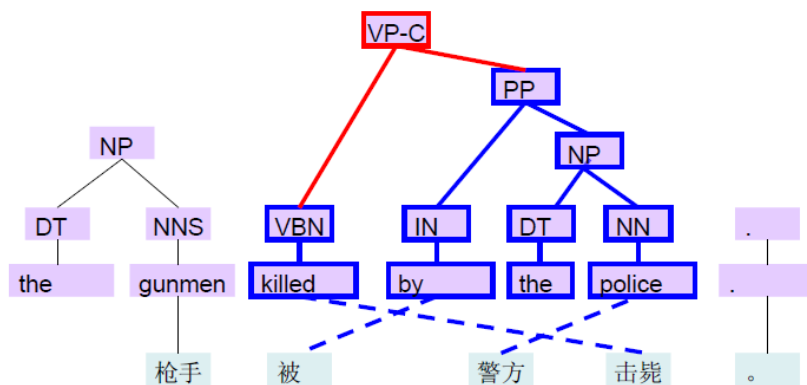
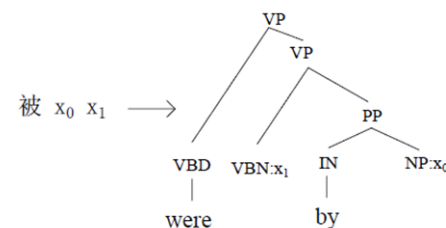
规则集和
语言模型



(3) 串到树模型

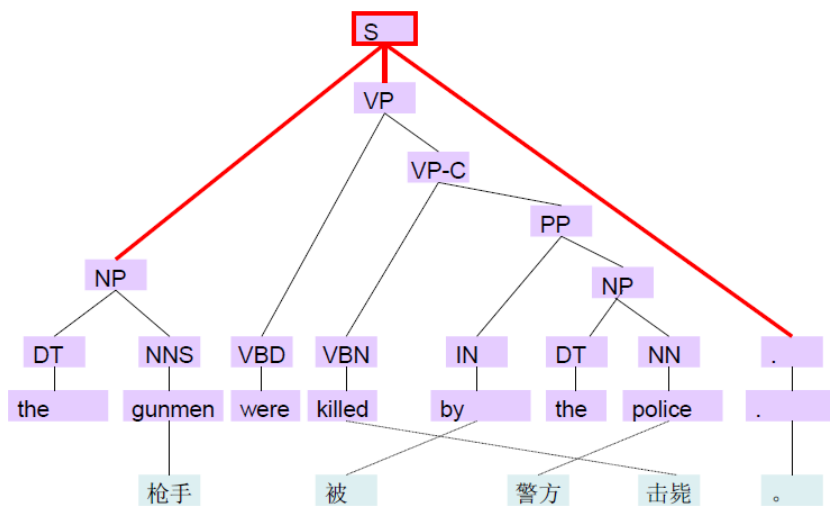
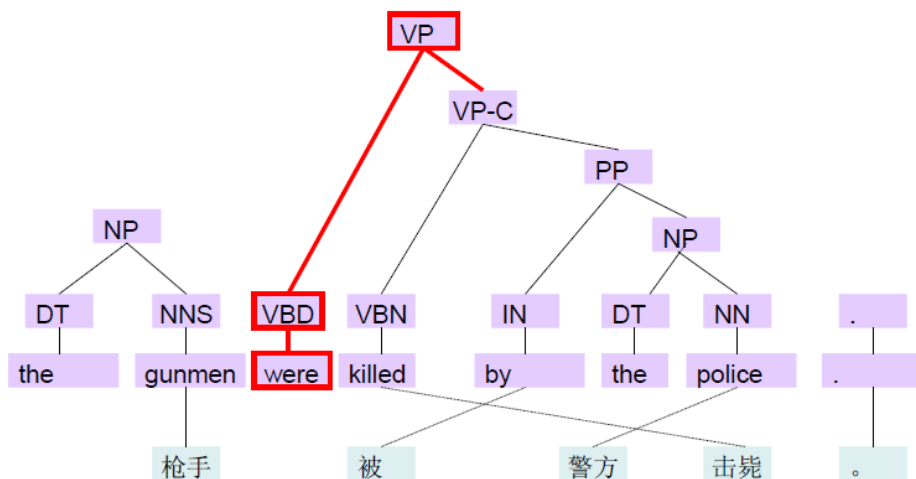


规则集和
语言模型



(3) 串到树模型

规则集和
语言模型

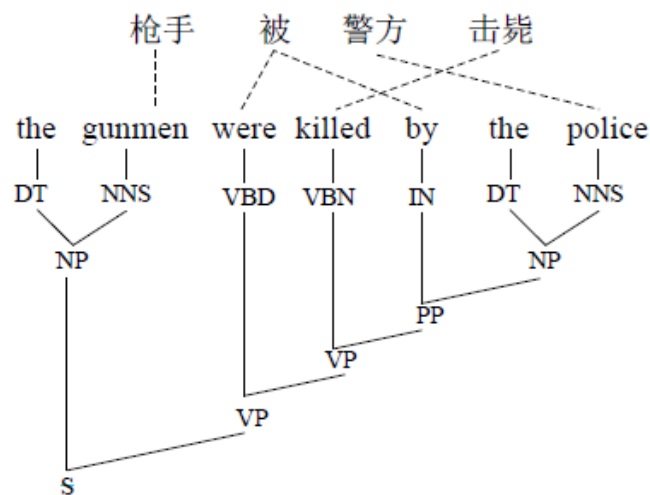


引自：刘群：机器翻译原理与方法(07) 基于句法的机器翻译方法

3.串到树翻译规则的学习

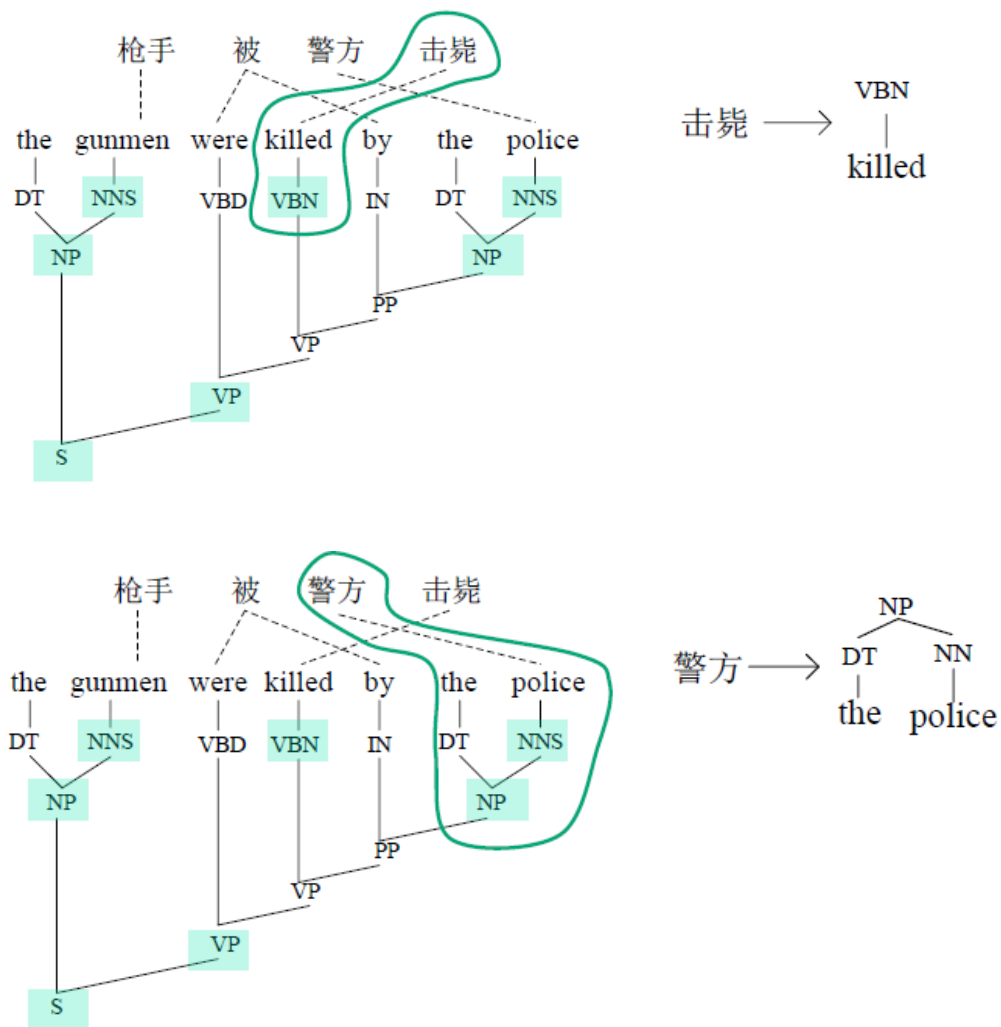
(3) 串到树模型

确定满足词语对齐的树节点： 目标语言句法树节点所能到达的
源语言子串与该树节点覆盖的目标语言子串满足词语对齐约束

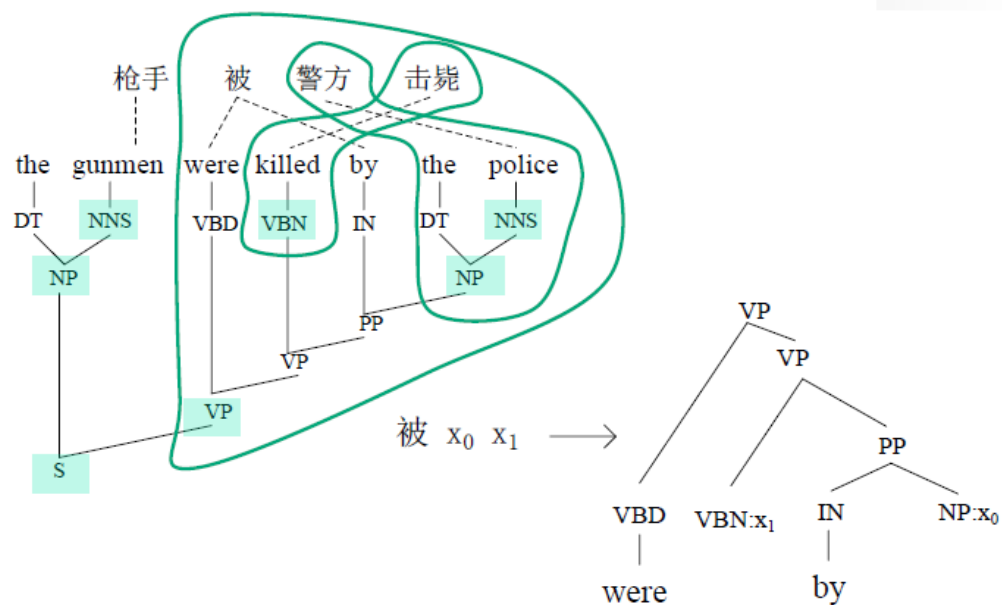


带词对齐关系的 串-树句对

(3) 串到树模型



(3) 串到树模型



(3) 串到树模型

□ 串到树模型的优势

- 搜索空间大，保证译文符合语法，翻译质量高

□ 串到树模型的不足

- 解码速度受限
- 未使用源语言端句法知识，存在词义消歧问题

内 容 提 要

- 附录. 1 基于词的统计机器翻译方法
- 附录. 2 基于短语的统计机器翻译方法
- 附录. 3 基于层次化短语的统计机器翻译方法
- 附录. 4 基于树的统计机器翻译方法
- 附录. 5 译文评估方法

附录.5 译文评估方法

常用的评测指标

□ 主观评测:

- (1) 流畅度
- (2) 充分性
- (3) 语义保持性

□ 自动评测:

由评测系统依据一定的数学模型对译文句子自动计算得分。

常用的自动打分方法有：

附录.5 译文评估方法

- **BLEU 评价方法**
- **NIST评测方法**
- **mWER 方法**
- **GTM方法**
- **METEOR 评测：**

参考文献:

宗成庆, 统计自然语言处理 (第2版) 课件

刘群, 机器翻译原理与方法讲义 (05) -(07)

张浩, 面向短语统计机器翻译解码算法的研究,
硕士学位论文,2012

梁华参, 基于短语的统计机器翻译模型训练中若干关键问题的研究,
博士学位论文, 2013

马永亮, 层次短语翻译模型的实现与分析, 硕士学位论文, 2011

蒋宏飞, 基于同步树替换文法的统计机器翻译方法研究,
博士学位论文, 2010

在此表示感谢!

附录：统计机器翻译

完