

统计机器翻译的 研究现状与发展趋势

刘群

liuqun@ict.ac.cn

2009-10-10 于郑州YOCSEF



中科院计算所

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
——基于词的**IBM**模型
- 最成熟的统计机器翻译方法
——基于短语的模型
- 目前统计机器翻译研究的热点
——基于句法的模型
- 中科院计算所的工作
- 统计机器翻译面临的问题和展望

统计机器翻译的研究热潮

- 历史回顾：一些重要事件回放
- 一种新的研究范式
- 统计机器翻译论文发表数量的增长
- 近年来国际机器翻译评测的最好成绩
- 统计机器翻译目前的水平

历史回顾：一些重要事件回放 (1)

- **1980年代末IBM**首次开展统计机器翻译研究
- **1992年IBM**首次提出统计机器翻译的信源信道模型
- **1993年IBM**提出五种基于词的统计翻译模型**IBM Model 1-5**
- **1994年IBM**发表论文给出了**Candide**系统与**Systran**系统在**ARPA**评测中的对比测试报告
- **1999年JHU**夏季研讨班重复了**IBM**的工作并推出了开放源代码的工具
- **2001年IBM**提出了机器翻译自动评测方法**BLEU**
- **2002年NIST**开始举行每年一度的机器翻译评测
- **2002年**第一个采用统计机器翻译方法的商业公司**Language Weaver**成立

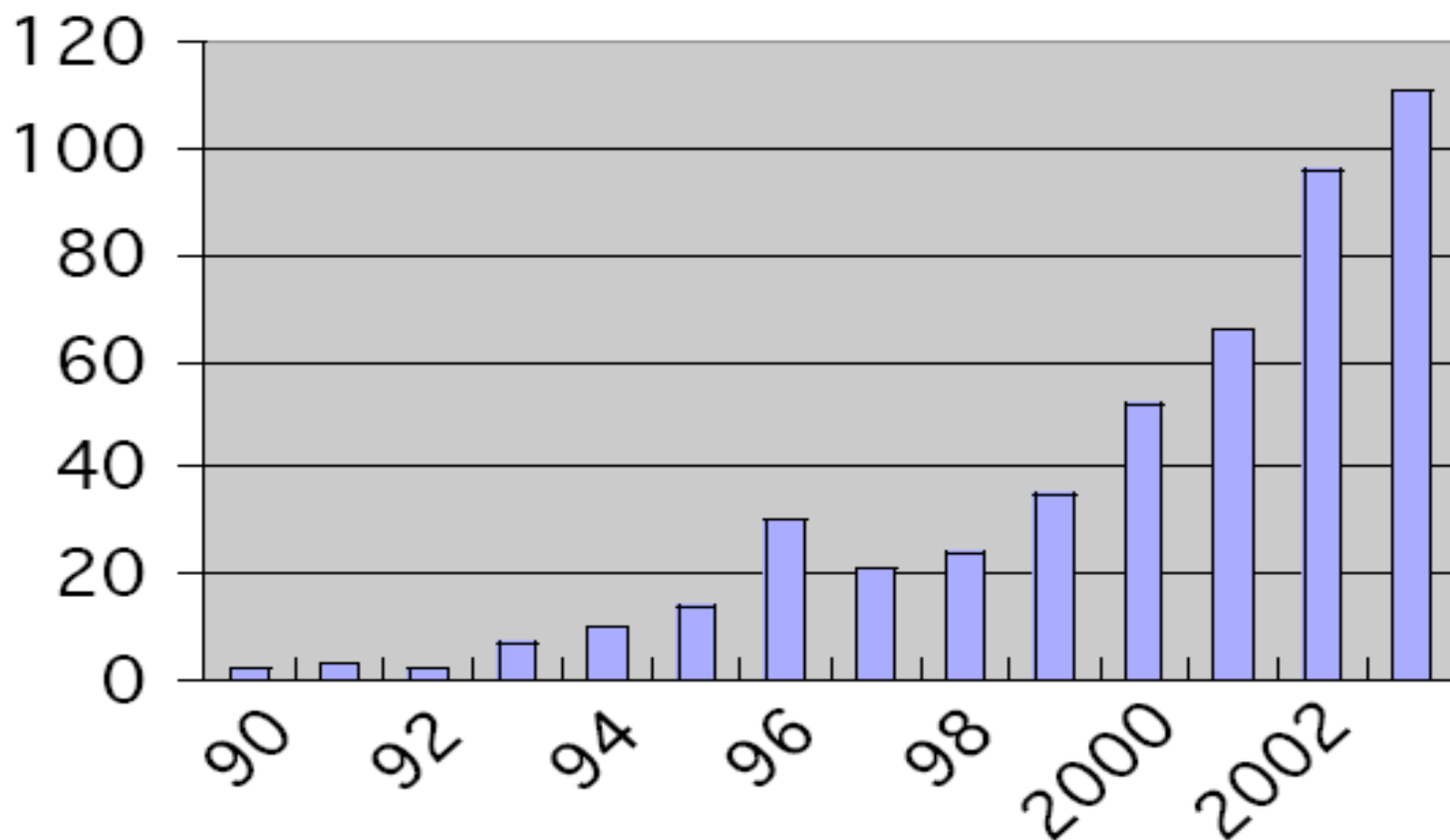
历史回顾：一些重要事件回放 (2)

- **2002年Franz Josef Och**提出统计机器翻译的对数线性模型
- **2003年Franz Josef Och**提出对数线性模型的最小错误率训练方法
- **2004年Philipp Koehn**推出**Pharaoh**（法老）标志着基于短语的统计翻译方法趋于成熟
- **2005年David Chiang**提出层次短语模型并代表**UMD**在**NIST**评测中取得好成绩
- **2005年Google**在**NIST**评测中大获全胜，随后**Google**推出基于统计方法的在线翻译工具，其阿拉伯语-英语的翻译达到了用户完全可接受的水平，目前已经可以支持**50**多种语言的互译
- **2006年NIST**评测中**USC-ISI**的树到串句法模型第一次超过**Google**（仅在汉英受限翻译项目中）

统计机器翻译：一种新的研究范式

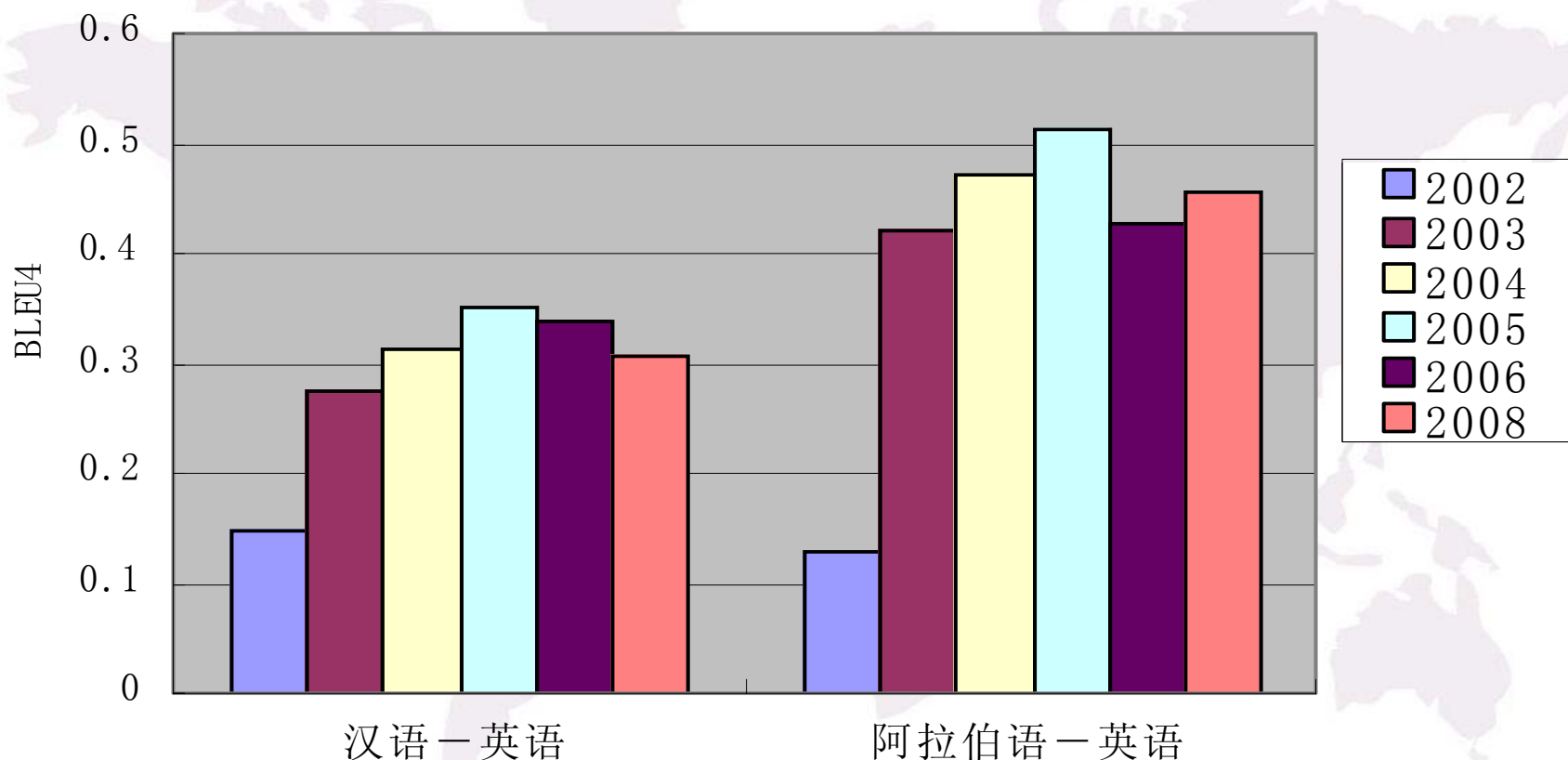
- 统计机器翻译的成功在于采用了一种新的研究范式（**paradigm**）
- 这种研究范式已在语音识别等领域中被证明是一种成功的翻译，但在机器翻译中是首次使用
- 这种范式的特点：
 - 公开的大规模的训练数据
 - 周期性的公开评测和研讨
 - 开放源码的工具

近年来统计机器翻译论文发表数量



引自 Franz Josef Och, Statistical Machine Translation: Foundations and Recent Advances, Tutorials on MT Summit X, September 13-15, 2005, Phuket, Thailand

近年来国际NIST评测最好成绩



统计机器翻译目前的水平

- 以**Google Translator**为例，实地考察一下统计机器翻译的水平
 - 阿拉伯语—英语
 - 汉语—英语
 - 英语—汉语

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
—基于词的**IBM**模型
- 最成熟的统计机器翻译方法
—基于短语的模型
- 目前统计机器翻译研究的热点
—基于句法的模型
- 中科院计算所的工作
- 统计机器翻译面临的问题和展望

基于词的统计机器翻译方法

- 统计机器翻译—为翻译建立概率模型
- **IBM**的信源信道模型
- 语言模型—**n**元语法模型
- 翻译模型—**IBM**模型1-5
- 搜索算法
- **Candide**系统

为翻译建立概率模型

- 假设任意一个英语句子 **e** 和一个法语句子 **f**, 我们定义 **f** 翻译成 **e** 的概率为:

$$\Pr(e | f)$$

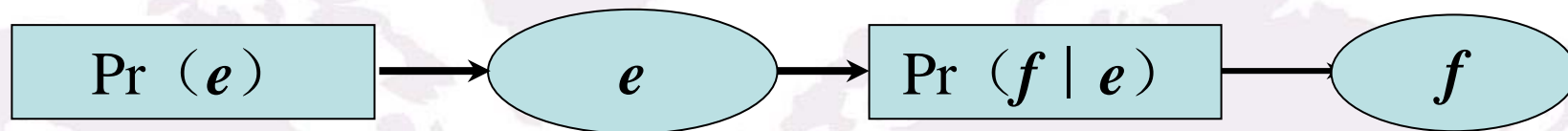
其归一化条件为:

$$\sum_e \Pr(e | f) = 1$$

- 于是将 **f** 翻译成 **e** 的问题就变成求解问题:

$$\hat{e} = \operatorname{argmax}_e \Pr(e | f)$$

信源信道模型 (1)



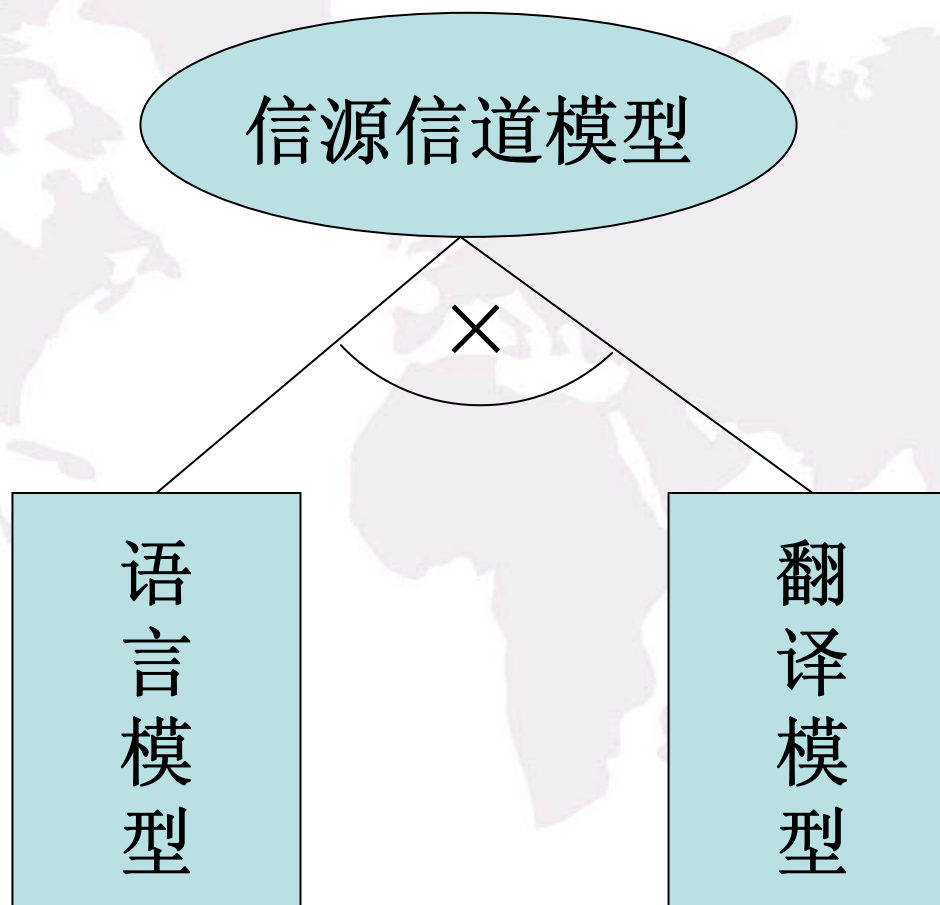
- 假设我们看到的源语言文本 F 是由一段目标语言文本 E 经过某种奇怪的编码得到的，那么翻译的目标就是要将 F 还原成 E ，这也就是就是一个解码的过程。
- 注意，在信源信道模型中：
 - 噪声信道的源语言是翻译的目标语言
 - 噪声信道的目标语言是翻译的源语言这与整个机器翻译系统翻译方向的刚好相反

信源信道模型 (2)

$$\hat{e} = \arg \max_e \Pr(e) \Pr(f | e)$$

- **P.Brown**称上式为统计机器翻译基本方程式
 - 语言模型: $P(E)$
 - 翻译模型: $P(F|E)$
- 语言模型反映“**E**像一个句子”的程度: 流利度
- 翻译模型反映“**F**像**E**”的程度: 忠实度
- 联合使用两个模型效果好于单独使用翻译模型, 因为后者容易导致一些不好的译文。

信源信道模型 (3)



信源信道模型 (4)

- 统计机器翻译分解为以下三个问题：
 - 语言模型的定义和参数估计
 - 翻译模型的定义和参数估计
 - 解码

语言模型 — n 元语法模型

- 语言模型在机器翻译中具有极为重要的作用
- 到目前位置，统计机器翻译中最常用、而且最有效的模型仍然是 n 元语法模型
- 模型的阶数越来越高：**3元、4元、5元**
- 模型的训练语料越来越大：
 - **Google**提供了公开的**Web 1T**语料库，其中的 n 元共现词频数据是从**web**中得到的**1T**英文词的语料库中统计得到的（剪切掉了低频组合）
 - **Google**号称使用了**2T**英文词训练的语言模型
 - 大规模的数据为系统实现带来很大的困难

翻译模型

- 翻译模型 $P(\mathbf{F}|\mathbf{E})$ 反映的是一个源语言句子 \mathbf{E} 翻译成一个目标语言句子 \mathbf{F} 的概率
- 由于源语言句子和目标语言句子几乎不可能在语料库中出现过，因此这个概率无法直接从语料库统计得到，必须分解成词语翻译的概率和句子结构（或者顺序）翻译的概率

翻译模型与对齐

- 翻译模型的计算，需要引入隐含变量：
对齐 **A**:

$$P(F|E) = \sum_A P(F, A|E)$$

- 翻译概率 $P(F|E)$ 的计算转化为对齐概率 $P(F, A|E)$ 的估计
- 对齐：建立源语言句子和目标语言句子的词与词之间的对应关系和句子结构之间的对应关系

词语对齐的表示 (1)

● 图形表示

- ✓ 连线
- ✓ 矩阵（见下页）

● 数字表示

- ✓ 给每个目标语言单词标记其所有对应的源语言单词



词语对齐的表示 (2)

achievement										
economic										
marked										
cities										
board										
open										
14										
's										
China										
	中国	十四	个	边境	开放	城市	经济	建设	成就	显著

IBM Model 1

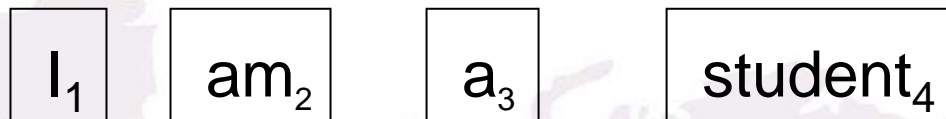
- 最简单的理解，可以句子 e 翻译成 f 的概率，就是 e 中每一个词语翻译成 f 中对应词语的概率的乘积
- 这就是**IBM Model 1**的基本思想
- **IBM**提出了复杂度递增的**5**个统计翻译模型，**IBM Model 1**是其中最简单的模型

IBM Model 1-5

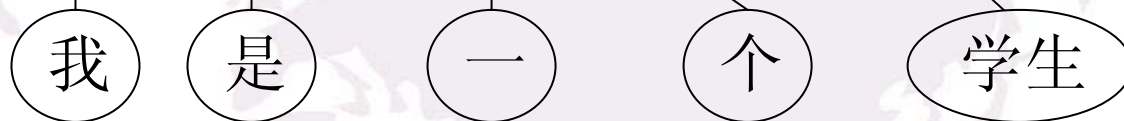
- **IBM Model 1**仅考虑词对词的互译概率
- **IBM Model 2**加入了词的位置变化的概率
- **IBM Model 3**加入了一个词翻译成多个词的概率
- **IBM Model 4**
- **IBM Model 5**

IBM Model 1 & 2 推导方式 (1)

源语言句子E：



目标语言句子F：



词语对齐A：

1 2 3 3 4

IBM模型1&2的推导过程：

1. 猜测目标语言句子长度；
2. 从左至右，对于每个目标语言单词：
 - 首先猜测该单词由哪一个源语言单词翻译而来；
 - 再猜测该单词应该翻译成什么目标语言词。

IBM Model 1 & 2 推导方式 (2)

假设翻译的目标语言句子为: $F = f_1^m = f_1 f_2 \cdots f_m$

假设翻译的源语言句子为: $E = e_1^l = e_1 e_2 \cdots e_l$

假设词语对齐表示为:

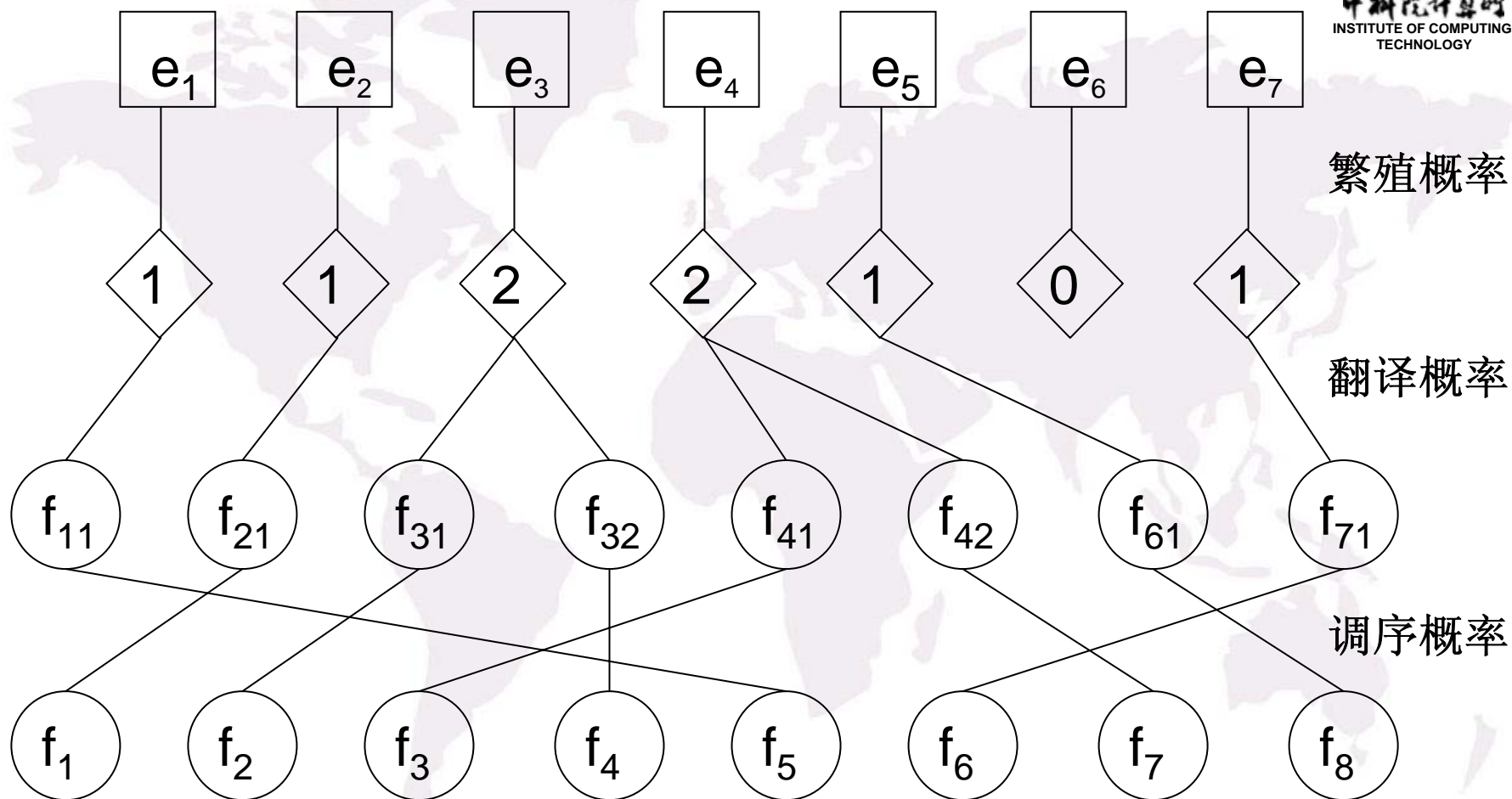
$$A = a_1^m = a_1 a_2 \cdots a_m, \forall i \in \{1, \cdots, m\}, a_i \in \{0, \cdots, l\}$$

那么词语对齐的概率可以表示为:

$$\Pr(F, A | E) = \Pr(m | E) \prod_{j=1}^m \Pr(a_j | a_1^{j-1}, f_1^{j-1}, m, E) \Pr(f_j | a_1^j, f_1^{j-1}, m, E)$$

注意: 在**IBM Model**中, 词语对齐只考虑了源语言到目标语言的单向一对多形式, 不考虑多对一和多对多的形式。

IBM Model 3 & 4 & 5 推导方式 (1)



IBM Model 3 & 4 & 5 推导方式 (2)



1. 首先根据源语言词语的繁殖概率，确定每个源语言词翻译成多少个目标语言词；
2. 根据每个源语言词语的目标语言词数，将每个源语言词复制若干次；
3. 将复制后得到的每个源语言词，根据翻译概率，翻译成一个目标语言词；
4. 根据调序概率，将翻译得到的目标语言词重新调整顺序，得到目标语言句子。

IBM模型的参数训练：EM算法

- **EM**参数训练算法是经典的无指导学习的算法：
 1. 给定初始参数；
 2. **E**步骤：用已有的参数计算每一个句子对的所有可能的对齐的概率；
 3. **M**步骤：用得到的所有对齐的概率重新计算参数；
 4. 重复执行**E**步骤和**M**步骤，直到收敛。
- 由于**EM**算法的**E**步骤需要穷尽所有可能的对齐，通常这会带来极大的计算量，除非我们可以对计算公式进行化简（就像前面**IBM Model 1**所做的那样），否则这种计算量通常是不可承受的。

IBM模型的参数训练

- **IBM Model 1**
 - 任何初始值均可达到全局最优
- **IBM Model 2~5:**
 - 存在大量局部最优，任意给定的初值很容易导致局部最优，而无法到达全局最优的结果
 - **IBM的训练策略:**
 - 依次训练**IBM Model 1-5**
 - 对于与上一级模型相同的参数初始值，直接取上一个模型训练的结果；
 - 对于新增加的参数，取任意初始值。

IBM Model 1的EM训练示例 (0)

我们用一个简单的例子来演示EM训练的过程

- 假设有两个句子对: $(a\ b|x\ y)$ 和 $(a\ y)$
- 先假设所有词语翻译概率平均分布 $P(f|e)$:

$\Pr(a x)$	1/2	$\Pr(a y)$	1/2
$\Pr(b x)$	1/2	$\Pr(b y)$	1/2

我们这里为方便起见, 对**IBM Model 1**做了简化:

- 只考虑词语一对一的情况, 不考虑词语一对多或者对齐到空的情况;
- 对齐概率计算的时候, 忽略了词语长度和词语对齐概率, 仅考虑词语翻译概率。




IBM Model 1的EM训练示例(1E)

	对所有可能的对齐 计算 $P(F, A E)$	对 $P(F, A E)$ 归一化 得到 $P(A F, E)$
$\begin{array}{cc} a & b \\ & \\ x & y \end{array}$	$P(F, A E) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$P(A F, E) = \frac{1}{4} / \frac{2}{4} = \frac{1}{2}$
$\begin{array}{cc} a & b \\ \diagdown & \diagup \\ x & y \end{array}$	$P(F, A E) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$P(A F, E) = \frac{1}{4} / \frac{2}{4} = \frac{1}{2}$
$\begin{array}{c} a \\ \\ y \end{array}$	$P(F, A E) = \frac{1}{2}$	$P(A F, E) = \frac{1}{2} / \frac{1}{2} = 1$

IBM Model 1的EM训练示例(1M)

计算 $c(f e)$	重新计算 $\Pr(f e)$
$c(a x) = \frac{1}{2}$	$\Pr(a x) = \frac{1}{2} / (\frac{1}{2} + \frac{1}{2}) = \frac{1}{2}$
$c(b x) = \frac{1}{2}$	$\Pr(b x) = \frac{1}{2} / (\frac{1}{2} + \frac{1}{2}) = \frac{1}{2}$
$c(a y) = \frac{1}{2} + 1 = \frac{3}{2}$	$\Pr(a y) = \frac{3}{2} / (\frac{3}{2} + \frac{1}{2}) = \frac{3}{4}$
$c(b y) = \frac{1}{2}$	$\Pr(b y) = \frac{1}{2} / (\frac{3}{2} + \frac{1}{2}) = \frac{1}{4}$

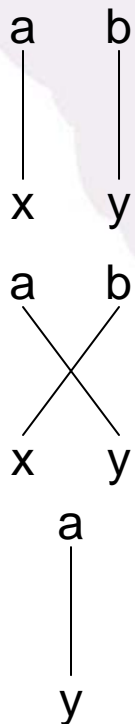
IBM Model 1的EM训练示例(2E)

	对所有可能的对齐 计算 $P(F, A E)$	对 $P(F, A E)$ 归一化 得到 $P(A F, E)$
	$P(F, A E) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	$P(A F, E) = \frac{\frac{1}{8}}{\frac{4}{8}} = \frac{1}{4}$
	$P(F, A E) = \frac{1}{2} \times \frac{3}{4} = \frac{3}{8}$	$P(A F, E) = \frac{\frac{3}{8}}{\frac{4}{8}} = \frac{3}{4}$
	$P(F, A E) = \frac{3}{4}$	$P(A F, E) = \frac{\frac{3}{4}}{\frac{3}{4}} = 1$

IBM Model 1的EM训练示例(2M)

计算 $c(f e)$	重新计算 $\Pr(f e)$
$c(a x) = \frac{1}{4}$	$\Pr(a x) = \frac{1}{4} / (\frac{1}{4} + \frac{3}{4}) = \frac{1}{4}$
$c(b x) = \frac{3}{4}$	$\Pr(b x) = \frac{3}{4} / (\frac{1}{4} + \frac{3}{4}) = \frac{3}{4}$
$c(a y) = \frac{3}{4} + 1 = \frac{7}{4}$	$\Pr(a y) = \frac{7}{4} / (\frac{7}{4} + \frac{1}{4}) = \frac{7}{8}$
$c(b y) = \frac{1}{4}$	$\Pr(b y) = \frac{1}{4} / (\frac{7}{4} + \frac{1}{4}) = \frac{1}{8}$

IBM Model 1的EM训练示例(n)



$$P(A | F, E) = 0.00...1$$

$$\Pr(a | x) = 0.00...1$$

$$\Pr(b | x) = 0.99...9$$

$$P(A | F, E) = 0.99...9$$

$$\Pr(a | y) = 0.99...9$$

$$P(A | F, E) = 1$$

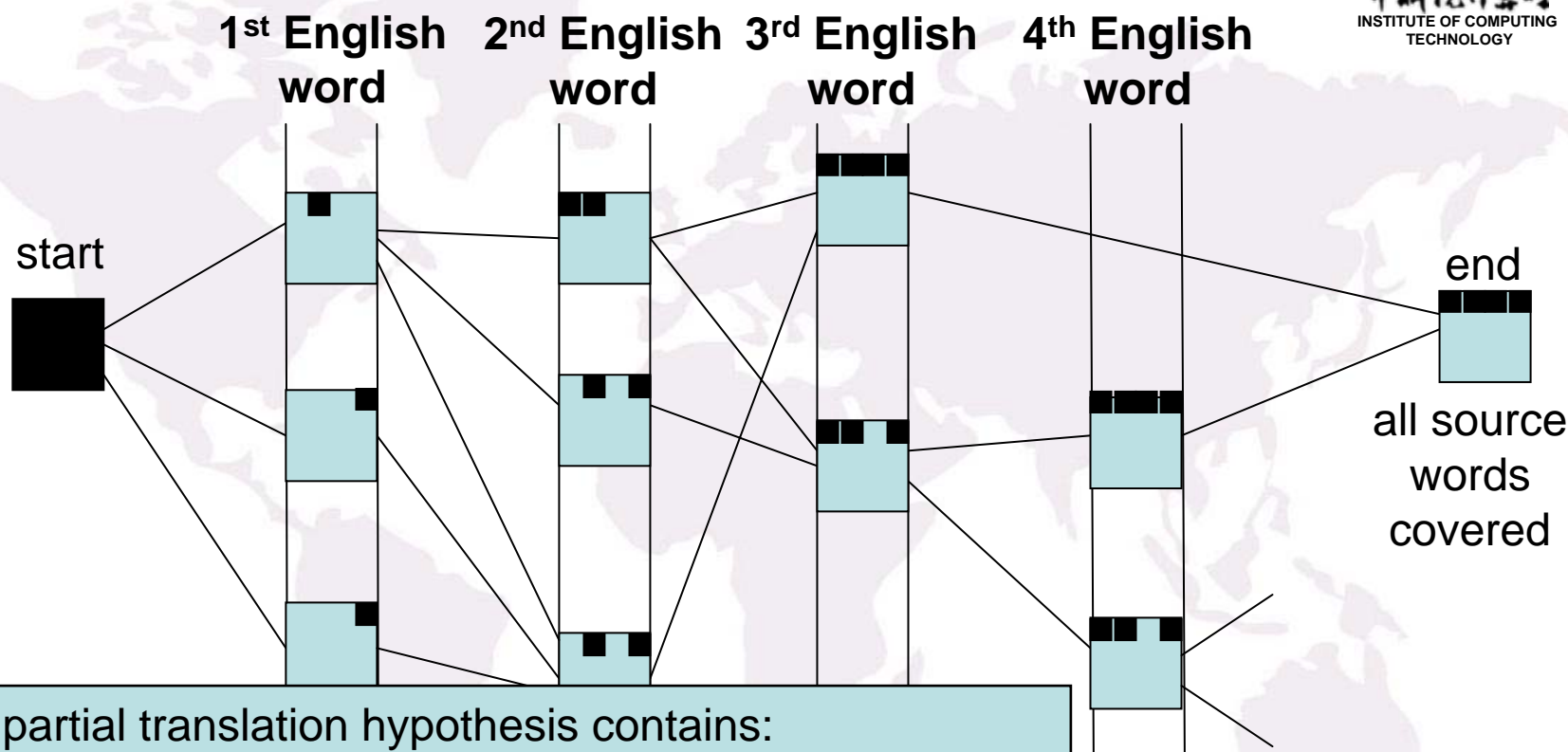
$$\Pr(b | y) = 0.00...1$$

统计机器翻译的解码

- 给定**F**，求**E**，使得 $P(E) * P(F|E)$ 最大
- 解码问题实际上是一个搜索问题，搜索空间巨大，不能保证总能找到全局最优，但通常一些局部最优也是可以接受的
- 如果考虑所有的词语对齐可能性，那么这个问题是一个**NP**完全问题 [**Knight 99**]
- 经典的算法：
 - 单调解码（不调整词序）
 - 堆栈搜索
 - 贪婪算法
 -

堆栈搜索解码算法 (1)

[Brown et al US Patent #5,477,451]



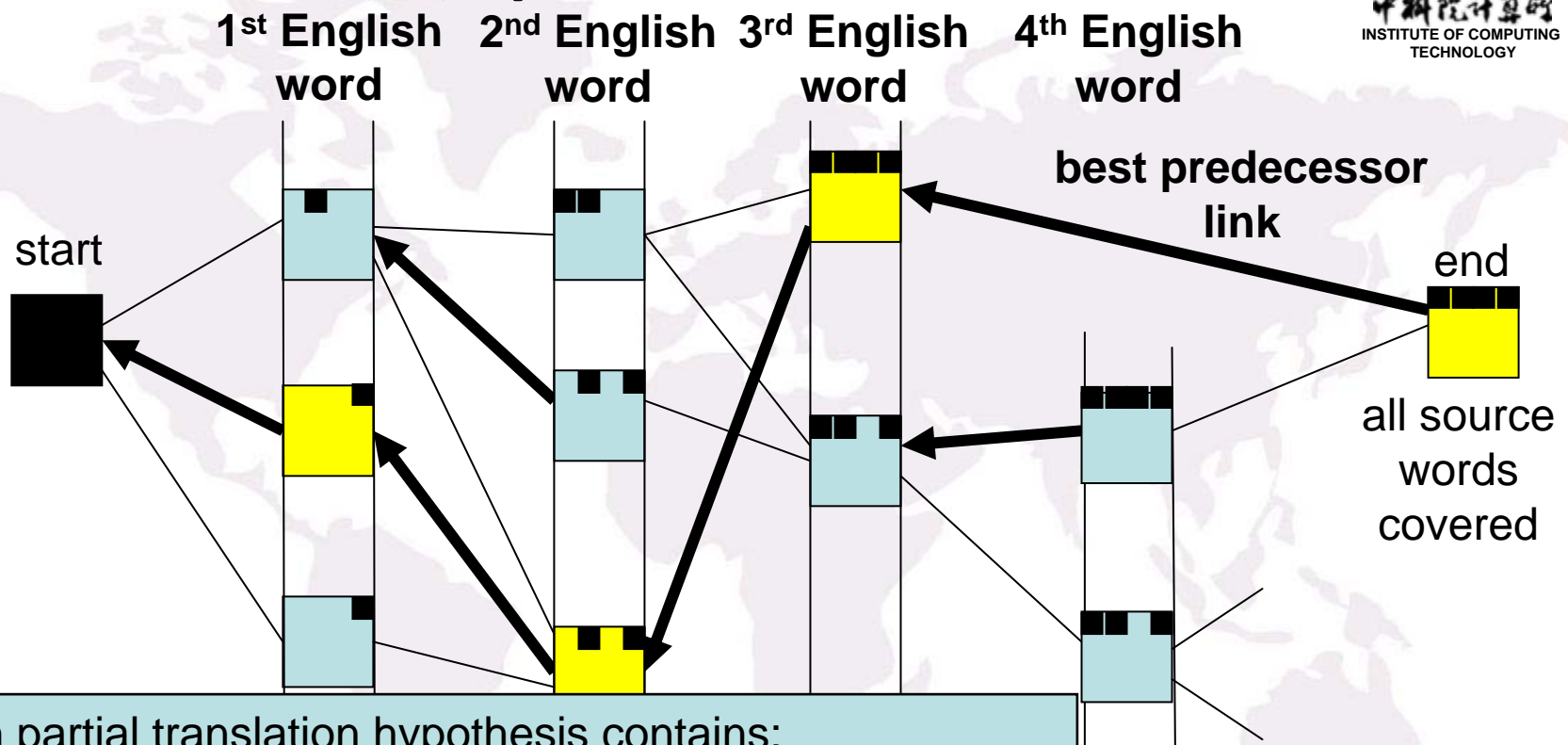
Each partial translation hypothesis contains:

- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence ■ ■ ■
- Language model and translation model scores (so far)

[Jelinek 69;
Och, Ueffing, and Ney, 01]

堆栈搜索解码算法 (2)

[Brown et al US Patent #5,477,451]



Each partial translation hypothesis contains:

- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek 69;
Och, Ueffing, and Ney, 01]

IBM公司的Candide系统(1)

- 基于统计的机器翻译方法
- 分析—转换—生成
 - 中间表示是线性的
 - 分析和生成都是可逆的
- 分析（预处理）：
 1. 短语切分
 2. 专名与数词检测
 3. 大小写与拼写校正
 4. 形态分析
 5. 语言的归一化

IBM公司的Candide系统(2)

- 转换（解码）：基于统计的机器翻译
- 解码分为两个阶段：
 - 第一阶段：使用粗糙模型的堆栈搜索
 - 输出**140**个评分最高的译文
 - 语言模型：三元语法
 - 翻译模型：**EM Trained IBM Model 5**
 - 第二阶段：使用精细模型的扰动搜索
 - 对第一阶段的输出结果先扩充，再重新评分
 - 语言模型：链语法
 - 翻译模型：最大熵翻译模型（选择译文词）

IBM公司的Candide系统(3)

- ARPA的测试结果：

	Fluency		Adequacy		Time Ratio	
	1992	1993	1992	1993	1992	1993
Systran	.466	.540	.686	.743		
Candide	.511	.580	.575	.670		
Transman	.819	.838	.837	.850	.688	.625
Manual		.833		.840		

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
—基于词的**IBM**模型
- 最成熟的统计机器翻译方法
—基于短语的模型
- 目前统计机器翻译研究的热点
—基于句法的模型
- 中科院计算所的工作
- 统计机器翻译面临的问题和展望

基于短语的统计机器翻译方法

- 从信源信道模型到对数线性模型
- 翻译模型的发展—基于短语的模型
- 短语的自动抽取
- 短语翻译概率的计算
- 短语语序的调整
- 几个基于短语的开源系统

统计机器翻译的对数线性模型(1)

- **Och**于**ACL2002**提出，思想来源于**Papineni**提出的基于特征的自然语言理解方法，该论文获得**ACL2002**的最佳论文称号
- 是一个比信源—信道模型更具一般性的模型，信源—信道模型是其一个特例
- 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。
- 与**NLP**中通常使用的最大熵方法的区别：使用连续量（实数）作为特征，而不是使用离散的布尔量（只取**0**和**1**值）作为特征

统计机器翻译的对数线性模型(2)

假设 e 、 f 是机器翻译的目标语言和源语言句子, $h_1(e, f), \dots, h_M(e, f)$ 分别是 e 、 f 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, 那么翻译概率可以用以下公式模拟:

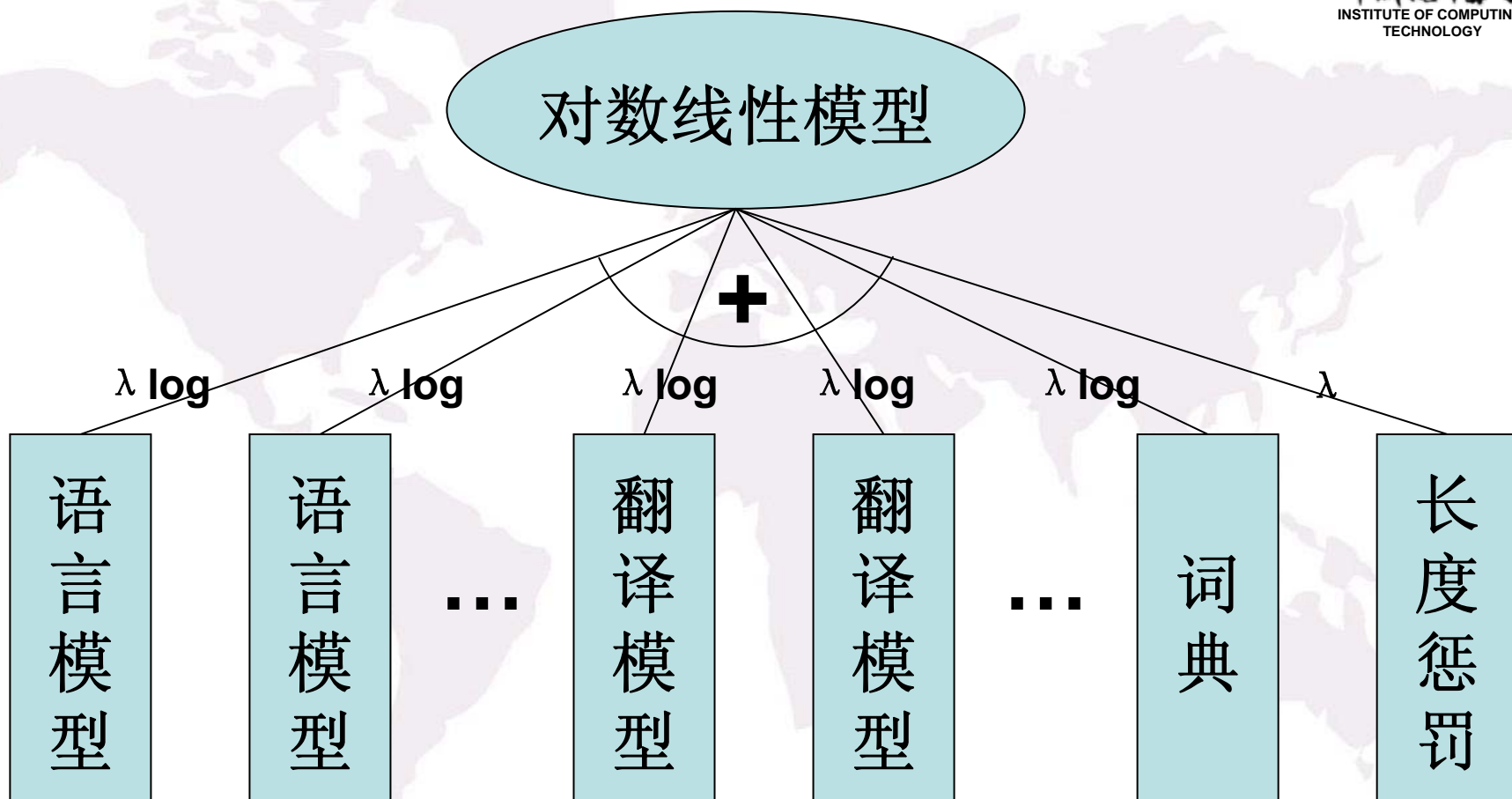
$$\begin{aligned} \Pr(e | f) &\approx p_{\lambda_1 \dots \lambda_M}(e | f) \\ &= \frac{\exp[\sum_{m=1}^M \lambda_m h_m(e, f)]}{\sum_{e'} \exp[\sum_{m=1}^M \lambda_m h_m(e', f)]} \end{aligned}$$

统计机器翻译的对数线性模型(3)

对于给定的 f , 其最佳译文 e 可以用以下公式表示:

$$\begin{aligned}\hat{e} &= \arg \max_e \{ \Pr(e | f) \} \\ &\approx \arg \max_e \left\{ \sum_{m=1}^M \lambda_m h_m(e, f) \right\}\end{aligned}$$

统计机器翻译的对数线性模型(4)



对数线性模型vs.噪声信道模型

- 取以下特征和参数时，对数线性模型等价于噪声信道模型：
 - 仅使用两个特征
 - $h_1(e, f) = \log p(e)$
 - $h_2(e, f) = \log p(f|e)$
 - $\lambda_1 = \lambda_2 = 1$

对数线性模型：Och的实验 (1)

- 方案

- 首先将信源信道模型中的翻译模型换成反向的翻译模型，简化了搜索算法，但翻译系统的性能并没有下降；
- 调整参数 λ_1 和 λ_2 ，系统性能有了较大提高；
- 再依次引入其他一些特征，系统性能又有了更大的提高。

对数线性模型：Och的实验 (2)

- 其他特征
 - 句子长度特征 (WP)：对于产生的每一个目标语言单词进行惩罚；
 - 附加的语言模型特征 (CLM)：一个基于类的语言模型特征；
 - 词典特征 (MX)：计算给定的输入输出句子中有多少词典中存在的共现词对。

对数线性模型：Och的实验 (3)

- 实验结果

	objective criteria [%]					subjective criteria [%]	
	SER	WER	PER	mWER	BLEU	SSER	IER
Baseline($\lambda_m = 1$)	86.9	42.8	33.0	37.7	43.9	35.9	39.0
ME	81.7	40.2	28.7	34.6	49.7	32.5	34.8
ME+WP	80.5	38.6	26.9	32.4	54.1	29.9	32.2
ME+WP+CLM	78.1	38.3	26.9	32.1	55.0	29.1	30.9
ME+WP+CLM+MX	77.8	38.4	26.8	31.9	55.2	28.8	30.9

对数线性模型的优点

- 噪声模型只有在理想的情况下才能达到最优，对于简化的语言模型和翻译模型，取不同的参数值实际效果更好；
- 对数线性模型大大扩充了统计机器翻译的思路；
- 特征的选择更加灵活，可以引入任何可能有用的特征。

对数线性模型的参数训练

- 目的是得到各个特征的参数 $\lambda_1, \dots, \lambda_n$
- 可用的训练算法
 - GIS（最大熵模型的训练算法）
 - 感知机
 - 最小错误率(MER)：直接以评测指标（如BLEU）最好为训练目标
 - 最大互信息(MMI)：把导致总体BLEU值最高的译文定义为好的译文，其他译文定义为不好的译文，进行判别式训练
 - 单纯形算法
- 目前通常使用最小错误率训练算法或单纯形算法



对数线性模型的特征 (1)

- 无论在噪声信道模型还是在~~对数线性模型~~中，~~语言模型和翻译模型~~都是两个最主要的特征
- 对于语言模型，目前主流的做法都还是采用 n 元语法，还没有发现哪些方法能够超过这种简单的模型
- 对于翻译模型，研究者进行了大量的尝试
 - 最早期的**IBM Model 1-5**是基于词的翻译模型
 - 目前最成熟和稳定的模型是基于短语的翻译模型
 - 基于句法的翻译模型近年来也取得了较大进展
- 在对数线性模型中，多个翻译模型和语言模型可以同时使用

对数线性模型的特征 (2)

- 其他特征
 - 词典特征
 - 长度特征：句子单词数。这个特征可以一定程度上避免由于使用语言模型导致的过于偏向短句子的倾向
 -

翻译模型的发展—基于短语的模型

- 基于词的**IBM**翻译模型有明显的缺陷：一个词在翻译的时候基本上不考虑上下文，孤立地进行翻译，导致了大量的错误；词序调整模型近乎无礼，很难准确调整词序，对词序差别较大的语言之间的翻译效果太差。
- 人们很容易想到，将一个短语捆绑起来进行翻译，可以大大提高翻译的准确率
- 很多不同的研究人员尝试了各种各样的基于短语的翻译模型，最终形成了目前比较成熟的基于短语的翻译模型

基于短语的翻译模型 (1)

- 基本思想

- 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
- 这里所说的短语是任意连续的词串，不一定是独立的语言单位
- 翻译的时候将输入的句子与短语词典进行匹配，选择最好的短语划分，将得到的短语译文重新排序，得到最优的译文

- 问题：

- 短语如何抽取？
- 短语概率如何计算？

基于短语的翻译模型 (2)

- 假设 \mathbf{f} 和 \mathbf{e} 之间存在一个短语对齐 \mathbf{B} , 而且这个短语对齐是一一对应的, 那么:

$$\Pr(f_1^J | e_1^I) = \sum_B \Pr(f_1^J, B | e_1^I) = \sum_B \Pr(B | e_1^I) \Pr(f_1^J | B, e_1^I)$$

- 假设短语划分的概率 $\Pr(B | e_1^I)$ 为均匀分布:

$$\Pr(B | e_1^I) = \alpha(e_1^I)$$

- 于是:

$$\Pr(f_1^J | e_1^I) = \alpha(e_1^I) \sum_B \Pr(f_1^J | B, e_1^I)$$

基于短语的翻译模型 (3)

- 假设短语的翻译是互相独立的，并且各种短语顺序调整的概率完全相同，那么：

$$\Pr(f_1^J | e_1^I) = \alpha(e_1^I) \sum_B \prod_k p(\tilde{f}_k | \tilde{e}_k)$$

这里 \tilde{f}_k 和 \tilde{e}_k 是在 B 对齐下源语言和目标语言的短语
而 $\Pr(\tilde{f}_k | \tilde{e}_k)$ 可以通过对短语对齐的语料库统计得到：

$$p(\tilde{f}_k | \tilde{e}_k) = \frac{N(\tilde{f}_k, \tilde{e}_k)}{N(\tilde{e}_k)}$$

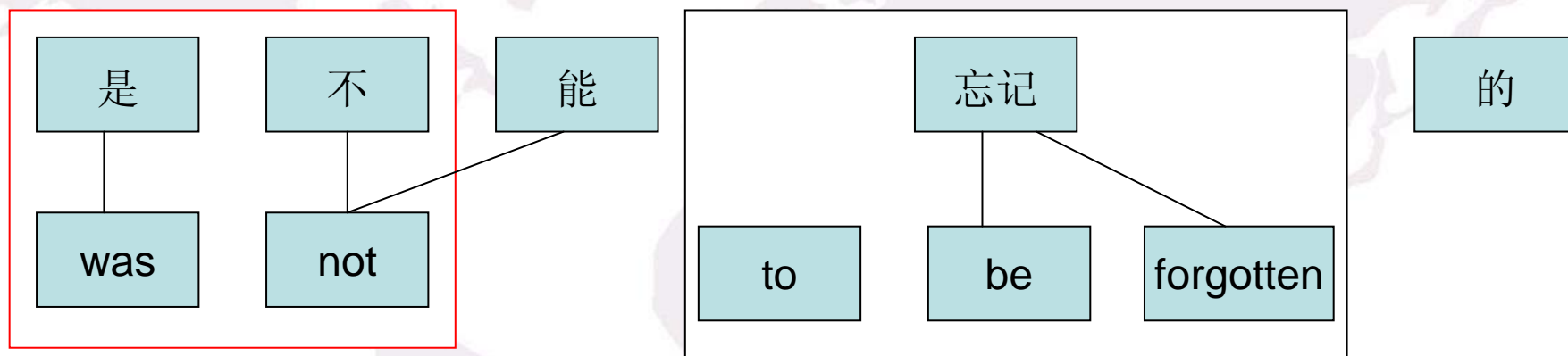


中科院计算所
INSTITUTE OF COMPUTING
TECHNOLOGY

基于短语的翻译模型 (4)

- 实际上，目前在计算短语翻译概率的时候，通常并不去真正生成一个短语对齐的语料库，而是直接从词语对齐的语料库上，去产生所有可能的短语对齐
- 所以，需要先利用**IBM Model**进行词语对齐，但：
 - **IBM Model**只能产生单向一对多的对齐
 - 为了产生更合理的对齐，需要实现多对多对齐，通常的做法是：
 - 先用**IBM Model**对两个方向分别进行一对多对齐
 - 将两个对齐进行某种合并（交集、并集、部分并集），这个操作称为“平衡化”
- 根据词语对齐的结果抽取短语并计算概率

基于词语对齐的短语自动抽取

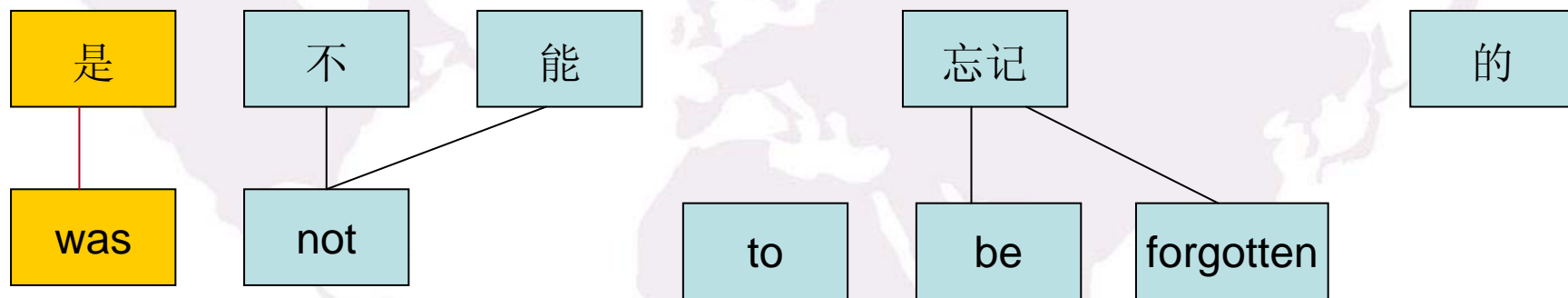


不相容

相容

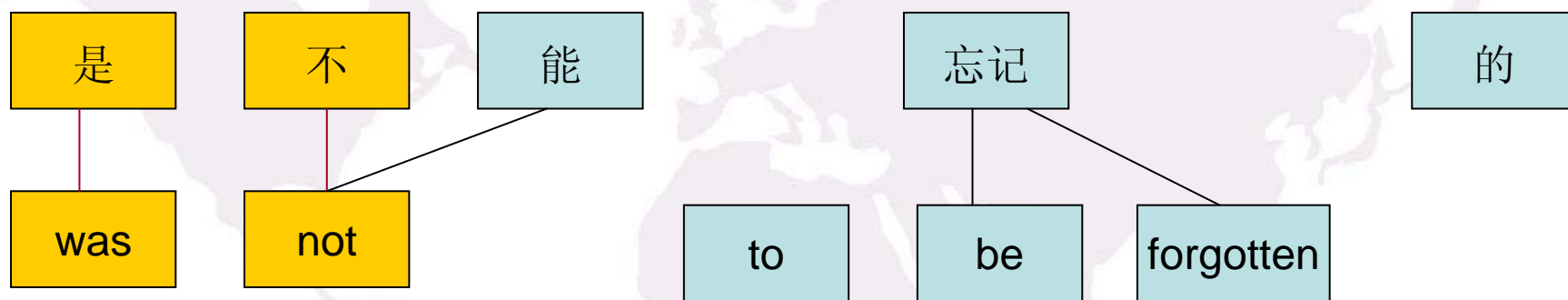
短语自动抽取算法运行示例 (1)

- 列举源语言所有可能的短语，
根据对齐检查相容性



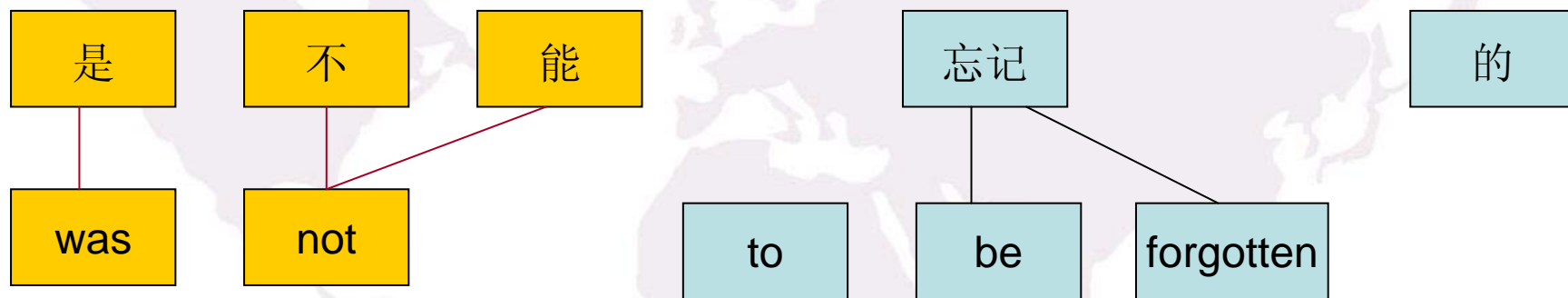
(是, was)

短语自动抽取算法运行示例(2)



不相容

短语自动抽取算法运行示例(3)



(是不能, was not)

短语自动抽取算法运行示例(4)



(是不能, was not to)

短语自动抽取算法运行示例(5)



(是不能忘记, was not to be forgotten)

短语自动抽取算法运行示例(6)



(是不能忘记的, was not to be forgotten)

短语表

- 是
- 是不能
- 是不能
- 是不能忘记
- 是不能忘记的
- 不能
- 不能
- 不能忘记
- 不能忘记的
- 忘记
- 忘记
- 忘记的
- 忘记的

was
was not
was not to
was not to be forgotten
was not to be forgotten
not
not to
not to be forgotten
not to be forgotten
be forgotten
to be forgotten
be forgotten
to be forgotten

短语翻译概率表

f e $p(f|e)$ $lex(f|e)$ $p(e|f)$ $lex(e|f)$

没有达成共识 ||| no consensus was reached ||| 1 0.00210153 1 8.87474e-05 2.718
 没有达成共识。 ||| no consensus was reached . ||| 1 0.0017517 1 8.83361e-05 2.718
 没有得到澄清 ||| clarified ||| 1 0.000592593 1 0.036396 2.718
 没有得到南方的响应 ||| no response ||| 0.5 1.49065e-06 1 0.00921419 2.718
 没有得到证实 ||| no evidence ||| 0.5 0.000178961 1 0.0021538 2.718
 没有兑现 ||| has sent ||| 0.2 6.64599e-05 1 0.00346412 2.718
 没有发生变化 ||| had not changed ||| 0.5 0.000141333 1 8.84361e-05 2.718
 没有发现明显 ||| is no obvious ||| 1 0.00114645 1 0.000308419 2.718
 没有犯罪 ||| no criminal ||| 1 0.0613205 1 0.0251376 2.718
 没有犯罪纪录 ||| no criminal record ||| 1 0.0196688 1 0.0123866 2.718
 没有放弃 ||| has not given up its ||| 1 0.000278368 1 8.34878e-06 2.718
 没有改变 ||| There is no change ||| 0.5 0.0148622 0.333333 1.50262e-05 2.718
 没有改变 ||| has not changed ||| 0.5 0.00505152 0.333333 0.00145408 2.718
 没有改变 ||| is no change ||| 1 0.0283586 0.333333 0.00201351 2.718
 没有改变。 ||| is no change . ||| 1 0.0236378 1 0.00200418 2.718
 没有改变， ||| has not changed , and ||| 1 0.000628986 1 8.72846e-05 2.718
 没有改变， 如果 ||| has not changed , and if ||| 1 0.000308107 1 5.71048e-05 2.718
 没有工作 ||| without work ||| 1 0.0559111 1 0.0130721 2.718
 没有工作许可证 ||| without work permits ||| 1 0.00559111 1 0.000344003 2.718
 没有归还 ||| not repaid till now ||| 1 0.00498227 1 6.22208e-05 2.718
 没有和平 ||| without a peaceful ||| 1 0.0398149 1 7.62298e-06 2.718

短语语序的调整

- 在基于短语的模型中，短语内部的顺序无需调整，只需要调整短语之间的顺序
- 短语的调序模型类似于基于词的模型，允许任意的语序调整
- 为了避免搜索空间的过于膨胀，通常限制语序调整的距离

法老（Pharaoh）

- 由**Philipp Koehn**开发
- 最经典的开源的基于短语的统计机器翻译系统
- 效果远远好于基于词的系统
- 性能稳定
- 推出后很快成为相关研究的基准（**baseline**）
- 缺点：解码器没有开放源代码

丝路（SilkRoad）

- 由国内五家单位联合开发的基于短语的开放源代码的统计机器翻译系统
- 完全开放源代码，包括训练部分和解码部分
- 多个解码器，基本原理与法老类似
- 有完整的中文文档，便于学习

摩西（Moses）

- 最新的开放源代码的基于短语的统计机器翻译系统
- 完全开放源码，包括训练部分和解码部分
- 在基于短语的模型中加入了要素模型（**Factored Model**）
- 采用了词汇化的短语语序调整模型
- 代码优化非常出色
- 性能比法老又有了明显提高

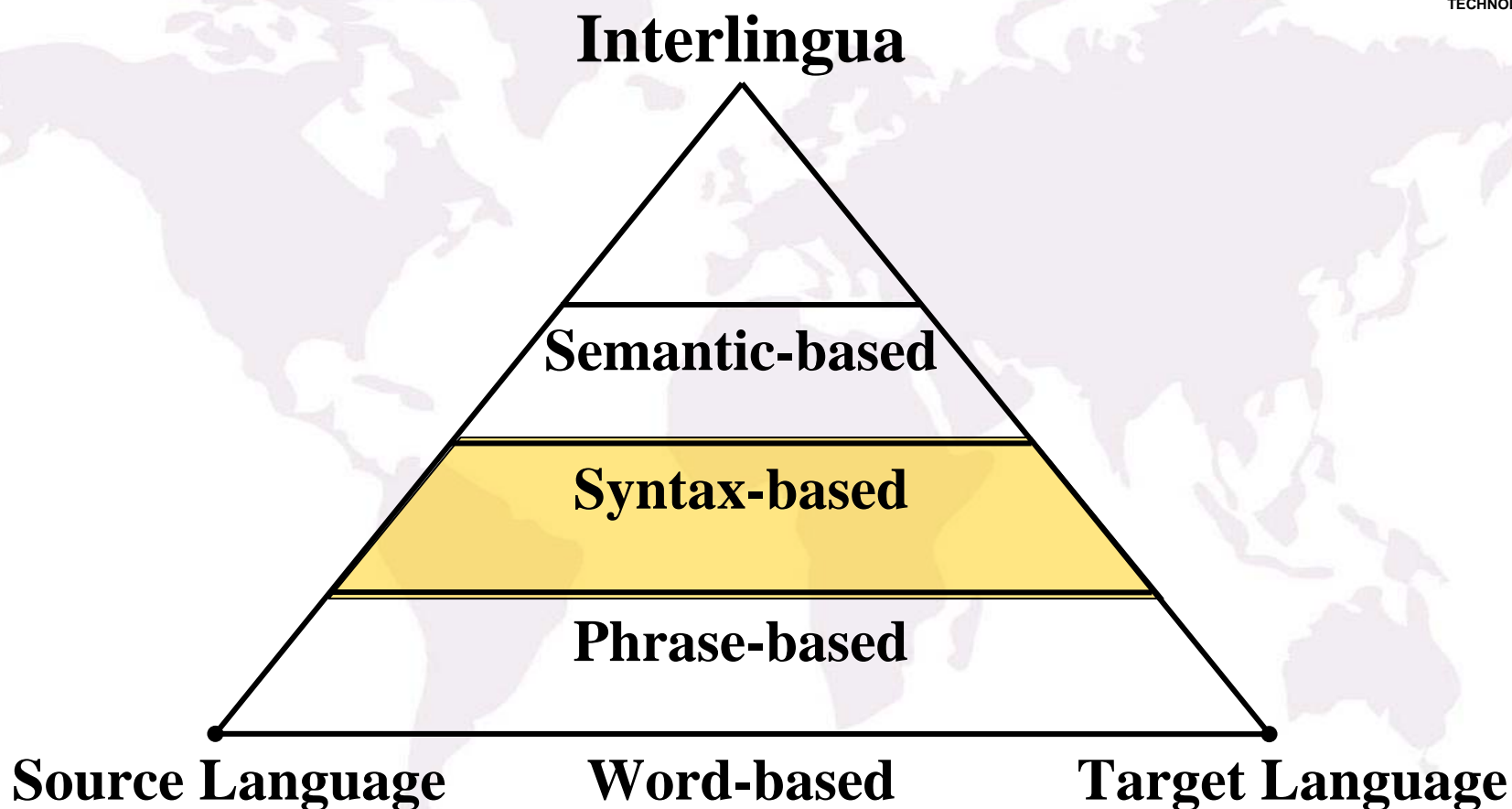
目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
——基于词的**IBM**模型
- 最成熟的统计机器翻译方法
——基于短语的模型
- 目前统计机器翻译研究的热点
——基于句法的模型
- 中科院计算所的工作
- 统计机器翻译面临的问题和展望

基于句法的模型

- 翻译模型的发展—基于句法的模型
- 基于句法的模型概述
- 形式上基于句法的模型
- 语言学上基于句法的模型
- 搜索算法—**CYK**形式的堆栈搜索

统计翻译模型的进展



翻译模型的发展—基于句法的模型

- 基于短语的模型比基于词的模型性能有了较大提高，但对于短语之间的语序调整，仍然没有提供合理的解决方案
- 经验表明，在基于短语的统计机器翻译系统中，绝大多数匹配的短语长度都是**2-3**个词，**1**个词的短语也占相当大的比例
- 要解决长距离语序的调整，引入句法信息是个必然的选择

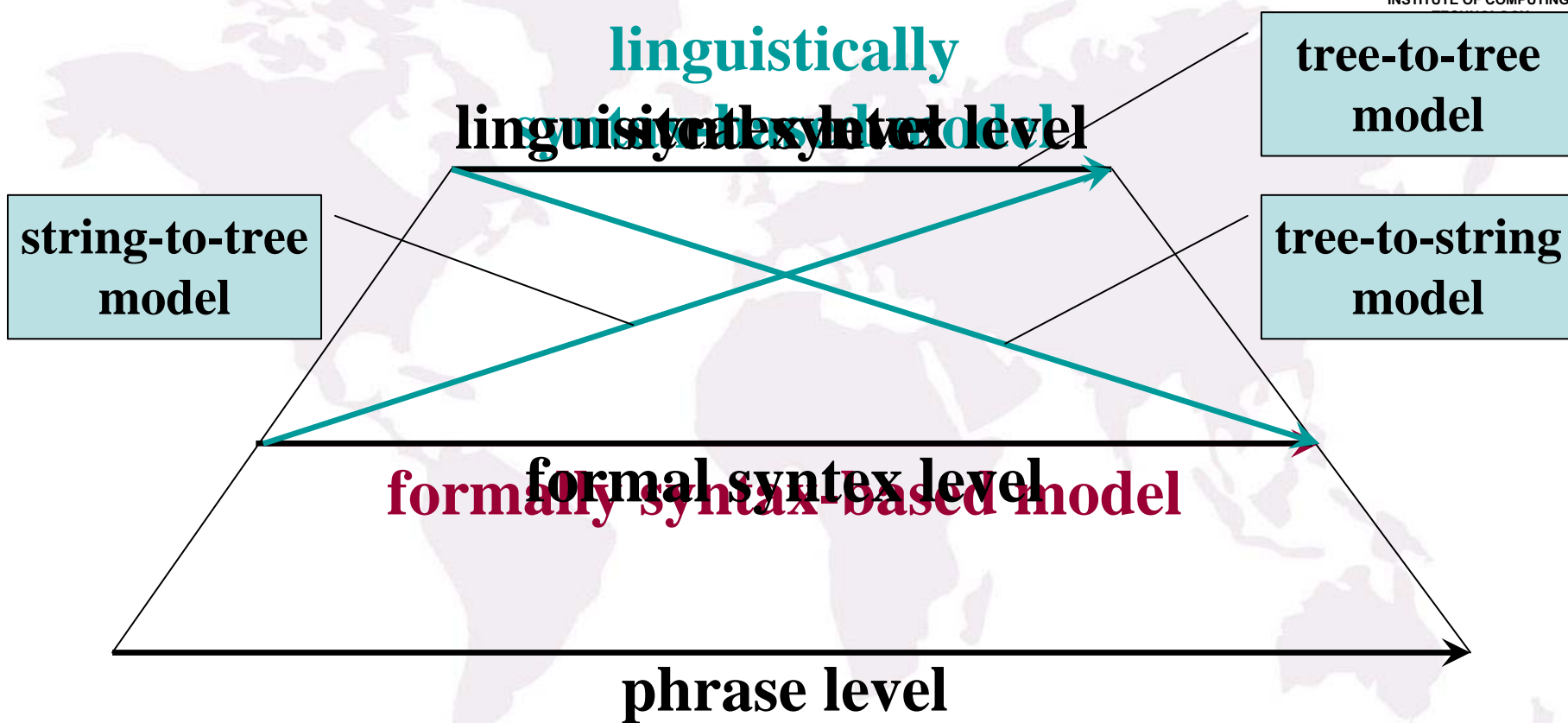
基于句法的统计翻译模型 (1)

- 基于句法的统计翻译模型，通常的做法都是分别为源语言和目标语言句子建立某种句法结构，并在这两种句法结构之间建立某种对应关系
- 基于句法的统计翻译模型有两种不同的做法
 - 形式上基于句法的统计翻译模型：并不采用语言学上的句法分析，而是从词语对齐的双语语料库中自动获取某种双语平行的句法结构
 - 语言学上基于句法的统计翻译模型：利用语言学上的句法分析，为源语言句子和目标语言句子建立句法结构，并借助词语对齐建立句法结构的对应关系

基于句法的统计翻译模型 (2)

- 语言学上基于句法的统计翻译模型又有三种不同的做法
 - 树到串模型：在源语言端进行句法分析并得到源语言句法结构，然后根据词语对齐建立对应的目标语言句法结构（可称为伪句法结构）
 - 串到树模型：在目标语言端进行句法分析并得到目标语言句法结构，然后根据词语对齐建立对应的源语言句法结构（也是伪句法结构）
 - 树到树模型：在源语言端和目标语言端分别进行句法分析并得到双语的句法结构，然后根据词语对齐建立这两种句法结构之间的对应关系

基于句法的统计翻译模型 (3)



形式上基于句法的模型

- 反向转录语法 (ITG) 和括号转录语法 (BTG)
Inversion (Bracketing) Transduction Grammar (ITG,BTG), Dekai Wu 1997
- 有限状态中心词转录机
Finite-State Head Transducer, Alshawhi 2000
- 基于层次短语的翻译模型
Hierarchical Phrase-based Model, David Chiang 2005
- 最大熵括号匹配语法的翻译模型
Maximal Entropy Bracket Transduction Grammar (ME-BTG), Deyi Xiong 2006

语言学上基于句法的模型

- 串到树模型 **String-to-Tree Model**
 - 美国南加州大学信息科学研究所（**ISI/CSU**）的工作
Yamada 2001, Galley 2006, Marcu 2006
- 树到串模型 **Tree-to-String Model**
 - 中科院计算所的工作
**Tree-to-string Alignment Template Model (TAT),
Liu Yang 2006**
 - 微软研究院的工作（依存模型）
Dependency Treelet Translation, Quirk 2005
- 树到树的模型 **Tree-to-Tree Model**

最大熵括号转录语法模型ME-BTG

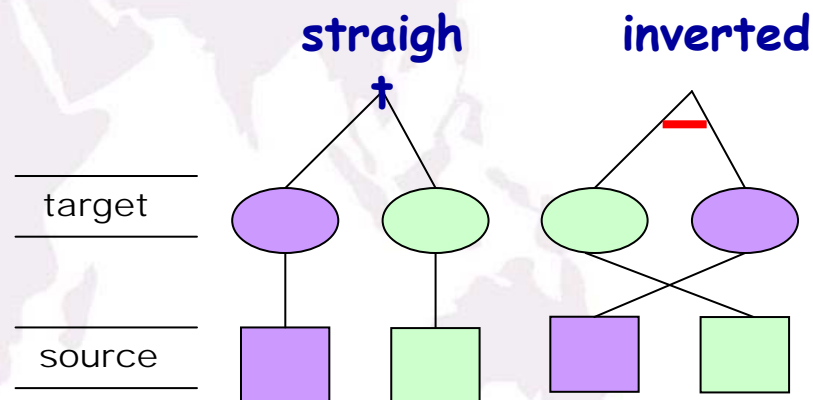
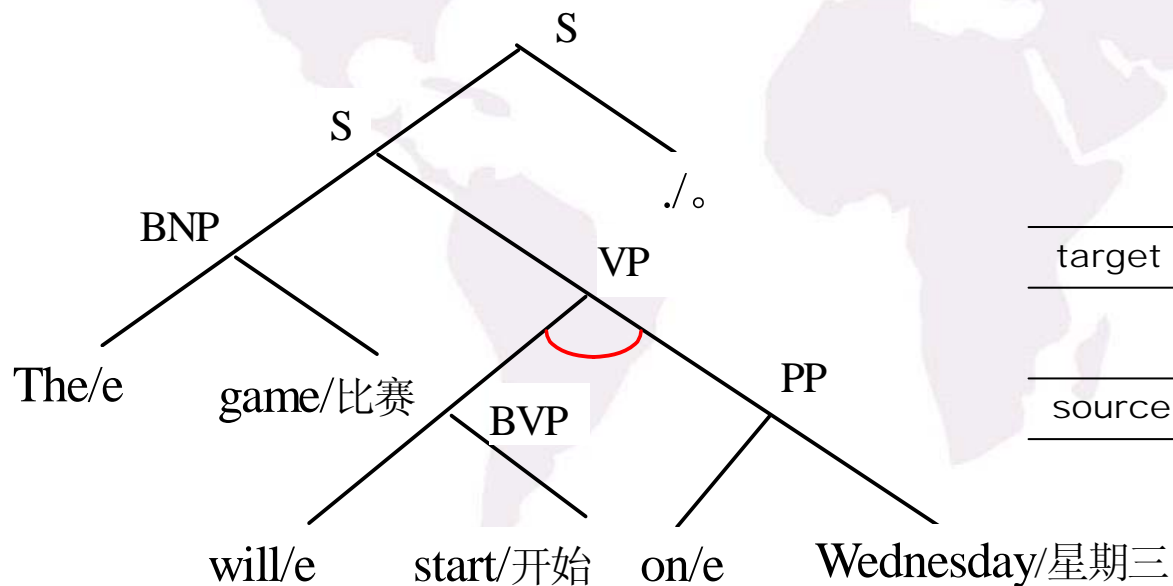
- 什么是**ITG**
- 基于**ITG**的机器翻译
- 什么是**BTG**
- 基于**BTG**建立统计翻译模型

什么是ITG (1)

- **ITG: Inversion Transduction Grammar**
- **ITG**是一种**Chomsky**范式形式的同步上下文无关语法
- **ITG**的规则有两种类型：
 - 非终结符规则（语法规则）
 - 终结符规则（词典）
- **ITG**规则采用**Chomsky**范式形式，因此所有非终结符规则都是二叉的，
- **ITG**的非终结符规则中，源语言规则到目标语言规则的对应关系只有两种：保序和交换

什么是ITG (2)

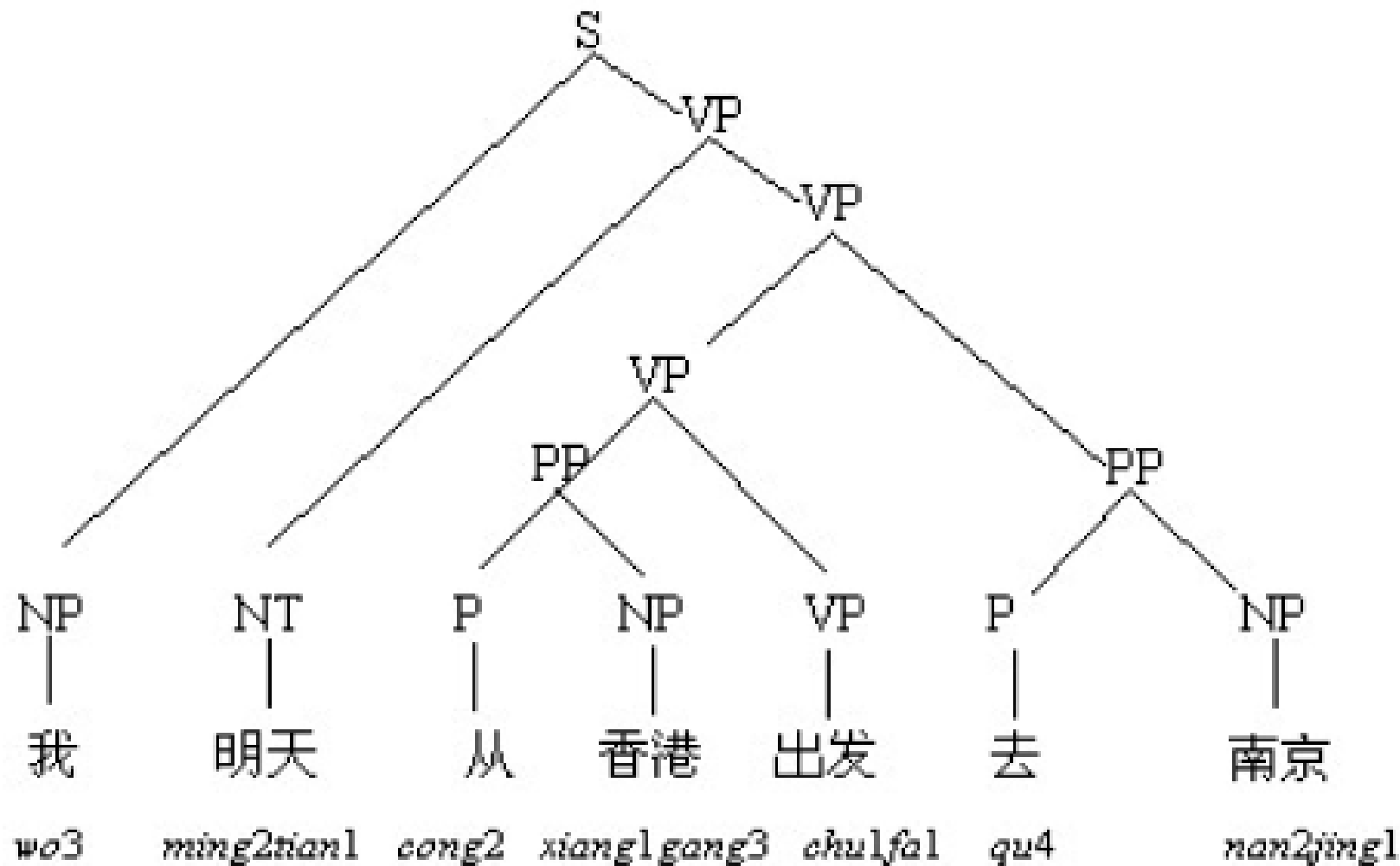
ITG rules	Source	Target
$A \rightarrow [B C]$	$A \rightarrow BC$	$A \rightarrow BC$
$A \rightarrow \langle B C \rangle$	$A \rightarrow BC$	$A \rightarrow CB$
$A \rightarrow x/y$	$A \rightarrow x$	$A \rightarrow y$



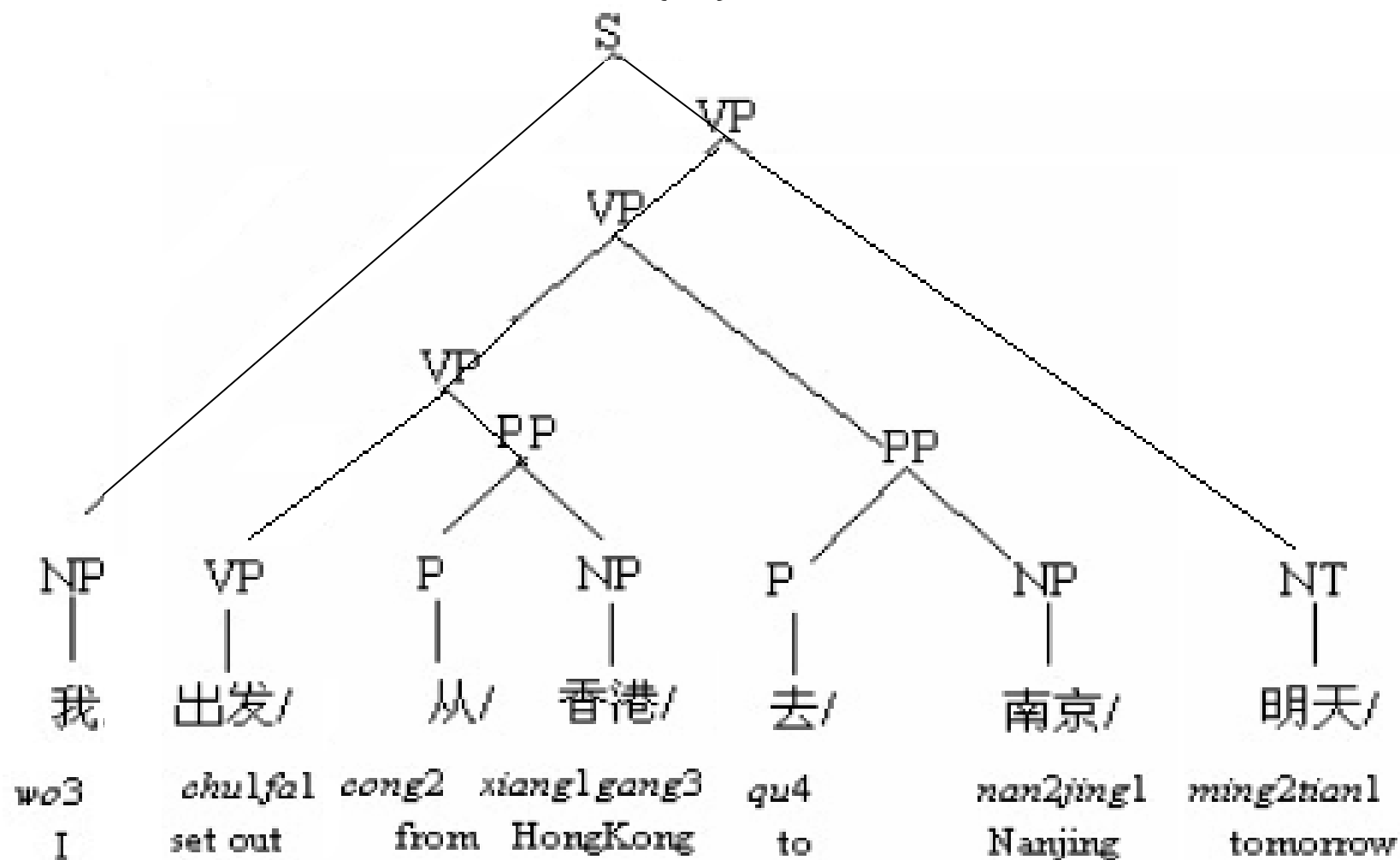
基于 ITG 的机器翻译 (1)

- 训练：
从词语对齐的双语语料库中自动获得**ITG**规则
- 解码：
类似于传统的基于规则的机器翻译方法
 - 先用**ITG**的源语言端规则对源语言进行句法分析
 - 根据**ITG**规则的映射关系，确定源语言句法树中每条源语言句法规则对应的目标语言句法规则
 - 生成目标语言句法树

基于 ITG 的机器翻译 (2)



基于 ITG 的机器翻译 (3)



基于 ITG 的机器翻译 (4)

- 在**ITG**中，仍然使用了**NP**、**VP**之类的句法标记，这对于训练语料库提出了比较高的要求
- 如果我们不考虑标记，也就是说，认为所有的标记都是相同的，只有一个非终结符标记**X**，那么**ITG**就退化成**BTG**

什么是BTG

- **BTG: Bracketing transduction grammars**
- **BTG**是简化的**ITG**，也就是在**ITG**中，只定义一个非终结符**X**
- **BTG**为两种语言的句法结构之间的对应关系建立了一个最简单的模型
 - 没有标记，只有结构
 - 没有多叉，只有二叉
- **BTG**大大限制了统计机器翻译的解码空间

BTG约束 vs. IBM约束 (1)

- **IBM约束 (IBM constraint)**

- 目标语言词语可以对齐到源语言句子中的任何一个词
- 解码的时候，从目标语言自左向右猜测每一个词语，首先猜测该词对齐到源语言的那一个词，然后猜测该词是什么。
- 在**IBM**约束下，可能的对齐方式将是指数级的。
- 为了避免搜索空间的膨胀，通常是限制词语调序的距离，而这种限制显然是不合理的

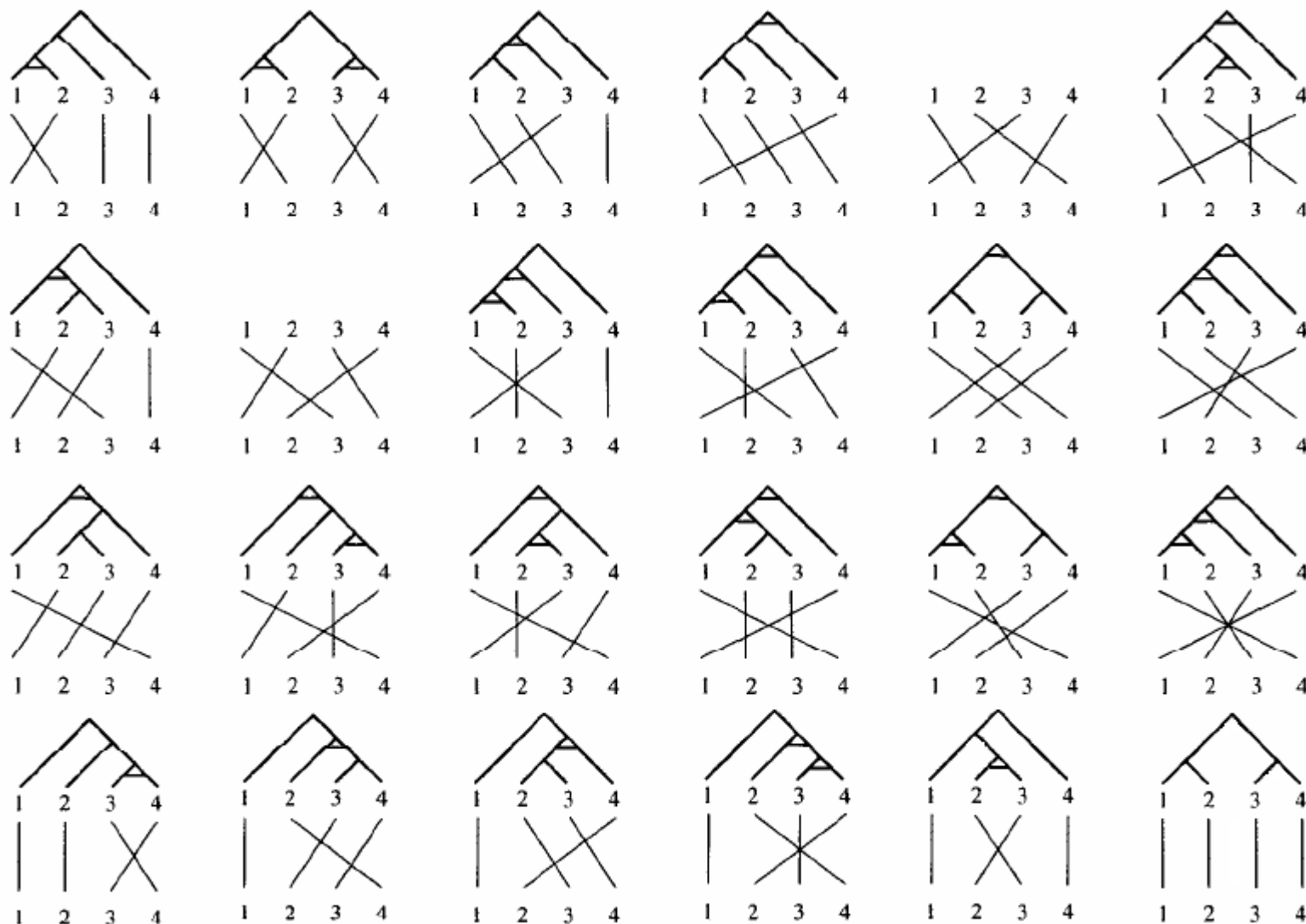
- **BTG约束 (BTG constraint)**

- 只有满足某种**BTG**对应关系的目标语言词序才是允许的，否则排除在搜索空间之外
- 解码的时候，采用类似于**CYK**句法分析的方式进行解码，就可以穷尽所有可能的**BTG**约束下的词序
- 在**BTG**约束下，可能的对齐方式是多项式级的
- 无需限制长距离的词序调整

BTG约束 vs. IBM约束 (2)

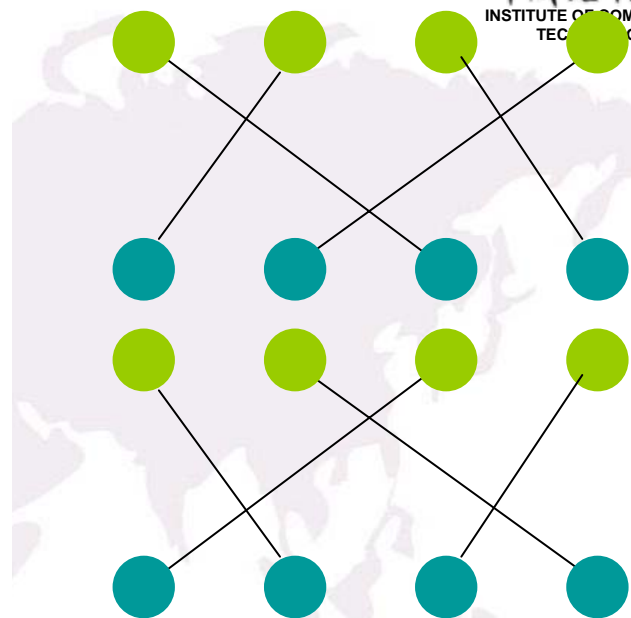


这里给出了四个词在IBM约束和BTG约束下所有可能的词序调整方案。其中有两种方案在IBM约束下是允许的，但在BTG约束下是不允许的。



BTG约束 vs. IBM约束 (3)

f	BTG	all matchings	ratio
0	1	1	1.000
1	1	1	1.000
2	2	2	1.000
3	6	6	1.000
4	22	24	0.917
5	90	120	0.750
6	394	720	0.547
7	1806	5040	0.358
8	8558	40320	0.212
9	41586	362880	0.115
10	206098	3628800	0.057
11	1037718	39916800	0.026
12	5293446	479001600	0.011
13	27297738	6227020800	0.004
14	142078746	87178291200	0.002
15	745387038	1307674368000	0.001
16	3937603038	20922789888000	0.000



**word reordering
which are not
permitted in BTG**

BTG约束 vs. IBM约束 (4) 一个反例

- For Chinese and English, almost true.
 - an exception:



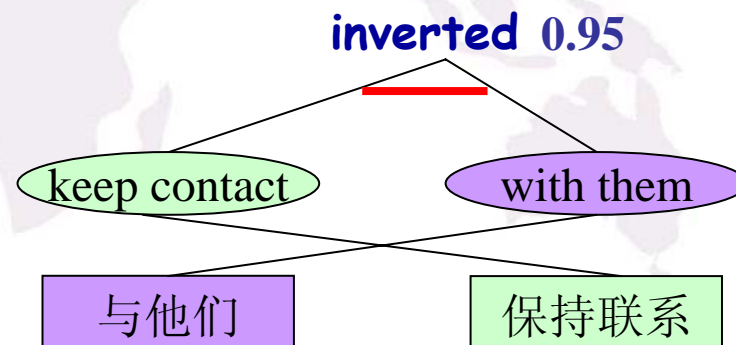
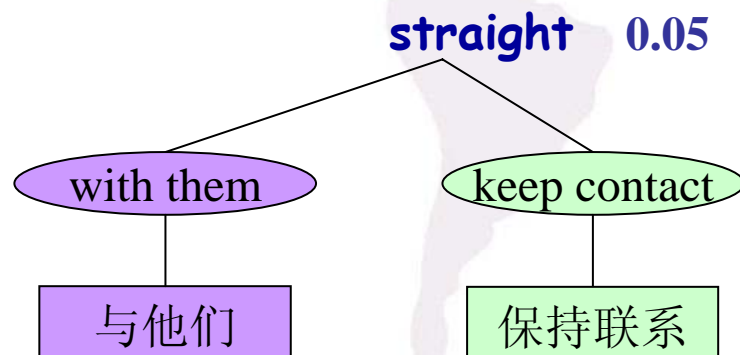
- For some other languages with free order, not true.

基于BTG的机器翻译

- 只有两条非终结符规则：
 $A \rightarrow [A A]$
 $A \rightarrow \langle A A \rangle$
- 吴德凯定义的**Stochastic BTG**给每条规则赋以先验概率
- **Stochastic BTG**是一种非常粗糙的调序模型，无法在细粒度上处理词语调序问题，实际应用效果也很不理想

MEBTG: 基本思想

- 在**BTG**框架下，将重排序问题看作是一个二元分类问题：
 - 条件：各种与重排序短语相关的特征
 - 类别：相邻语块的顺序 **{straight, inverted}**
- 引入最大熵模型作为分类模型，根据实际上下文计算合并规则的概率



MEBTG模型

- 模型

$$\Omega = p_{\theta}(o | A^1, A^2) = \frac{\exp(\sum_i \theta_i h_i(o, A^1, A^2))}{\sum_{o'} \exp(\sum_i \theta_i h_i(o', A^1, A^2))}$$

- 特征

$$h_i(o, A^1, A^2) = \begin{cases} 1 & \text{if } f(A^1, A^2) = T, o = O \\ 0 & \text{otherwise} \end{cases}$$

$$O \in \{straight, inverted\}$$

重排序特征

- 单目特征：单个源/目标语言边界单词
- 双目特征：两个源/目标语言边界单词的组合

<与 他们|with them; 保持联系|keep contact> → INVERTED

特征选择

$$h_{mono}(o, A^1, A^2) = \begin{cases} 1 & \text{if } A^2.t_1 = keep, o = inverted \\ 0 & \text{otherwise} \end{cases}$$

$$h_{bino}(o, A^1, A^2) = \begin{cases} 1 & \text{if } A^1.t_1 = with, A^2.t_1 = keep, o = inverted \\ 0 & \text{otherwise} \end{cases}$$

MEBTG模型小结

- 形式上基于句法的模型
- 性能明显超过基于短语的模型
- 完全兼容基于短语的模型
- 采用**BTG**语法形式，只有两条规则
- 规则的选用采用最大熵方法进行取舍

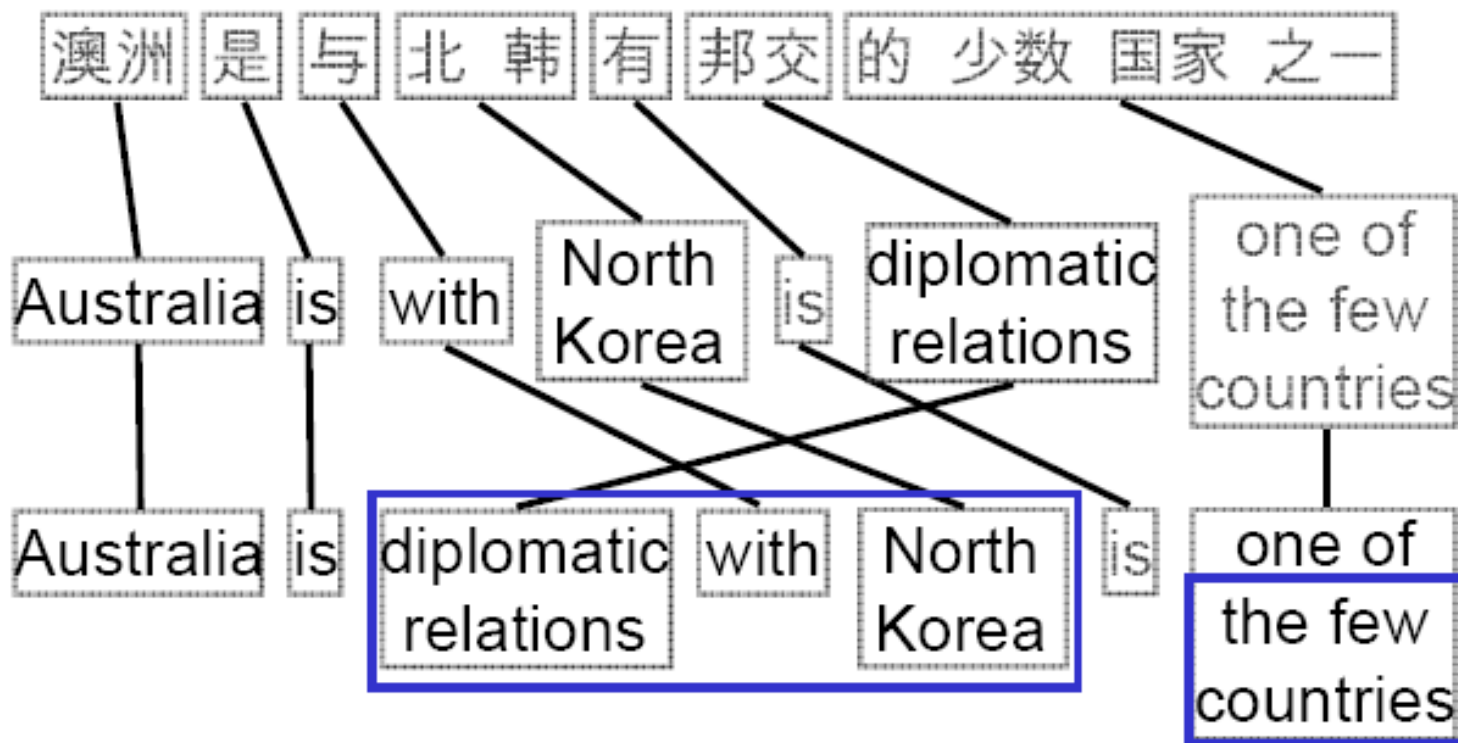
层次短语模型 (1)

- **David Chiang. A Hierarchical Phrase-Based Model for Statistical Machine Translation. ACL2005. (Best Paper Award)**
- 本讲义这一部分内容直接引用了以下讲义的部分内容，特此说明并向原作者表示感谢：
 - **David Chiang, Hiero: Finding Structure in Statistical Machine Translation, in National University of Singapore**

层次短语模型 (2)

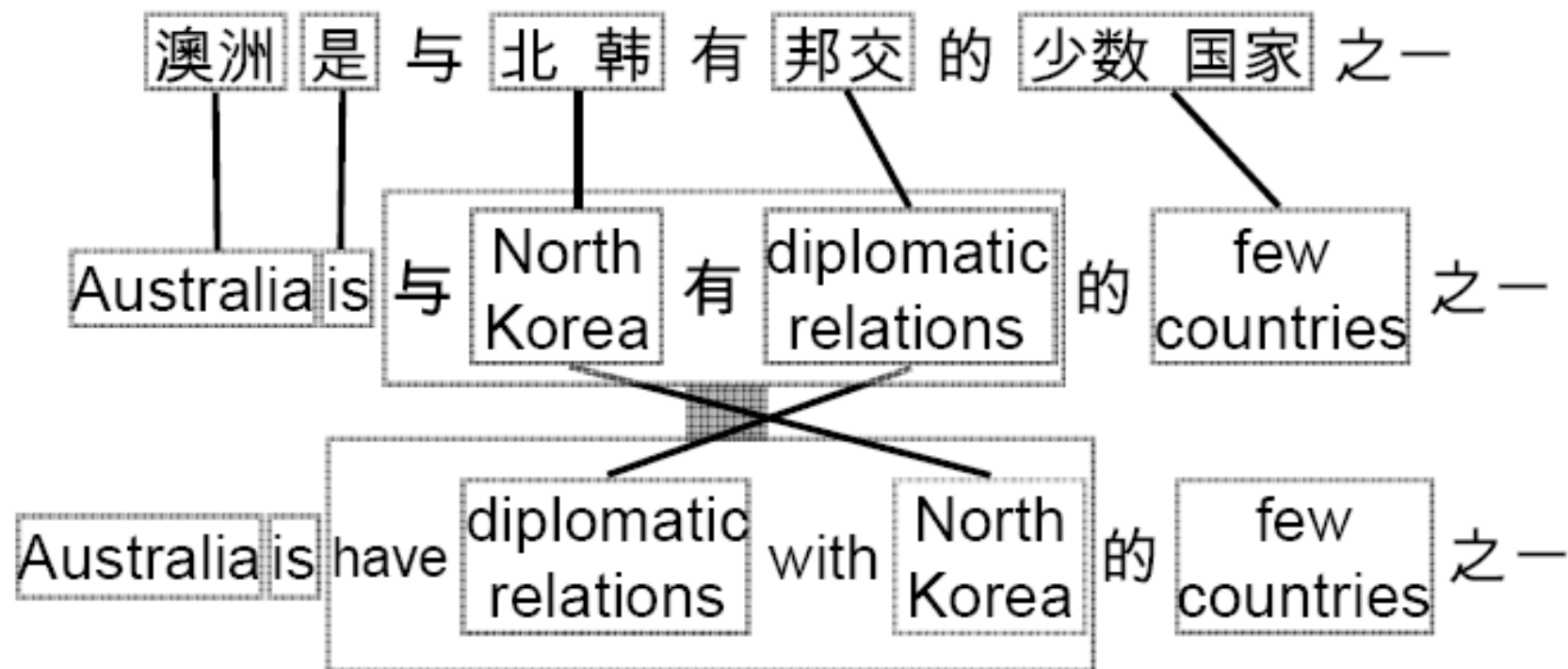
- 传统的基于短语的翻译模型中，短语是平面的，不能嵌套
- 在层次短语模型中，引入了嵌套的层次短语
- 采用平行上下文无关语法作为理论基础，但只使用唯一的非终结符标记
- 效果比传统的短语模型有很大提高

平面的短语

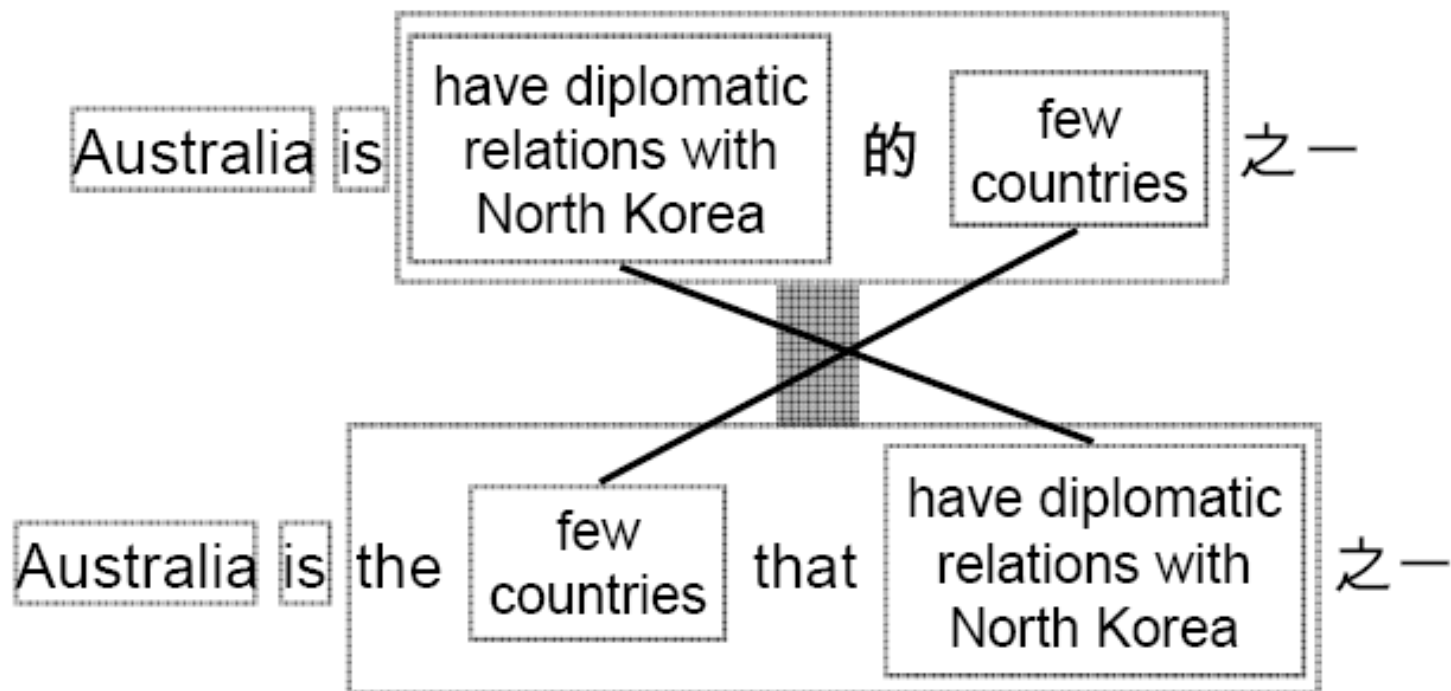


可以观察到短语是有层次的。

层次短语 (1)



层次短语 (2)

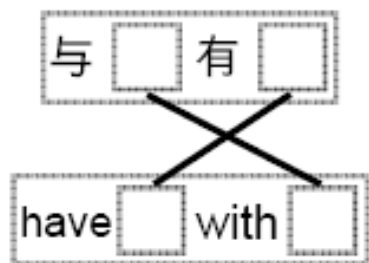


层次短语 (3)

Australia is the few countries that have diplomatic relations with North Korea 之一

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

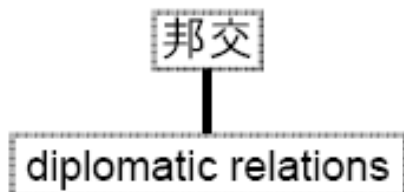
用同步语法表示层次短语 (1)



$(X \rightarrow \text{与 } X_1 \text{ 有 } X_2, X \rightarrow \text{have } X_2 \text{ with } X_1)$

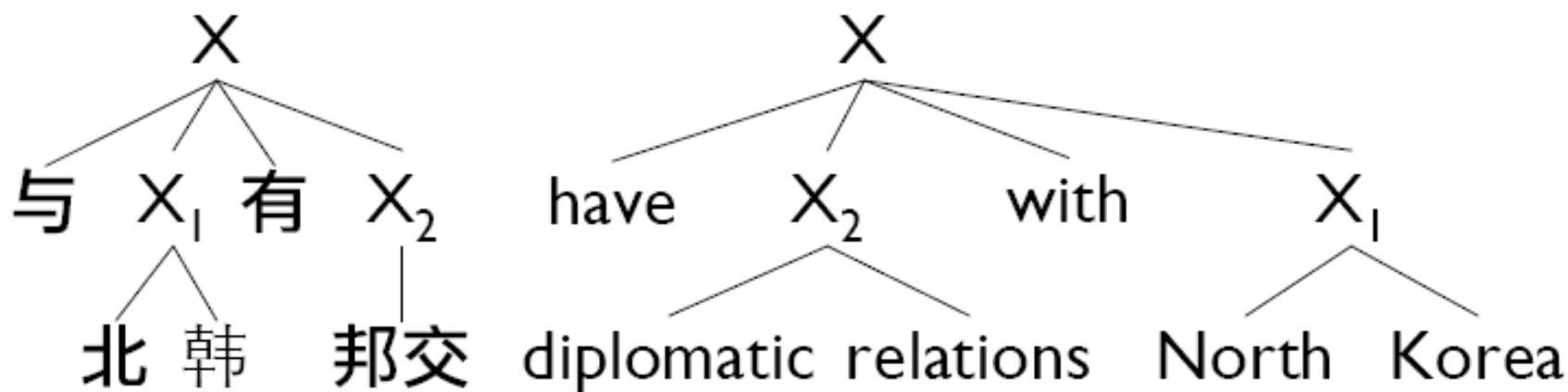


$(X \rightarrow \text{北 韩}, X \rightarrow \text{North Korea})$



$(X \rightarrow \text{邦交}, X \rightarrow \text{diplomatic relations})$

用同步语法表示层次短语 (2)



规则举例

$X \rightarrow \text{的}$

$X \rightarrow \text{'s}$

$X \rightarrow X_1 \text{ 的 } X_2$

$X \rightarrow \text{the } X_2 \text{ of } X_1$

$X \rightarrow X_1 \text{ 的 } X_2$

$X \rightarrow \text{the } X_2 \text{ that } X_1$

$X \rightarrow \text{在}$

$X \rightarrow \text{in}$

$X \rightarrow \text{在 } X_1 \text{ 下}$

$X \rightarrow \text{under } X_1$

$X \rightarrow \text{在 } X_1 \text{ 前}$

$X \rightarrow \text{before } X_1$

$X \rightarrow \text{今年 } X_1$

$X \rightarrow X_1 \text{ this year}$

$X \rightarrow X_1 \text{ 之一}$

$X \rightarrow \text{one of } X_1$

$X \rightarrow X_1 \text{ 总统}$

$X \rightarrow \text{president } X_1$

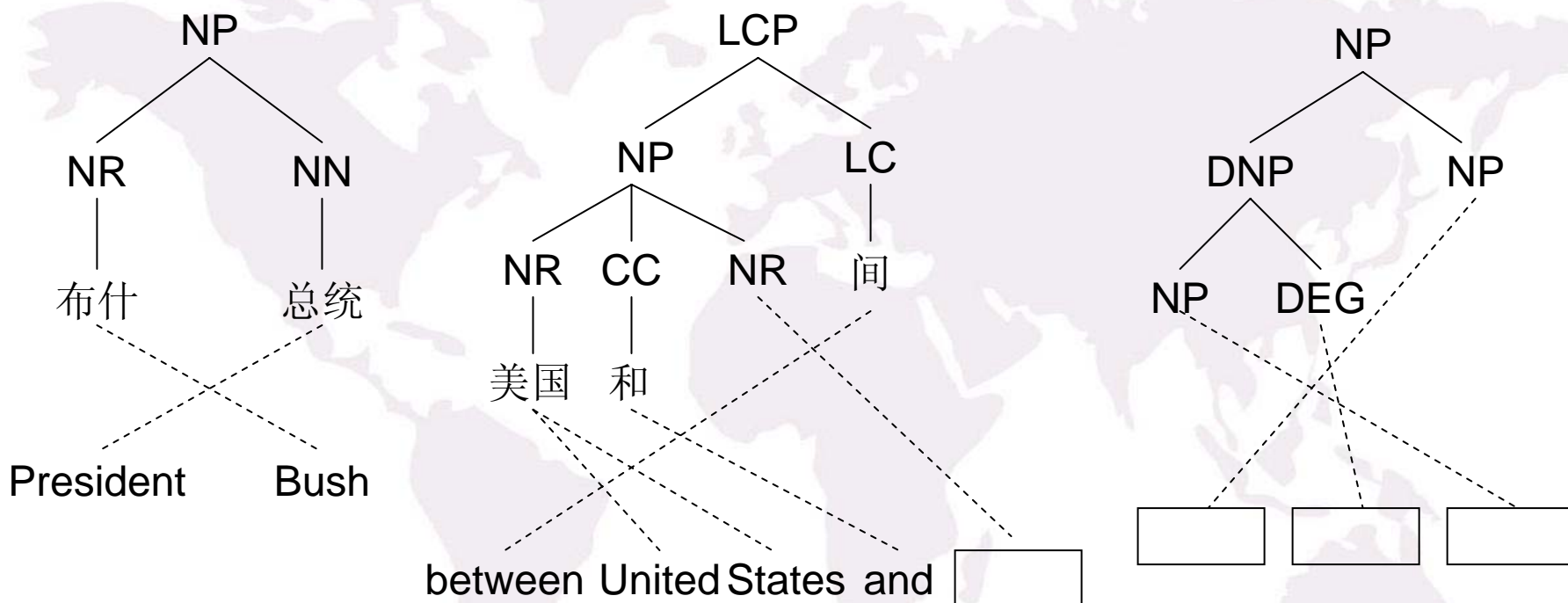
层次短语模型小结

- 形式上基于句法的模型
- 性能明显超过基于短语的模型
- 完全兼容基于短语的模型
- 规则采用同步上下文无关语法形式，但只有一个非终结符**X**
- 所有规则可以自动抽取
- 规则数量极为庞大

基于树到串对齐模板的翻译模型

- 基于树到串对齐模板（简称**TAT**）的统计翻译模型是一种在源语言进行句法分析的基于语言学句法结构的统计翻译模型
- 树到串对齐模板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**TAT**
- 自底向上的柱搜索算法

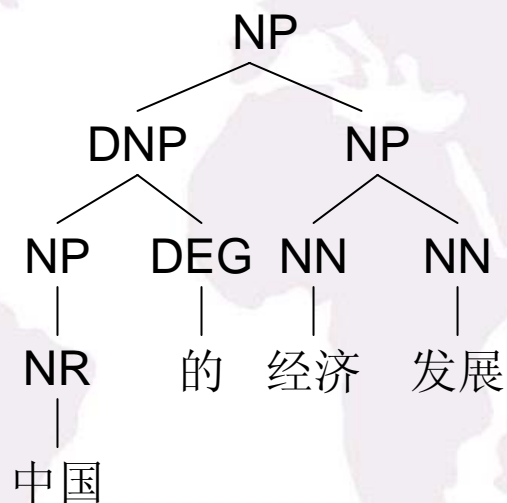
树到串对齐模板



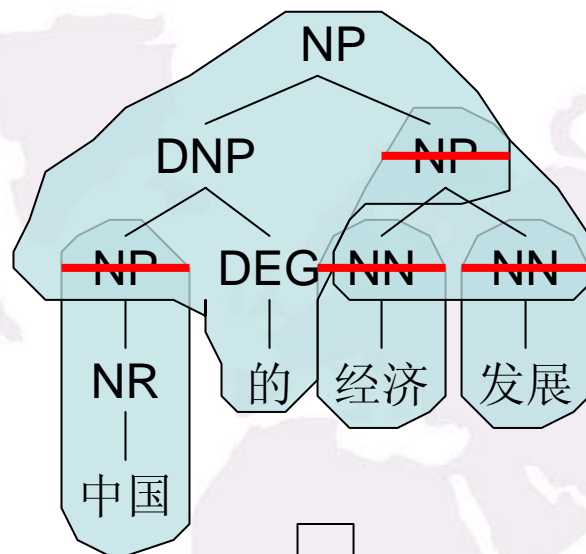
翻译过程：Parsing

中国的经济发展

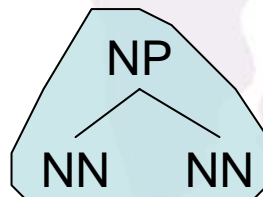
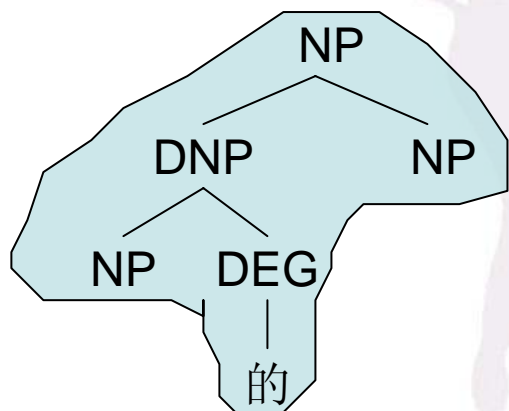
parsing



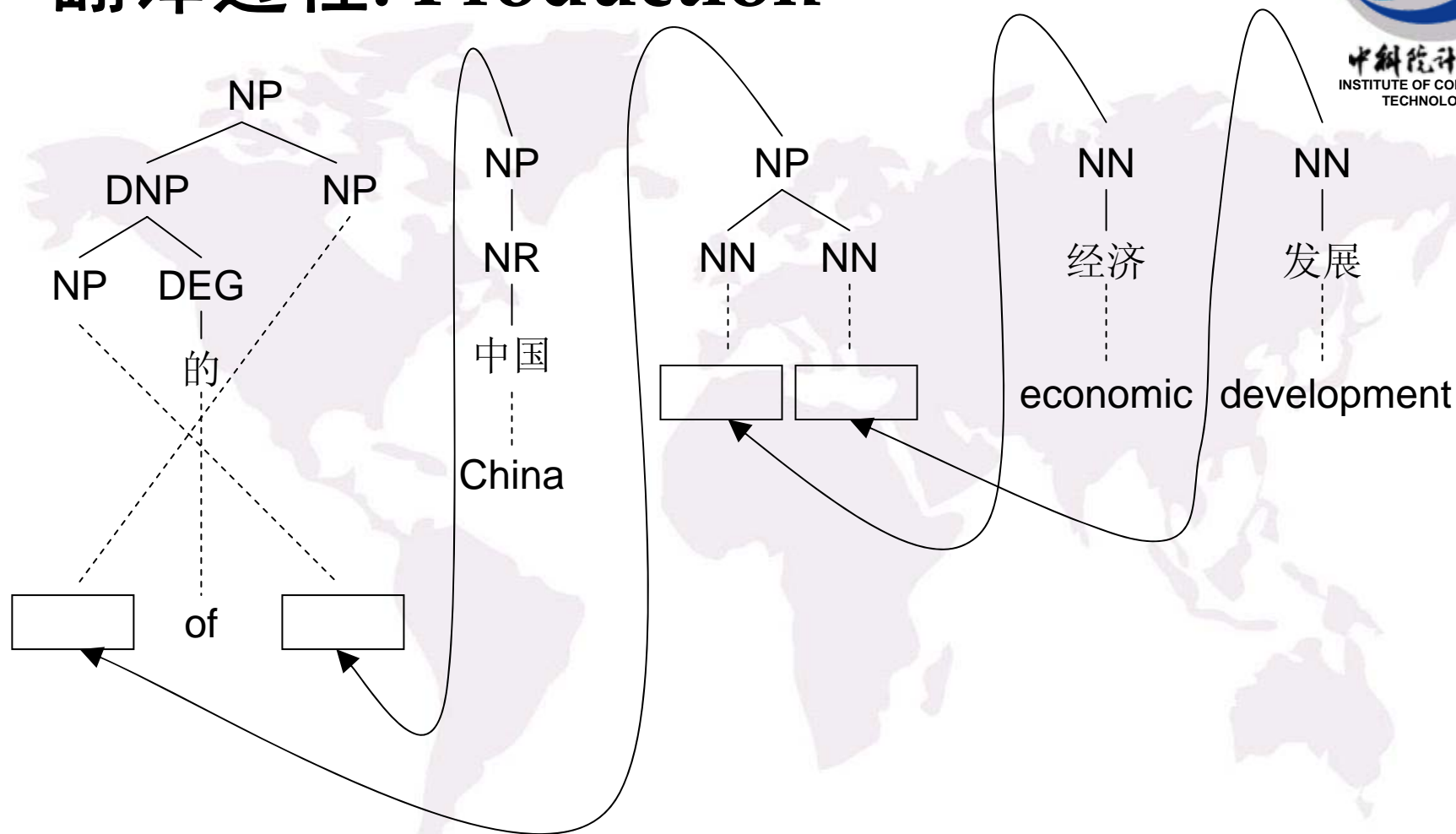
翻译过程: Detachment

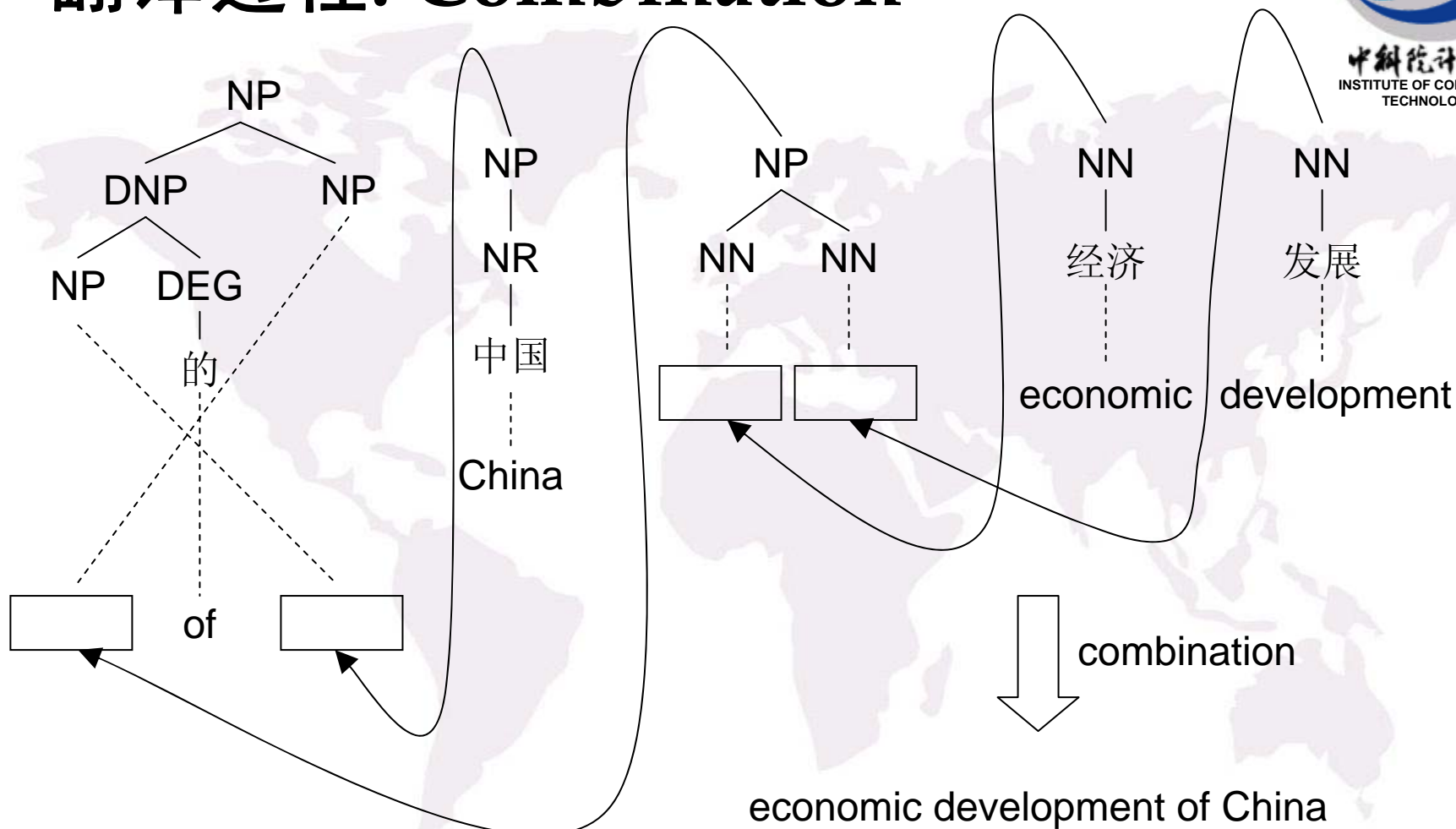


detachment



翻译过程: Production





基于树到串模板的统计翻译模型小结

- 一种语言学上基于句法的模型
- 训练时除了双语词语对齐外，还要对源语言进行句法分析
- 树到串对齐模板**TAT**的源语言端是一个子树，不是一条上下文无关语法的规则，等价于同步数替换语法（**STSG**）
- 不完全兼容于基于短语的模型（其扩展形式“基于森林到串模板的统计翻译模型”可以兼容于基于短语的模型）
- 所有规则全自动抽取
- 规则数量极为庞大
- 解码时要先利用传统的句法分析器进行源语言句法分析，然后采用基于句法树的堆栈搜索
- 性能比基于短语的模型有显著提高

串到树的统计翻译模型 (1)

- **USC-ISI**的系列工作
- 发表了大量论文，但还没有一个完整的论述
- 性能优异，在**NIST2006**汉英项目平常中超过了**Google**（**Google**使用的语言模型规模比**ISI**大得多）

串到树的统计翻译模型 (2)

- 基本思想
 - 在目标语言端进行句法分析
 - 根据目标语言端的句法结构，和词语对齐，建立源语言端的句法结构（伪树）
 - 利用两个句法结构自动抽取带概率的平行上下文无关语法
 - 对平行上下文无关语法进行二叉化
 - 解码时类似规则方法，复杂度等价于句法分析
 - 源文分析
 - 规则映射
 - 译文生成

翻译过程：串到树解码

枪手

被

警方

击毙

。

翻译过程：串到树解码

NNS
gunmen

枪手

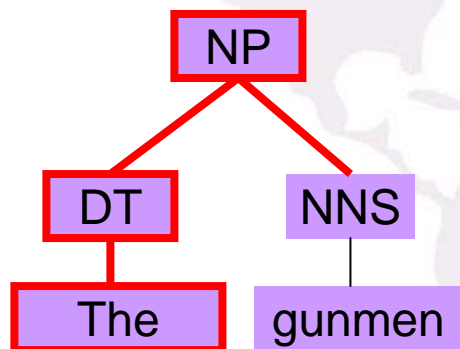
被

警方

击毙

。

翻译过程：串到树解码



枪手

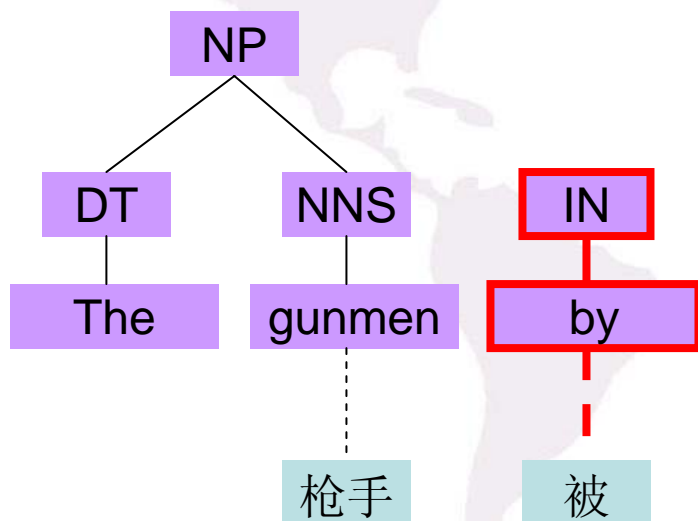
被

警方

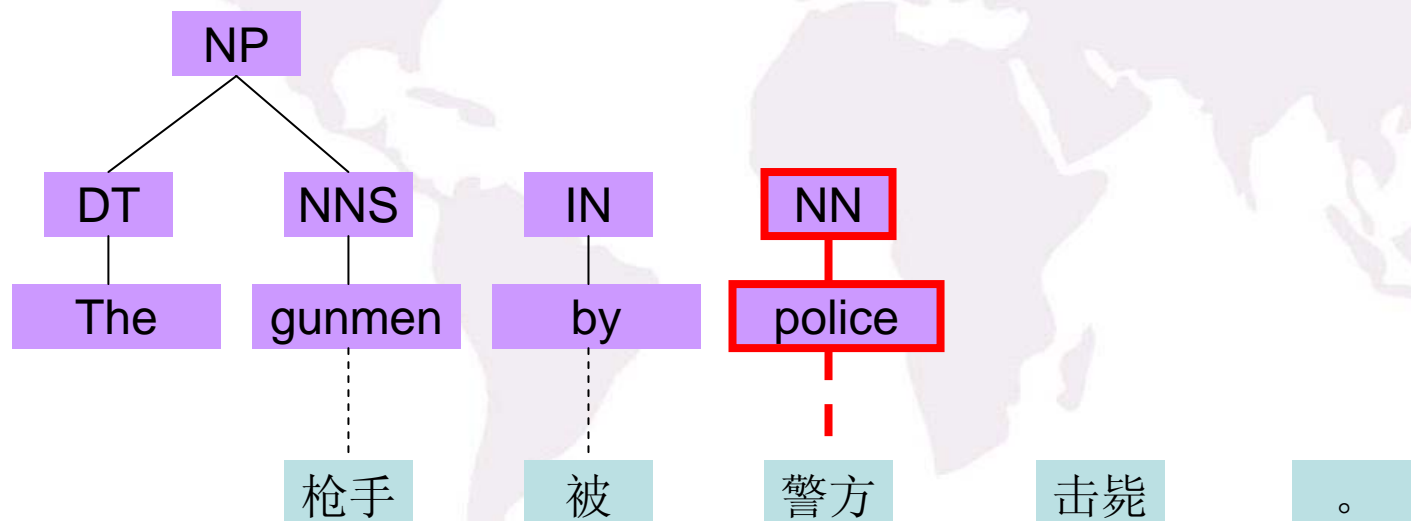
击毙

。

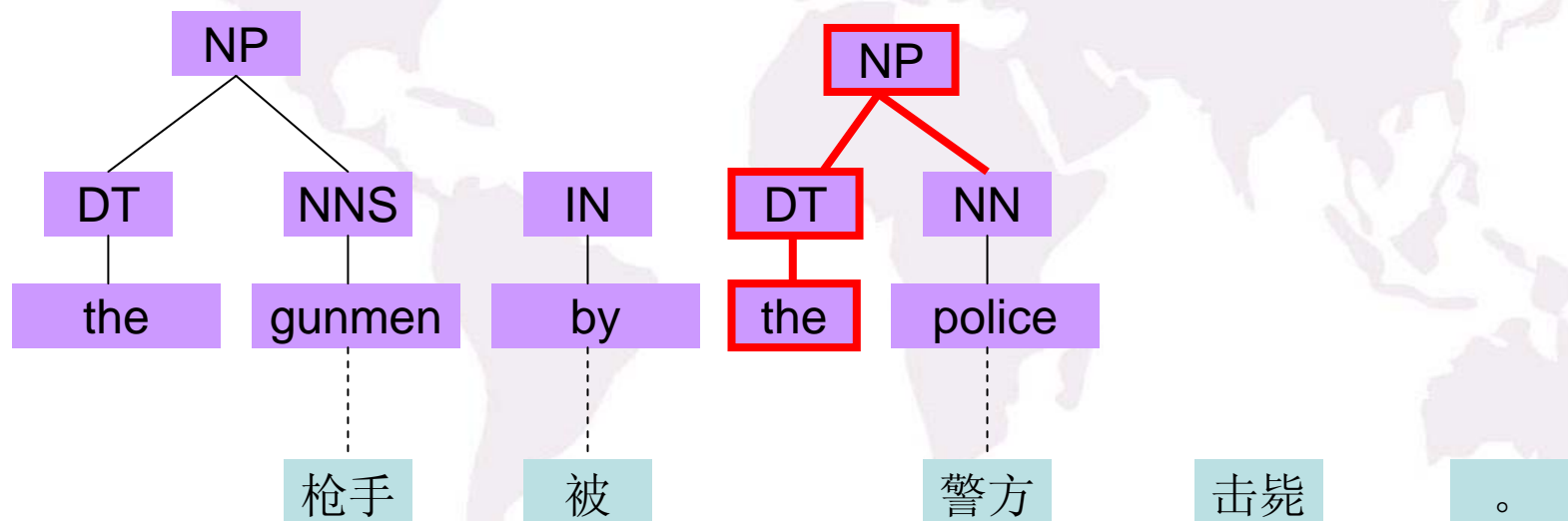
翻译过程：串到树解码



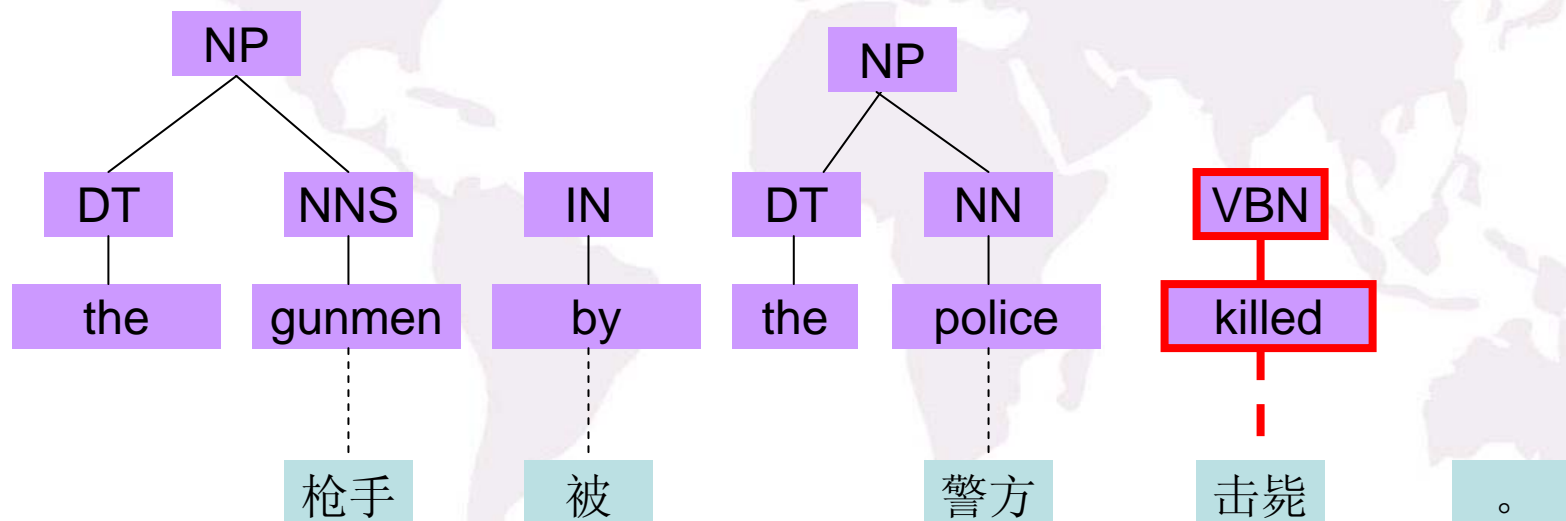
翻译过程：串到树解码



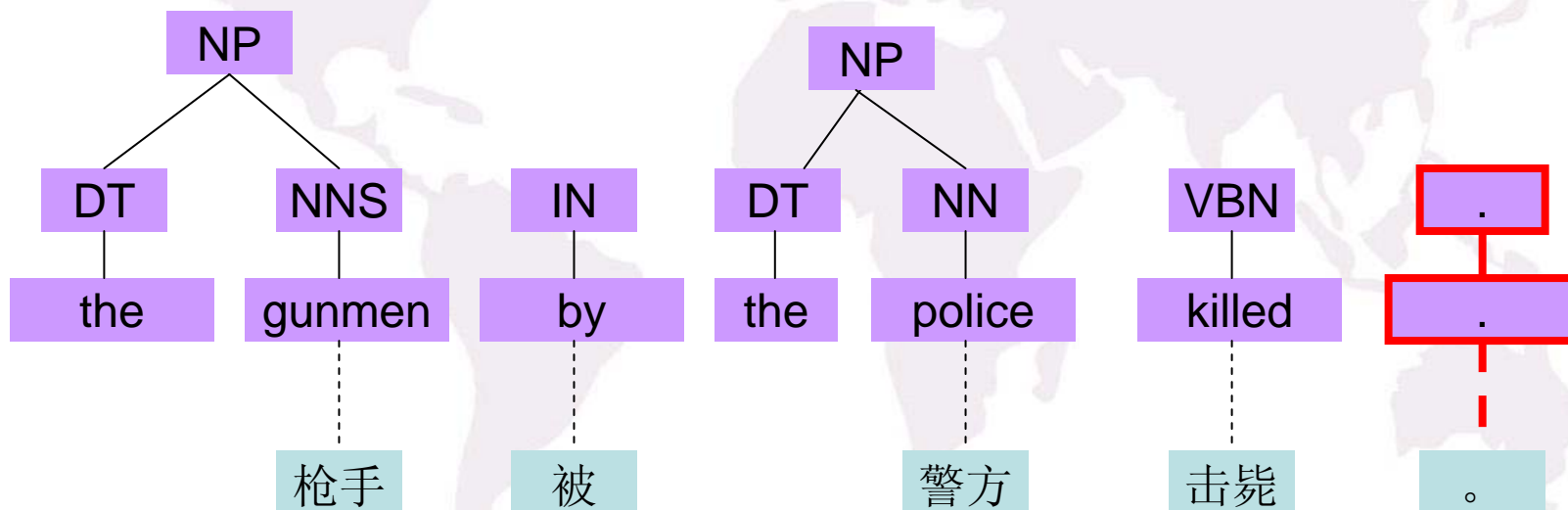
翻译过程：串到树解码

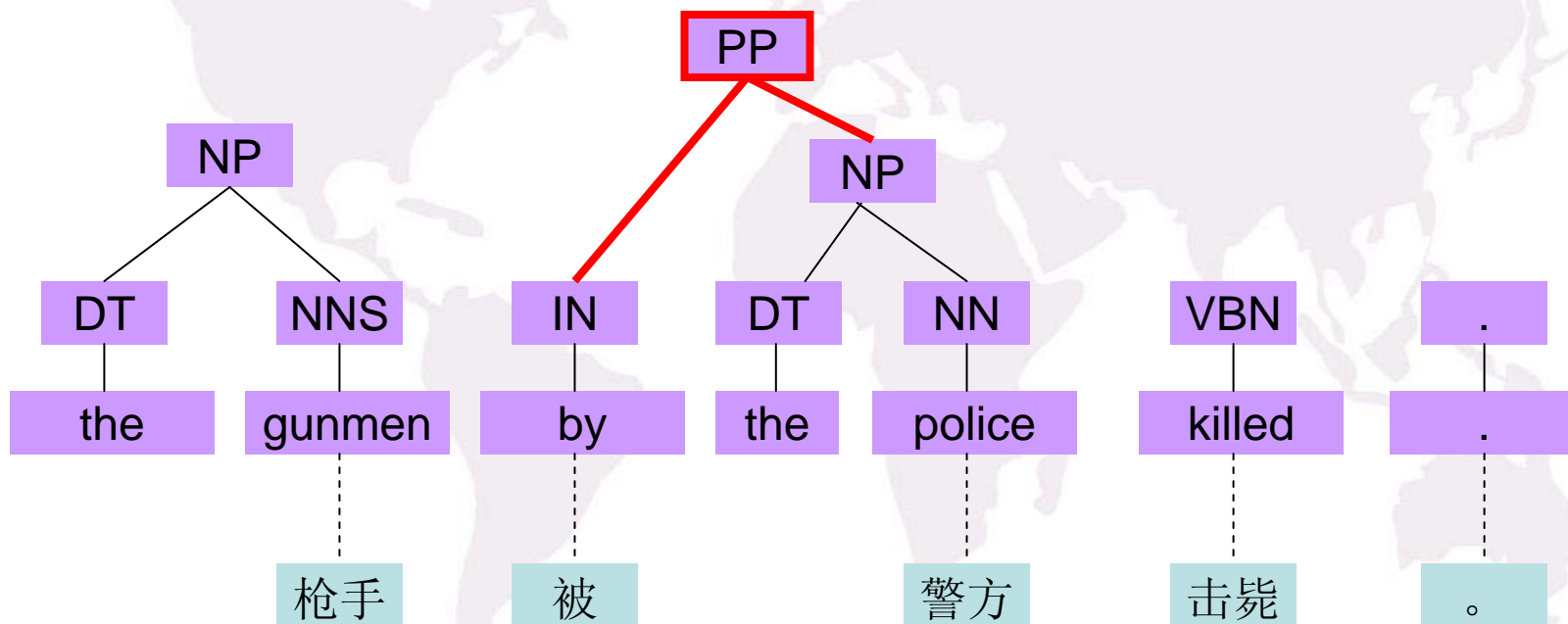


翻译过程：串到树解码

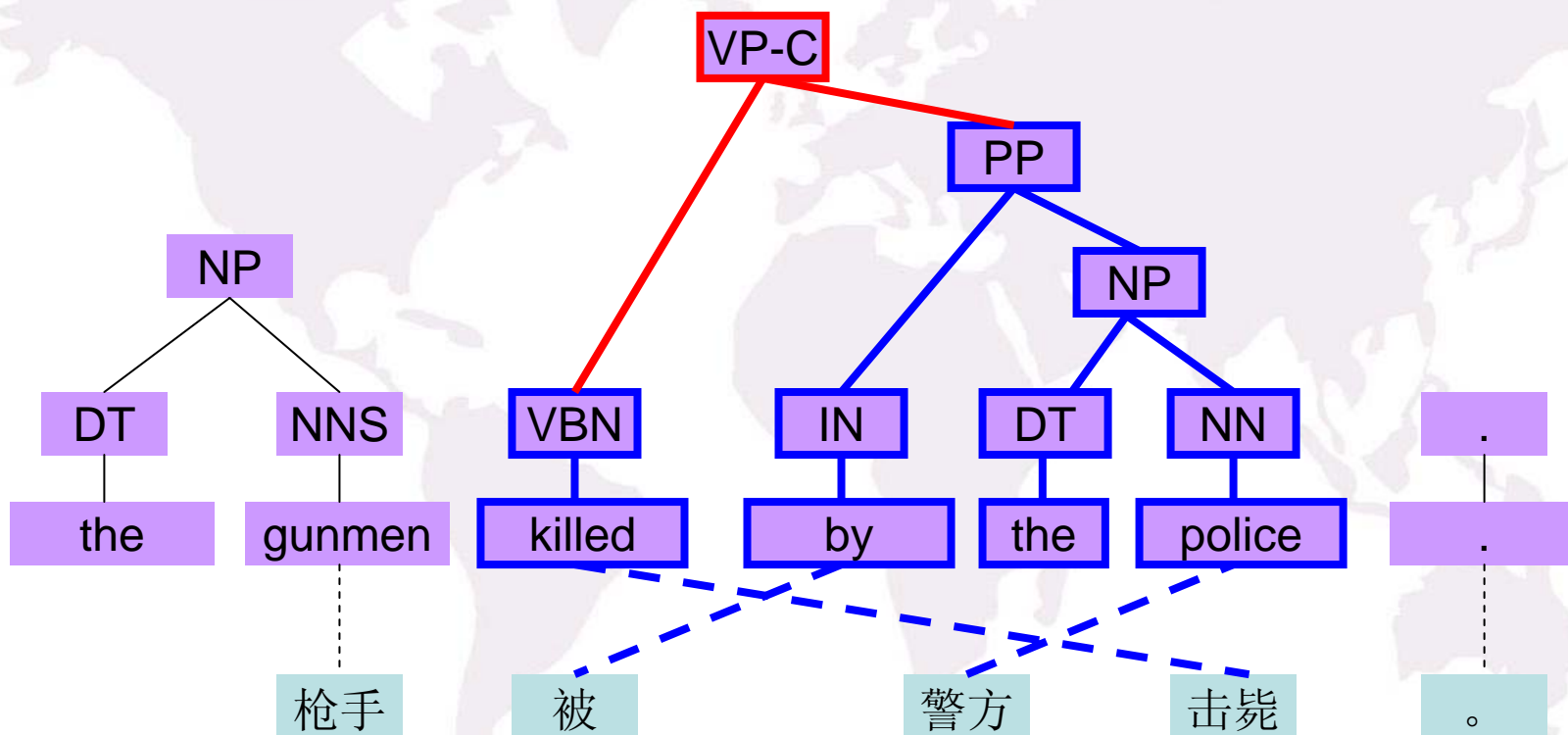


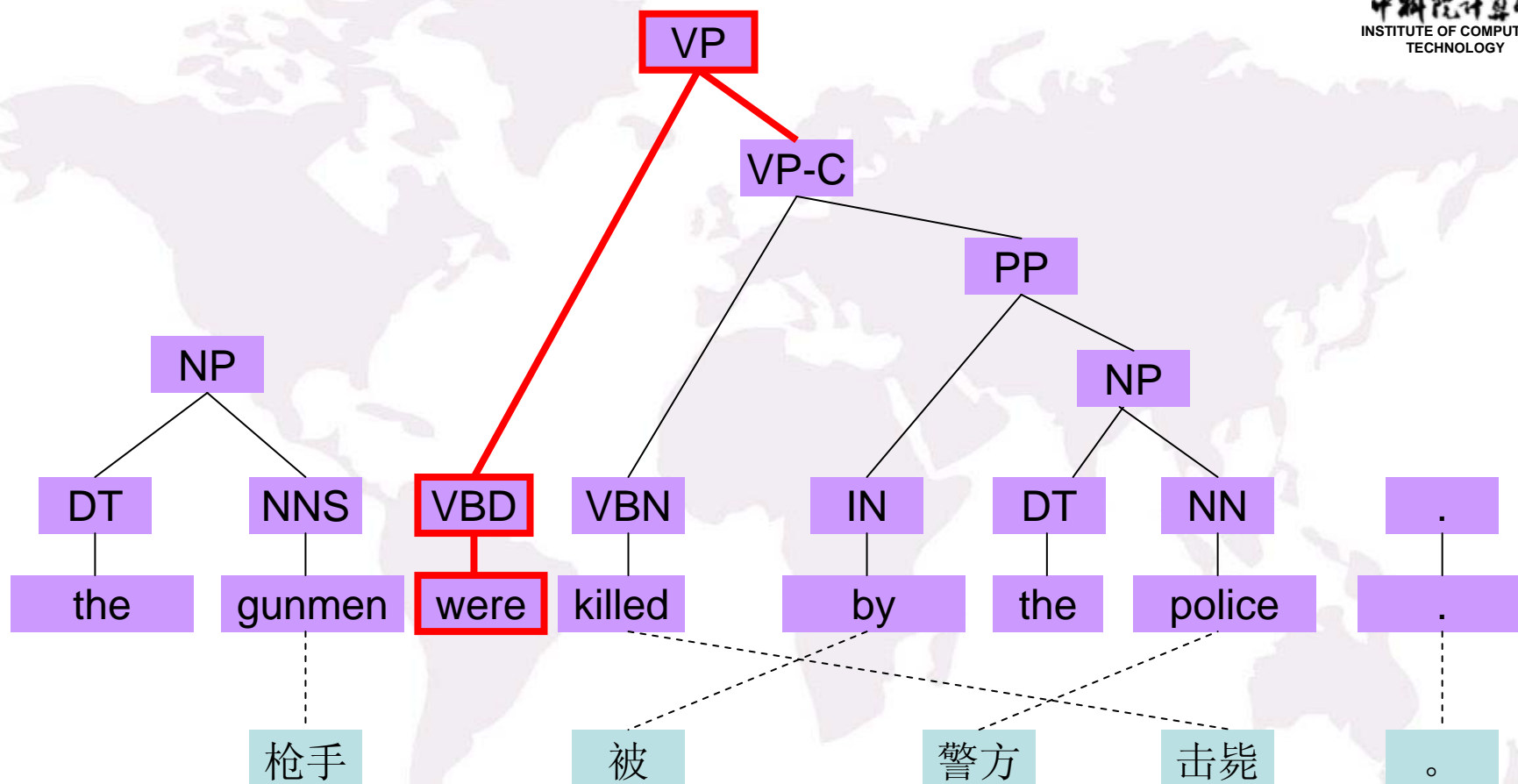
翻译过程：串到树解码



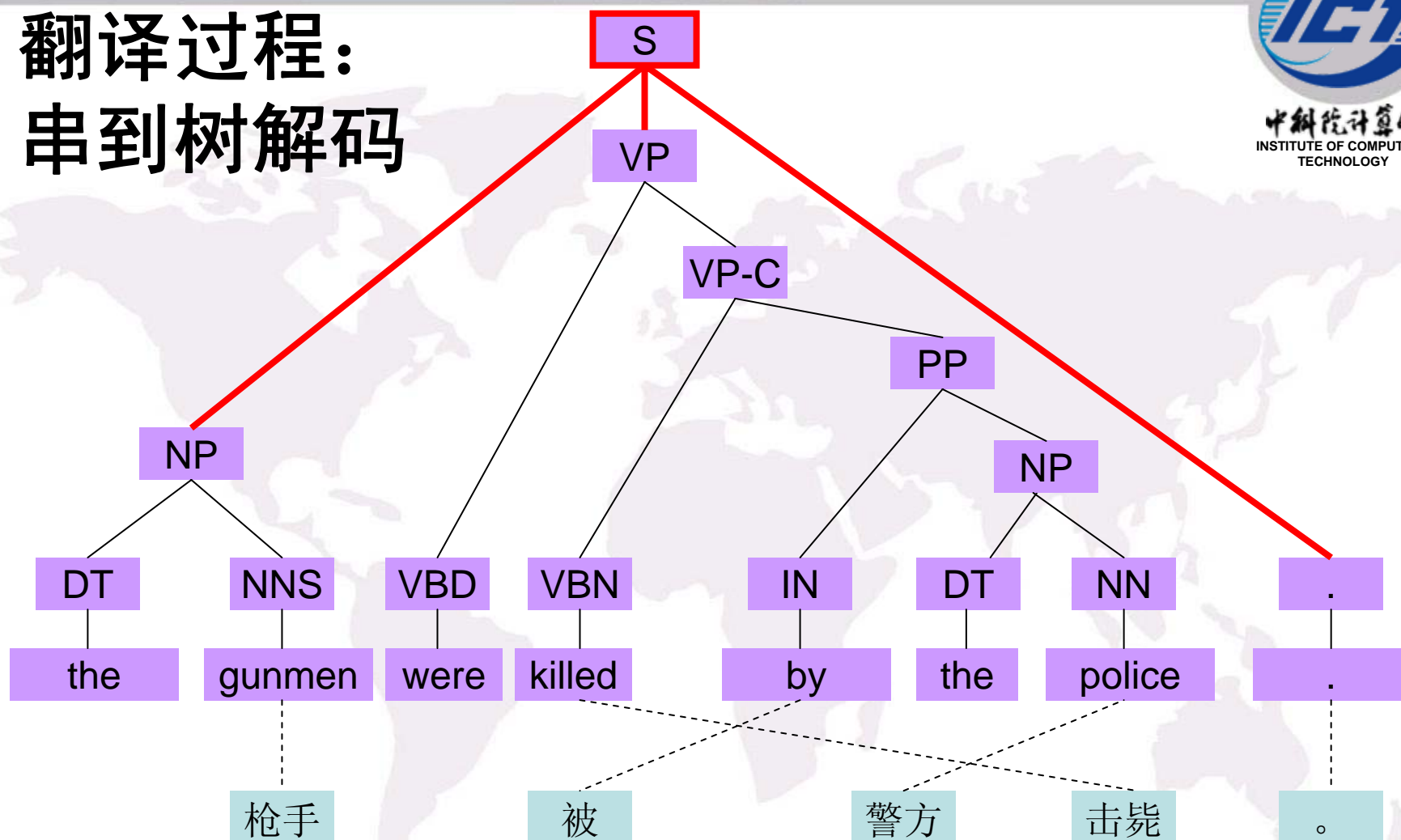


翻译过程：串到树解码





翻译过程： 串到树解码



串到树的统计翻译模型小结

- 一种语言学上基于句法的模型
- 训练时除了双语词语对齐外，还要对目标语言进行句法分析
- 规则形式是同步上下文无关语法形式
- 不完全兼容于基于短语的模型所有规则全自动抽取（修改后可兼容）
- 规则数量极为庞大
- 不需要利用传统的句法分析器进行句法分析
- 解码过程等价于句法分析过程
- 性能比基于短语的模型有显著提高

目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
——基于词的**IBM**模型
- 最成熟的统计机器翻译方法
——基于短语的模型
- 目前统计机器翻译研究的热点
——基于句法的模型
- 中科院计算所的工作
- 统计机器翻译面临的问题和展望

在本领域顶级会议上发表的论文

- **ACL: 11**

2005:1 2006:2 2007:1 2008:3 2009:4

- **EMNLP: 6**

2007:1 2008:2 2009:3

- **COLING: 4**

2006:2 2008:2

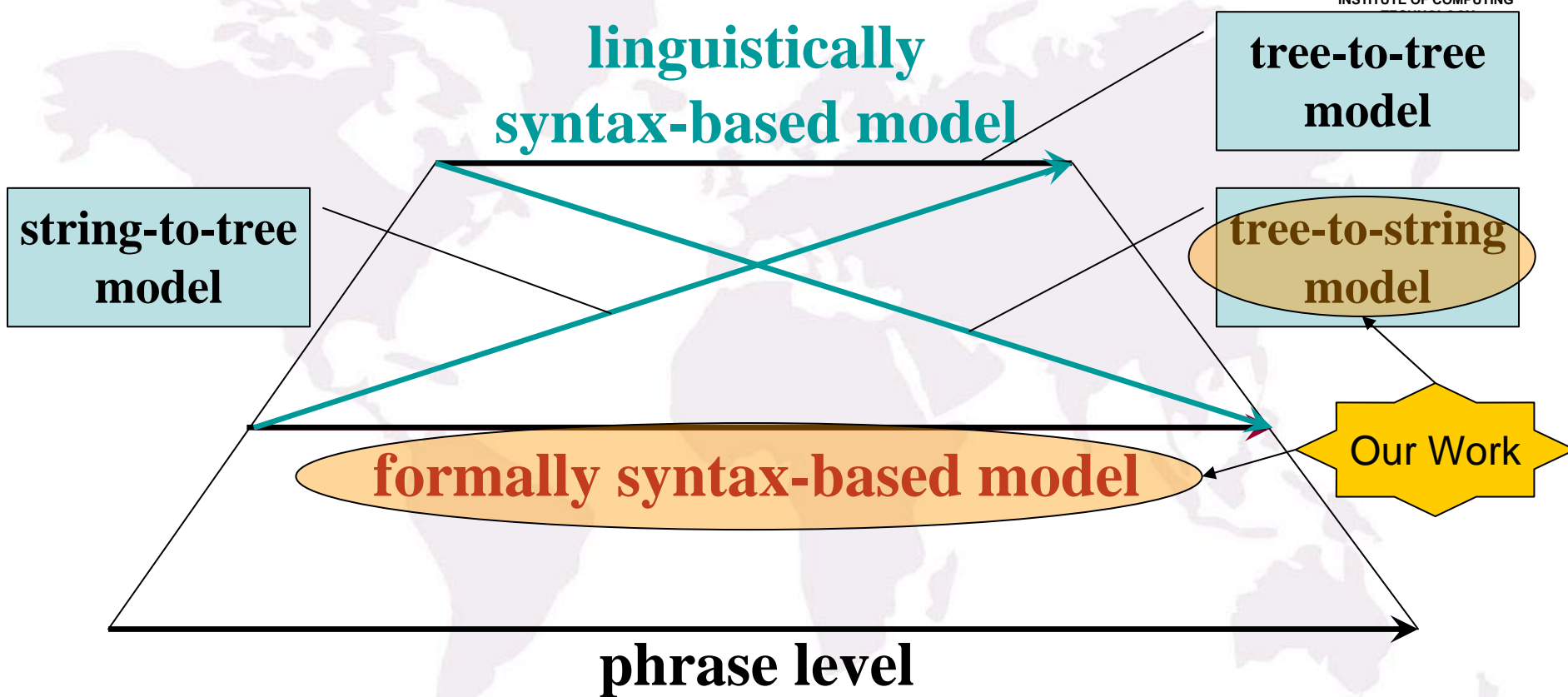
研究论文涉及的领域

- 统计翻译模型
 - 树到串模型
 - 树到树模型
 - 森林模型
 - 规则选择模型
- 词语对齐
- 中文分词与中文词性标注

统计翻译建模

- 基于句法的统计模型是目前机器翻译的研究热点
- 我们提出了统计机器翻译的树到串模型，是目前统计机器翻译研究中的主要几个句法模型之一

基于句法的统计翻译模型





树到串统计翻译模型

我们在计算语言学领域的国际顶级会议上发表了一系列关于树到串统计翻译模型的论文

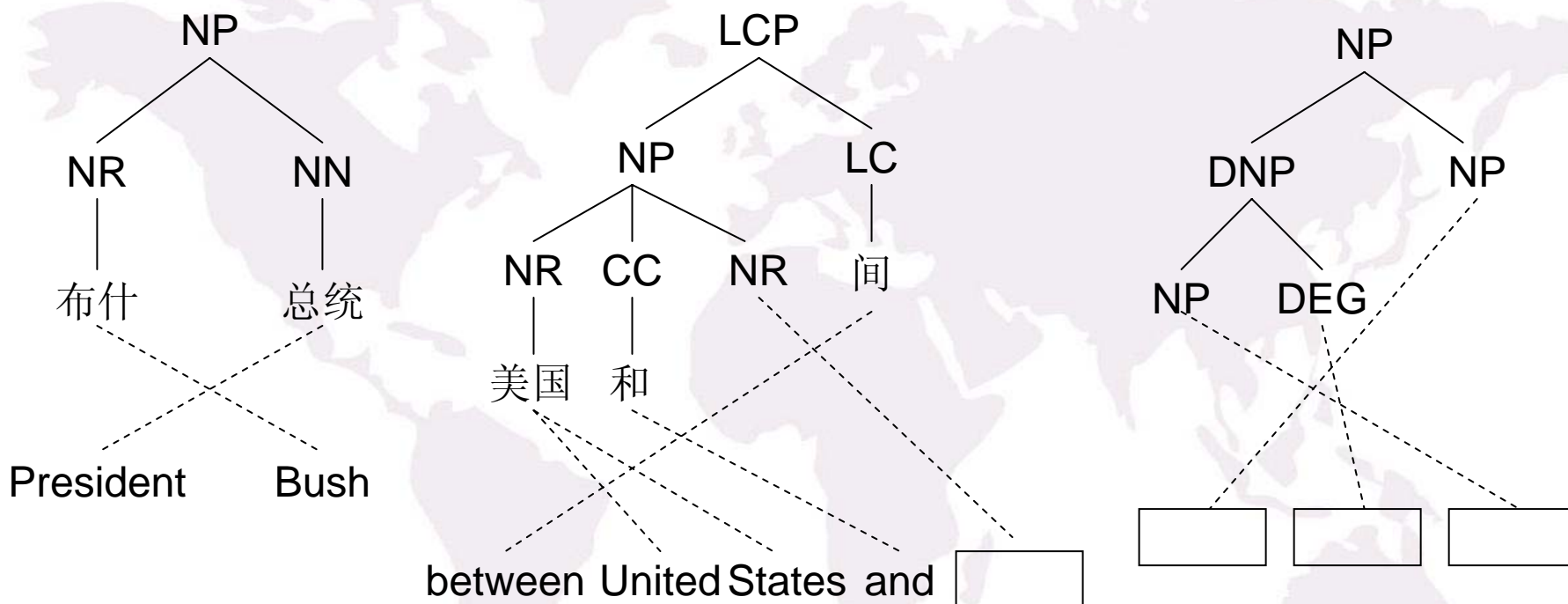
- LIU Yang, et al., ACL2006
- LIU Yang, et al., ACL2007
- MI Haitao, et al., ACL2008
- MI Haitao & HUANG Liang, EMNLP2008
- LIU Qun, et al., EMNLP2008
- LIU Yang, et al., ACL2009

基于树到串对齐模板的统计翻译模型



中科院计算所
INSTITUTE OF COMPUTING
TECHNOLOGY

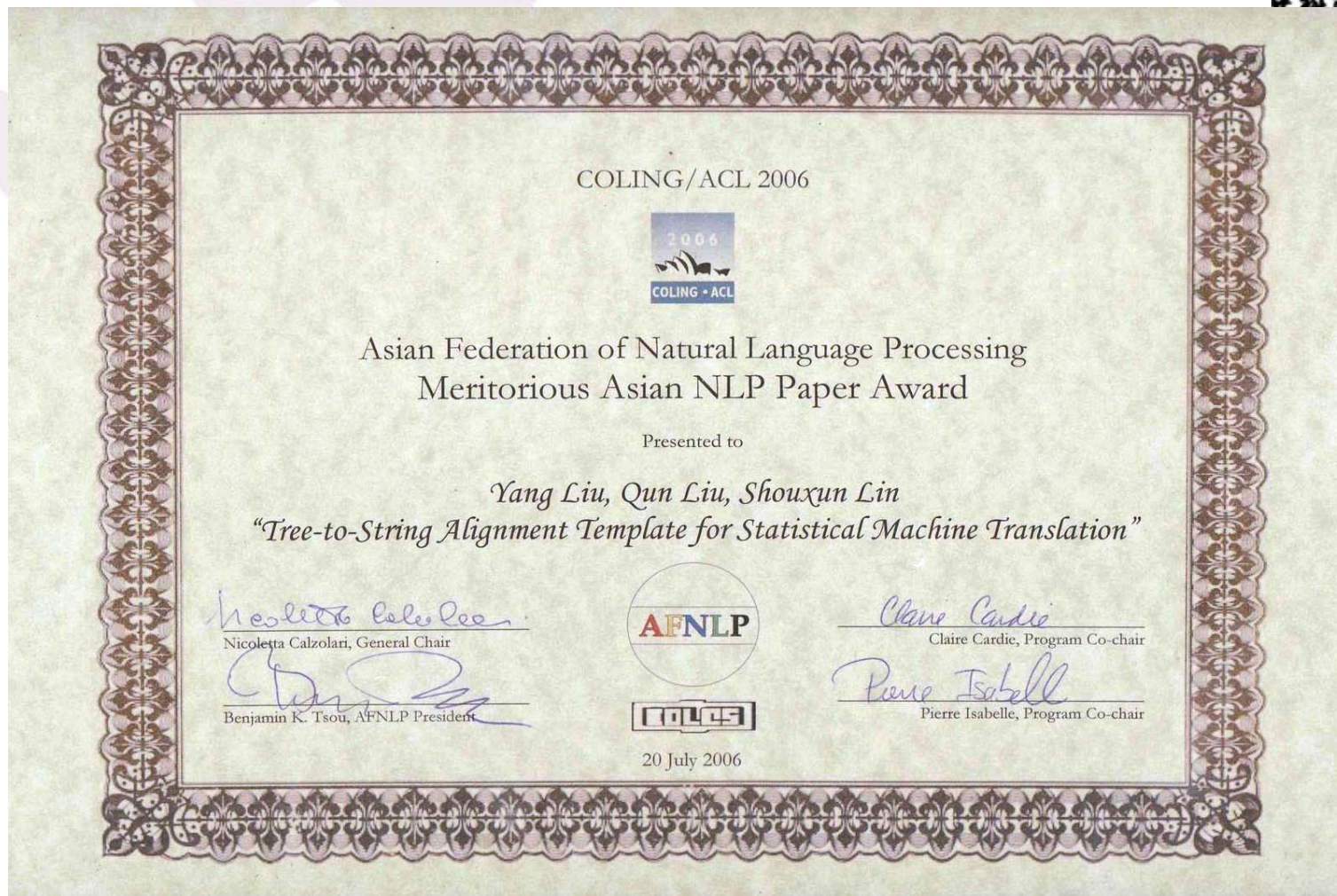
LIU Yang, et al., ACL2006



Meritorious Asian NLP Paper Award

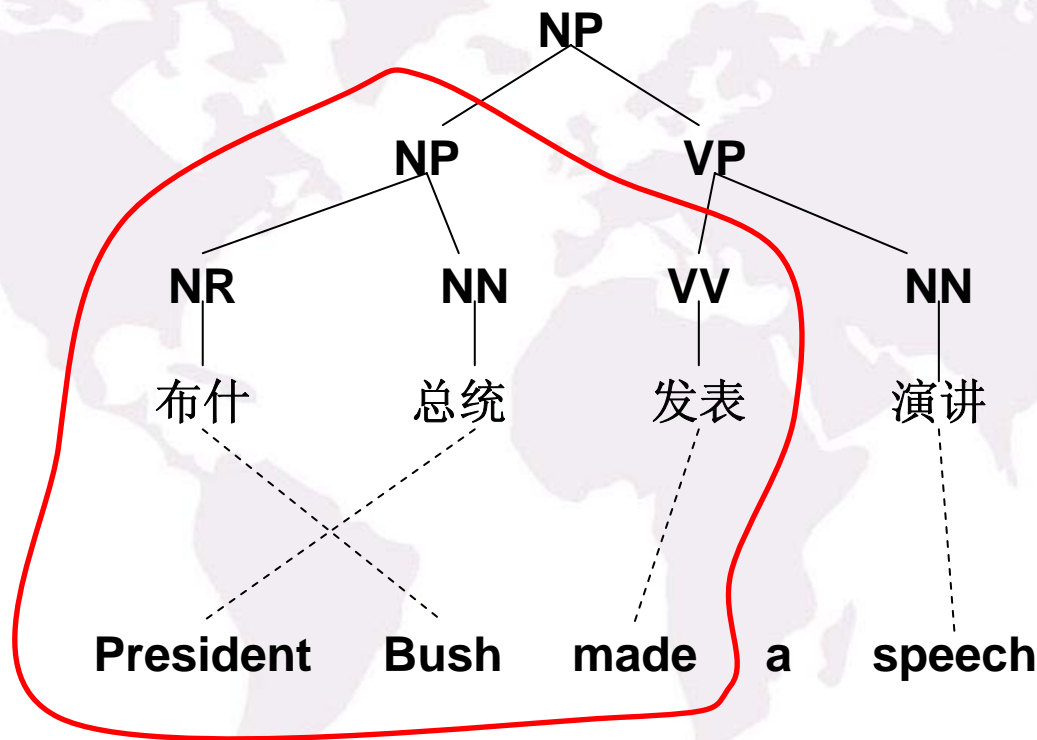


中国科学院
计算所
F COMPUTING
NOLOGY



加入树序列到串规则的树到串模型

LIU Yang, et al., ACL2007

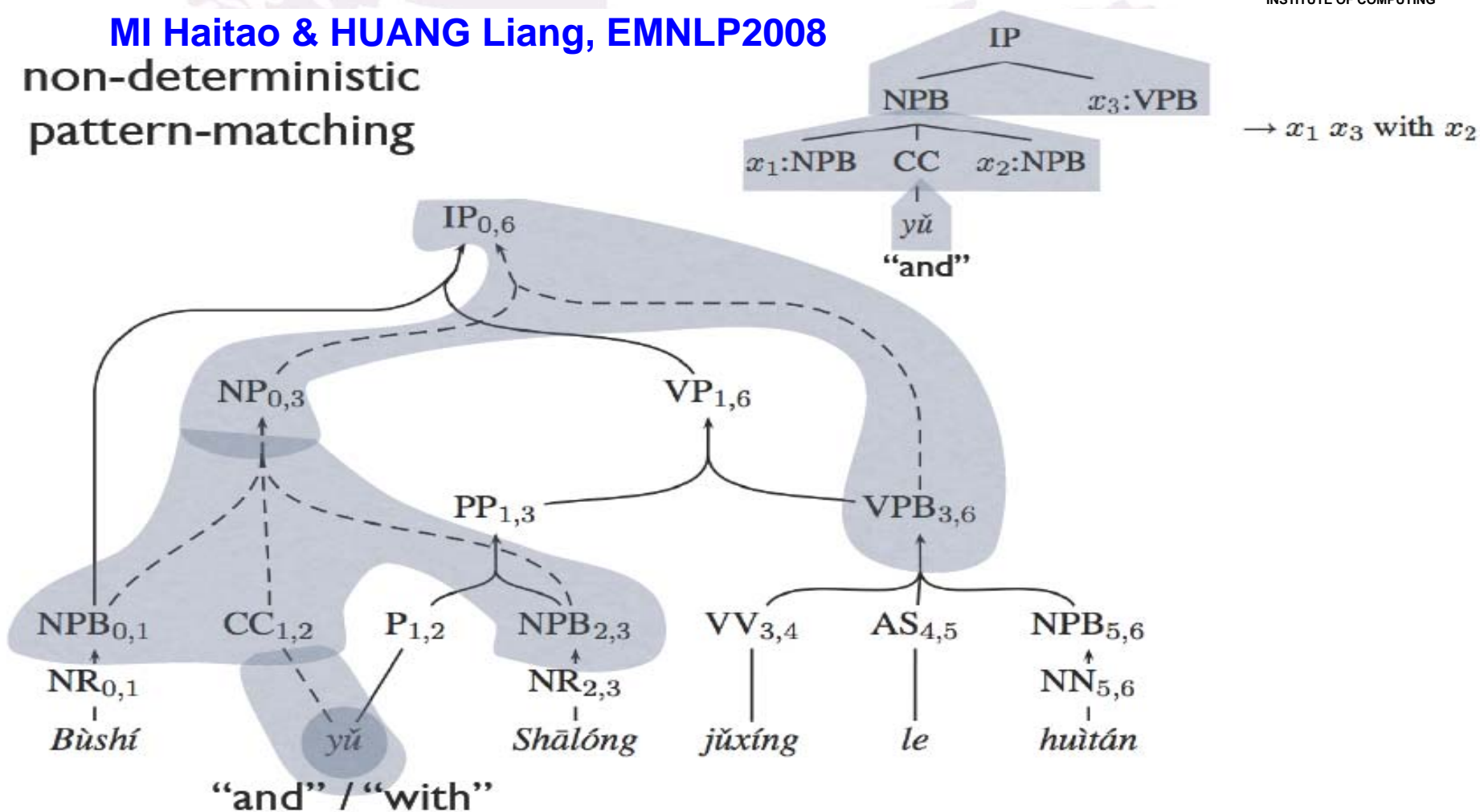


基于句法森林的统计翻译模型

MI Haitao, et al., ACL2008

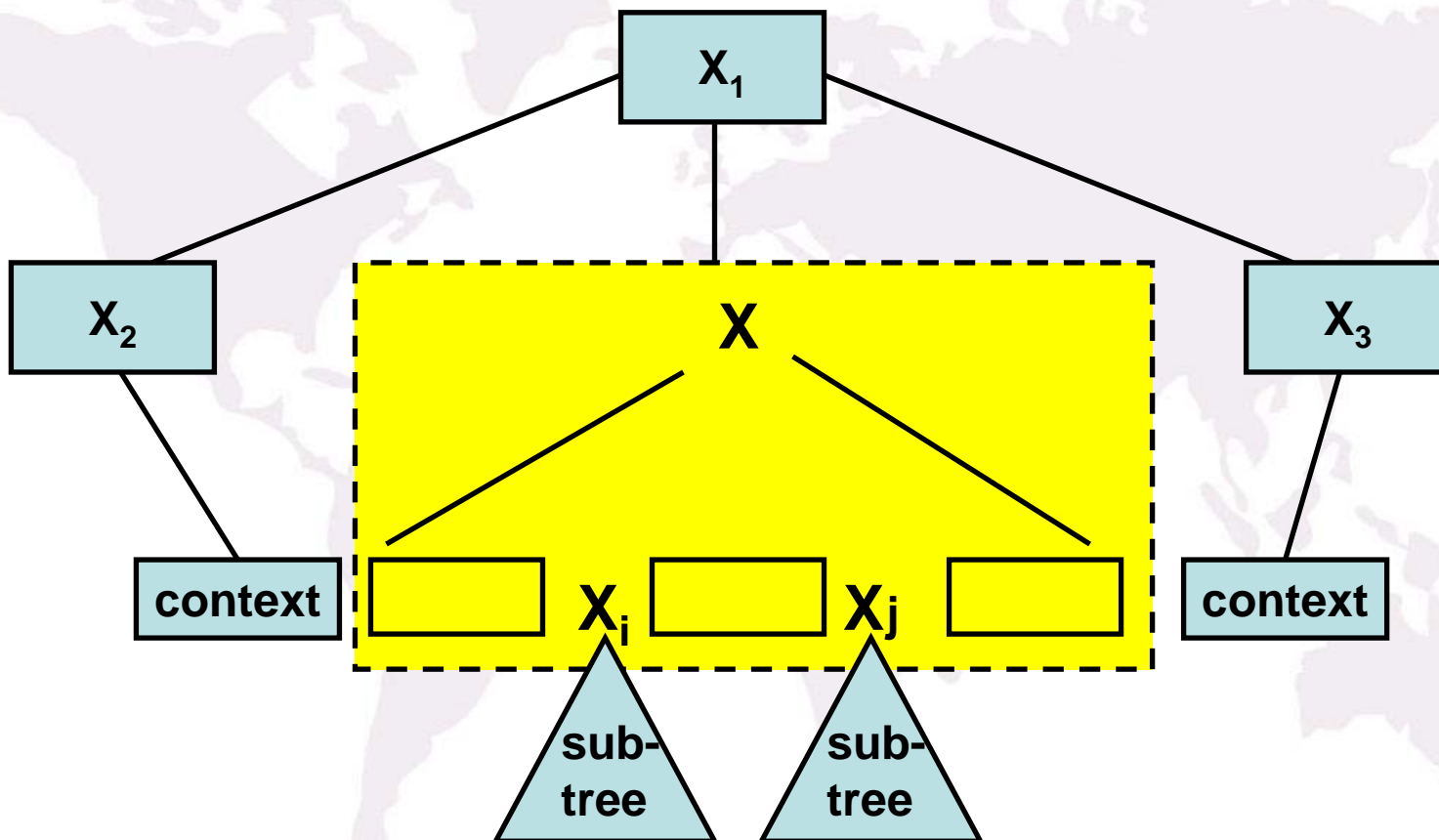
MI Haitao & HUANG Liang, EMNLP2008

non-deterministic
pattern-matching



基于最大熵的规则选择模型

LIU Qun, et al., EMNLP2008



参加机器翻译国际评测结果

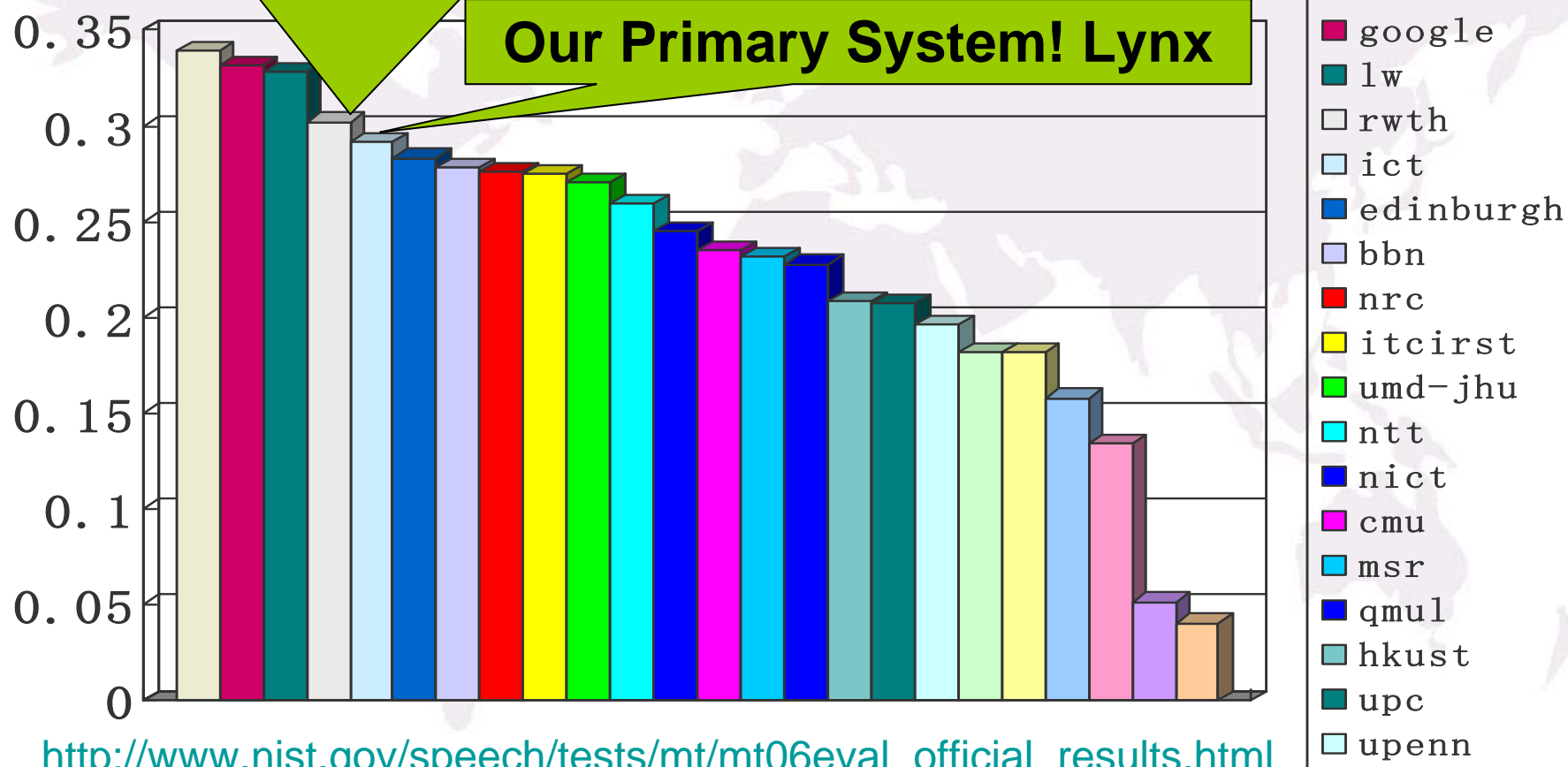
- **NIST**机器翻译评测是国际上影响最大也是竞争最激烈的机器翻译评测
- 我们在**NIST2006**的**24**个参评单位中排名第**5**，在**NIST2009**的**18**个参评单位中总成绩排名第**3**
- 这至今仍然是中国乃至亚洲研究机构在该项评测中取得的最好成绩

Results on NIST 2006 Evaluation: Large Data Track, NIST Subset



One of our contrast system! Bruin

Our Primary System! Lynx

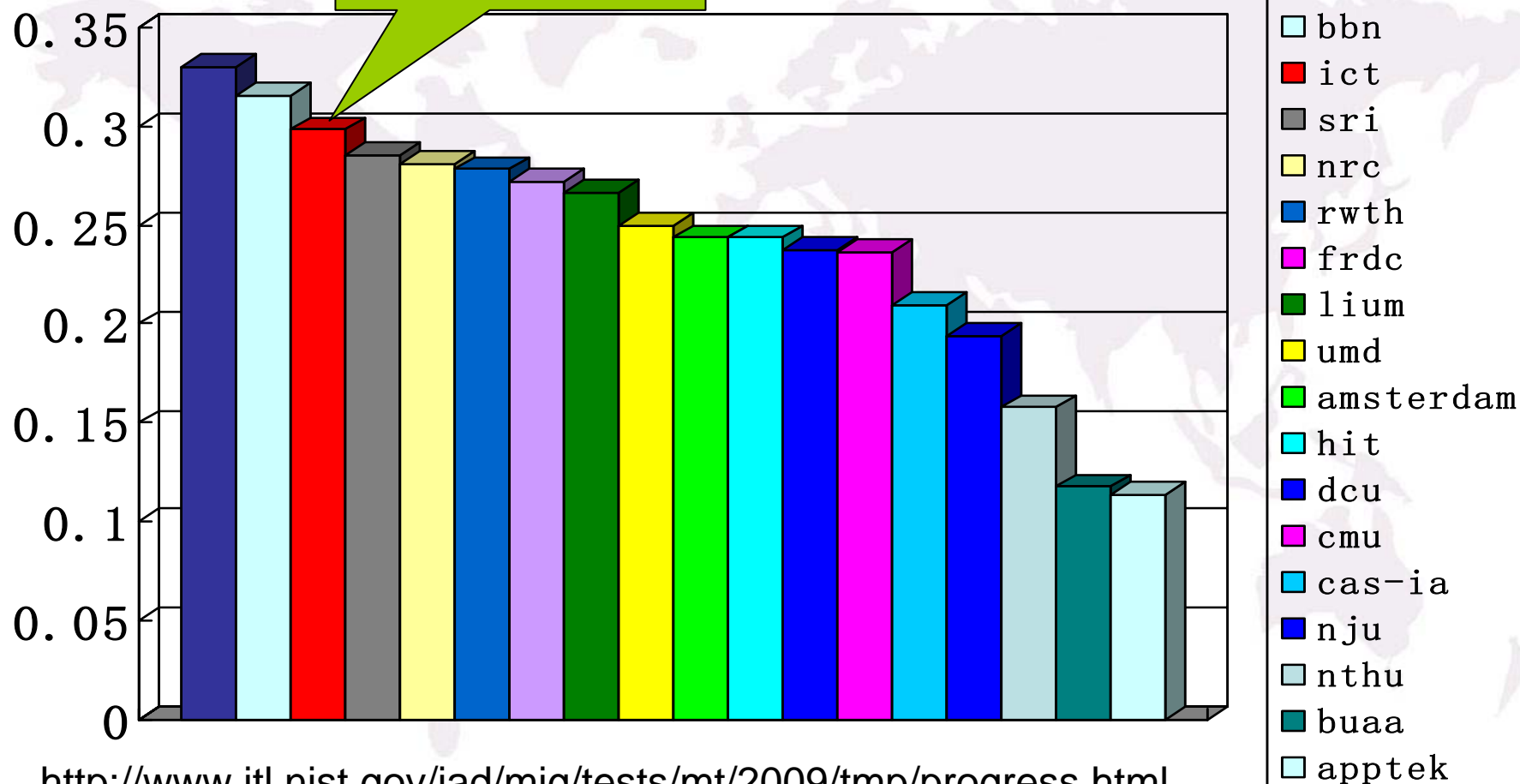


Results on NIST 2009 Evaluation: Progress test data, Chinese-English



中国科学院
INSTITUTE OF COMPUTING
TECHNOLOGY

Our System!



<http://www.itl.nist.gov/iad/mig/tests/mt/2009/tmp/progress.html>

承担国家重要项目

- 八六三重点项目：
 - 面向跨语言搜索的机器翻译关键技术研究
 - 承担单位：中科院计算所、自动化所、软件所、哈工大、厦门大学
- 自然科学基金重点项目：
 - 语言知识与统计模型相结合的机器翻译方法
 - 承担单位：中科院计算所、自动化所、哈工大

机器翻译应用

- 专利翻译
- 移动翻译

专利翻译

- 与某专利数据加工公司合作
- 面向专利翻译的计算机辅助翻译平台
- 概况
 - 服务器-客户端结构
 - 基于短语模型的模型 + 用户自定义模板
 - 八个领域，数百万句子对训练
 - 数万用户自定义模板
 - 用户提供的术语词典
 - 用户评价：准确率**70-85%**（根据领域不同）

移动设备机器翻译

- 正在开发应用于移动翻译设备的机器翻译系统
- 旅游领域
- 中、英、韩三国语言
- 与某跨国公司合作

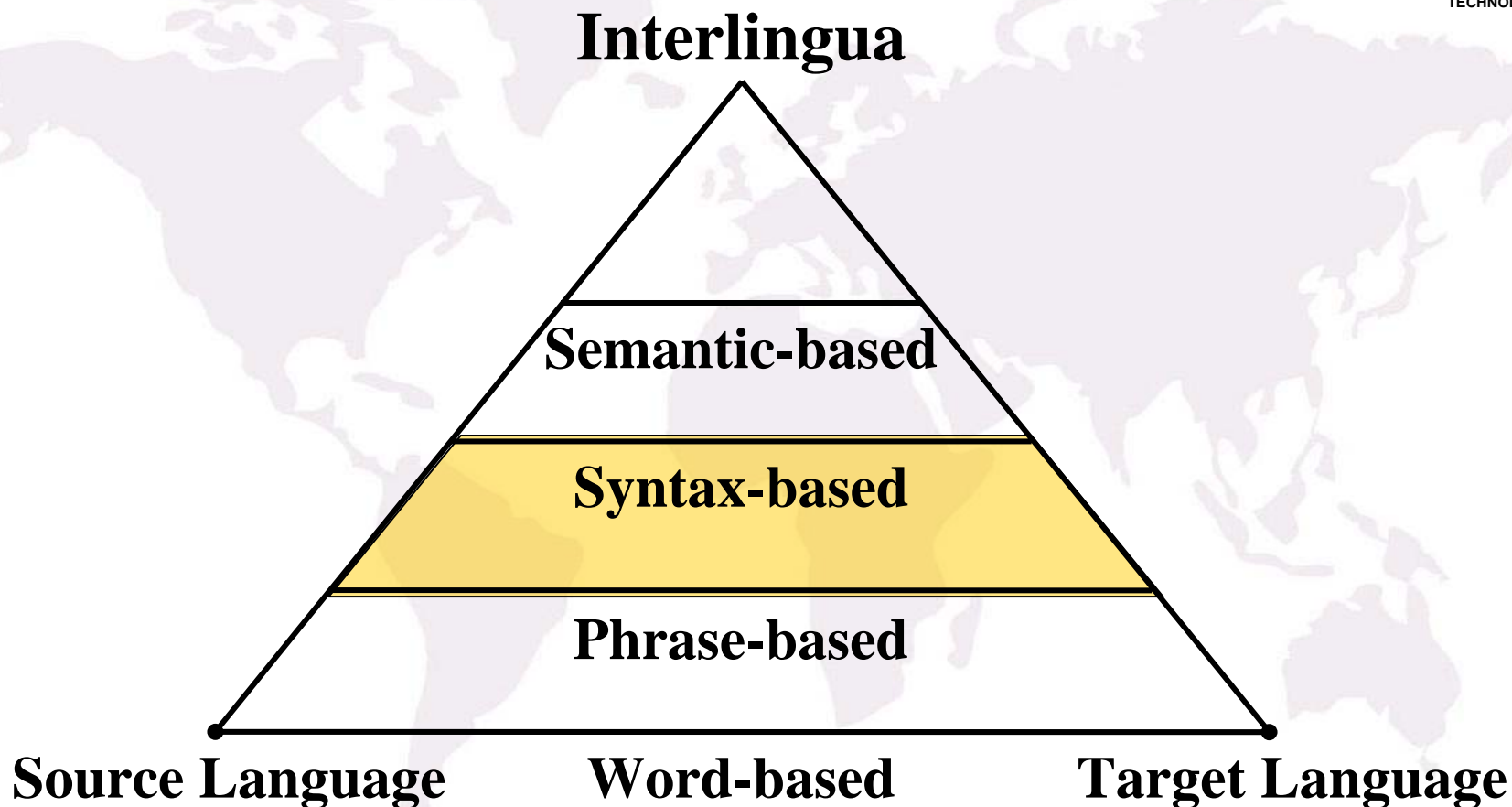
目录

- 统计机器翻译的研究热潮
- 经典的统计机器翻译方法
——基于词的**IBM**模型
- 最成熟的统计机器翻译方法
——基于短语的模型
- 目前统计机器翻译研究的热点
——基于句法的模型
- 中科院计算所的工作
- 统计机器翻译面临的问题和展望

总结

- 综上所述，统计机器翻译的发展可以清理出两条主线的进展
 - 框架模型的进展
 - 信源信道模型
 - 对数线性模型
 - 翻译模型的进展
 - 基于词的模型
 - 基于短语的模型
 - 基于句法的模型

统计翻译模型的进展 (1)



统计翻译模型的进展 (2)

- 我们看到，统计机器翻译与基于规则的机器翻译一样，也存在一个金字塔形的发展过程
- 基于词的模型是**1990**年前后由**IBM**公司提出来的
- 基于短语的模型是**Och**、**Zens**、**Koehn**等人**2003**年前后提出来的，是目前最成熟稳定、也是最普遍采用的模型
- 基于句法的模型是目前的研究热点，虽然很早就有人开展研究，但真正取得较好的结构是在**2005**年以后，典型的工作有**[Chiang 2005]**、**ISI**的工作、中科院计算所的工作。

统计翻译模型的进展 (3)

- 对于基于规则的方法

- 在理想的情况下，如果语言分析（主要是句法分析和语义分析）完全正确，那么，转换的层次越深，可以利用的信息就越多，应该可以达到更好的翻译效果。
- 由于语言分析的过程中总会引入各种各样的错误，而且这些错误还会随着分析层次的加深而逐渐积累起来，从而导致翻译的错误，因此，并不是在越深层次上进行转换所获得的翻译系统性能就越好，而是需要取一个折衷的“最优”。
- 目前，大部分基于规则的机器翻译系统都是在句法或者语义层面进行转换。

统计翻译模型的进展 (4)

- 对于统计方法

- 在统计方法中，沿金字塔向上攀升是一个艰难的过程，从基于词的方法到基于短语的方法，经过了十几年的发展才成熟起来，而基于短语的方法到基于句法的方法，经过几年的发展，虽然成为了目前研究的热点，但还没有成为主流的做法
- 早期很多基于短语的模型和基于句法的模型研究都已失败告终，有些虽然报告取得了较好的结果，但在没有人能够重复的情况下，依然无法被研究界所接受

统计翻译模型的进展 (5)

- 语言知识与统计模型的融合
 - 语言知识与语言模型结合成为研究的趋势
 - 统计模型中如果不引入语言知识，其性能很难再有大的提高
 - 语言知识如果不能与统计模型相结合，就回到了基于规则的老路上去，所面临的知识获取和冲突的问题无法从根本上得到解决

统计翻译模型的进展 (6)

- 在统计模型中加入语言知识是一个非常困难的工作。语言知识的加入会大大增加模型的复杂性，使得系统的性能和能够处理的规模都大大下降；在大部分情况下，由于语言知识的覆盖率和正确率不够，使得系统的整体性能反而下降了。一种融入语言知识的统计模型，只有经过精心设计、反复实验、不断改进，才能取得成功。而且这种成功只有经过其他研究人员的重复才能被学术界广泛接受。
- 在统计模型中加入语言知识很难一蹴而就，只能由浅层知识到深层知识，由简单到复杂，循序渐进地实现

统计翻译模型的进展 (7)

- 基于句法的统计翻译模型目前还是研究热点
- 在统计翻译中应用语义知识也有人开始研究，他们的研究结果表明**WSD**对统计翻译能够有所贡献，但并不太大
 - Dekai Wu, ACL2007
 - Hwee Tou Ng, EMNLP2007
- 上述在统计机器翻译中利用**WSD**的方法本质上只利用了词汇层的语义信息，并没有利用语义结构信息
- **[Shen Libin, ACL2008]**成功地将目标语言（英语）的依存结构引入统计翻译模型中，向语言结构信息的利用跨出了一小步

展望

- 能否建立更有效的基于句法的统计翻译模型？
- 能否建立有效的基于语义的统计翻译模型？
- 更多的丰富多彩、形式多样的语言知识如何融入统计翻译模型？
- 在将语言知识融入统计模型时，如何克服语言知识的错误和低覆盖率对机器翻译带来的负面影响？
- 框架模型是否还有更新的可能？
- 语言模型是否还能进行改进？（比如建立基于句法的语言模型）

资源

- 中文分词和词性标注**ICTCLAS**（开源共享）
- 中文句法分析
 - 短语结构分析
 - 依存分析
 - 达到**State-of-the-art**的水平
 - 还没有提供开源共享
- 机器翻译系统
 - “丝路”（开源共享，五个单位共同开发，短语模型）
 - 机器翻译引擎：**Bruin**、**Sylinus**、**Chario**、.....
几乎实现了目前国际上所有主流的机器翻译模型
 - 机器翻译工具：语言模型、词语对齐、规则抽取、参数调节.....

资源

- 词典
 - **40**多万词的通用词典
 - **200**多万次的专业词典（近**30**个专业）
- 语料库
 - 双语平行语料库
 - 机器翻译评测语料库
 - 双语词语对齐语料库及规范
- 平行语料库加工工具
- 基于**Internet**的双语语料库抽取工具

常用工具列表

- 语言模型**SRI**
- 词语对齐工具**Giza++**
- 机器翻译开源系统**Moses**
- 最大熵工具软件（**Zhang Le**）
- 中文分词和词性标注**ICTCLAS**
- 中文句法分析**ICTParser**、**Stanford Parser**、**Berkeley Parser**
- 机器翻译自动评价工具**MTEval**
- 基于**MapReduce**的分布式并行构架**Hadoop**

现场演示

- 演示系统说明：
 - 汉语 \leftrightarrow 英语
 - 采用**MEBTG**模型[Xiong, Coling-ACL2006]
 - 采用**260**万句对的双语语料库进行训练
 - 主要适合新闻领域翻译
- <http://mitel.ict.ac.cn/bruin/index.php>



谢谢！