

# 统计机器翻译简介

冯志伟  
zwfengde@hotmail.com

NLP课题组例会报告 2008-7-72002-12-6

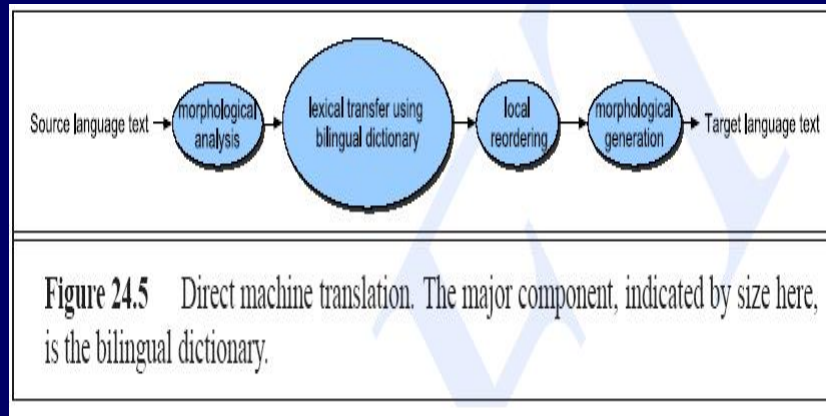
## Vauquois的MT金字塔

The diagram illustrates Vauquois's MT Pyramid, showing the flow of information from source to target text through various linguistic levels. The process starts with 'Source Language Text' (Words) and ends with 'Target Language Text' (Words). The pyramid consists of four levels: Words, Syntactic Structure, Semantic Structure, and Interlingua. The flow is as follows:

- Source Language Text (Words) → Morphological Analysis → Words (Syntactic Structure) → Parsing → Syntactic Structure
- Words (Syntactic Structure) → Syntactic Transfer → Syntactic Structure (Semantic Structure)
- Words (Syntactic Structure) → Direct → Words (Target Language Text)
- Words (Syntactic Structure) → Semantic Generation → Semantic Structure (Interlingua)
- Words (Syntactic Structure) → Semantic Transfer → Semantic Structure (Interlingua)
- Semantic Structure (Interlingua) → Conceptual Analysis → Interlingua
- Semantic Structure (Interlingua) → Conceptual Generation → Semantic Structure (Target Language Text)
- Semantic Structure (Target Language Text) → Syntactic Generation → Syntactic Structure (Target Language Text)
- Syntactic Structure (Target Language Text) → Morphological Generation → Target Language Text (Words)

NLP课题组例会报告 2008-7-72002-12-6

## 直接机器翻译法



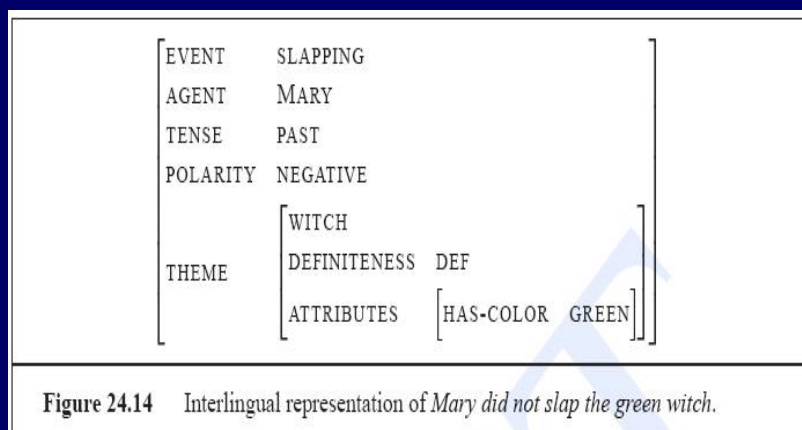
**Figure 24.5** Direct machine translation. The major component, indicated by size here, is the bilingual dictionary.

## 转换机器翻译法

English to Spanish:		
1.	$NP \rightarrow \text{Adjective}_1 \text{ Noun}_2$	$\Rightarrow NP \rightarrow \text{Noun}_2 \text{ Adjective}_1$
Chinese to English:		
2.	$VP \rightarrow PP[+\text{Goal}] V$	$\Rightarrow VP \rightarrow V PP[+\text{Goal}]$
English to Japanese:		
3.	$VP \rightarrow V NP$	$\Rightarrow VP \rightarrow NP V$
4.	$PP \rightarrow P NP$	$\Rightarrow PP \rightarrow NP P$
5.	$NP \rightarrow NP_1 \text{ Rel. Clause}_2$	$\Rightarrow NP \rightarrow \text{Rel. Clause}_2 NP_1$

**Figure 24.13** An informal description of some transformations.

## 中间语言机器翻译法



NLP课题组例会报告

2008-7-72002-12-6

## 统计机器翻译的分类

- 基于平行概率语法的统计机器翻译模型
- 基于噪声信道理论的统计机器翻译模型
  - IBM的Peter Brown等人首先提出
  - 目前影响最大
  - 几乎成为统计机器翻译的同义词
- 基于最大熵的统计机器翻译模型
  - 源于基于特征的自然语言理解
  - Och提出，获ACL2002最佳论文

NLP课题组例会报告

2008-7-72002-12-6

## 统计机器翻译的优缺点

- 优点
  - 无需人工编写规则，利用语料库直接训练得到机器翻译系统；（但可以使用语言资源）
  - 系统开发周期短；
  - 鲁棒性好；
  - 译文质量好；
- 缺点
  - 时空开销大；
  - 数据稀疏问题严重；
  - 对语料库依赖性强；
  - 算法研究不成熟。

NLP课题组例会报告

2008-7-72002-12-6

## 基于平行概率语法的统计机器翻译模型

- 基本思想
  - 两种语言建立一套平行的语法规则，
    - 规则一一对应
    - 两套规则服从同样的概率分布
  - 句法分析的过程决定了生成的过程
- 主要模型
  - Alshawi的基于Head Transducer的MT模型—中心词转录机
  - Synchronous Context-Free Grammar (SCFG) -- 同步上下文无关语法
  - 吴德恺的Inverse Transduction Grammar (ITG) -- 反向转录语法

NLP课题组例会报告

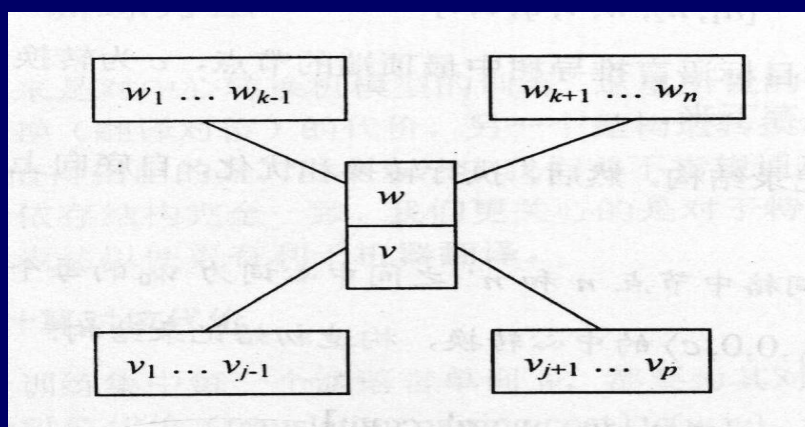
2008-7-72002-12-6



## Head Transducer MT (1)

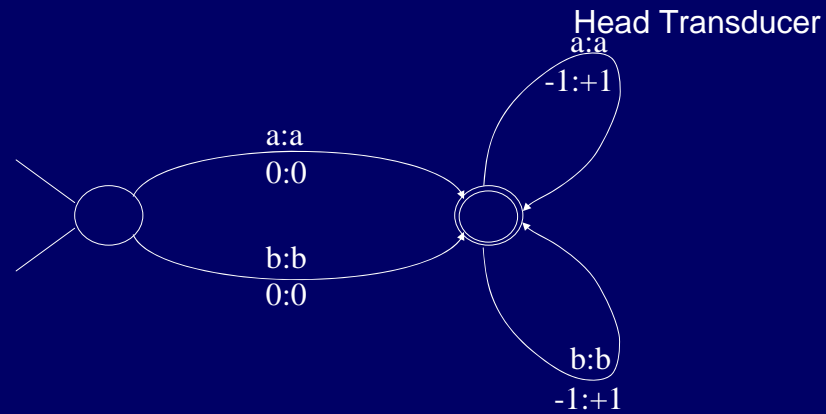
- Head Transducer (中心词转录机) 是一种 Definite State Automata (有限状态自动机)。
- 一个转录的定义:  $\langle q, q', w, v, \alpha, \beta, c \rangle$   
 $q$ : 输入状态,  $q'$ : 输出状态;  $w$ : 输入符号,  $v$ : 输出符号;  $\alpha$ : 输入位置,  $\beta$ : 输出位置;  $c$ : 转录的权重 (代价)。
- 与一般的有限状态识别器的区别:
  - 每一条边上不仅有输入, 而且有输出;
  - 不是从左至右输入, 而是从中心词往两边输入

## 输入符号 $w$ 和输出符号 $v$



## Head Transducer MT(2)

例子：一个可以将任何中心词 {a, b} 组成的串及其前后词转录的



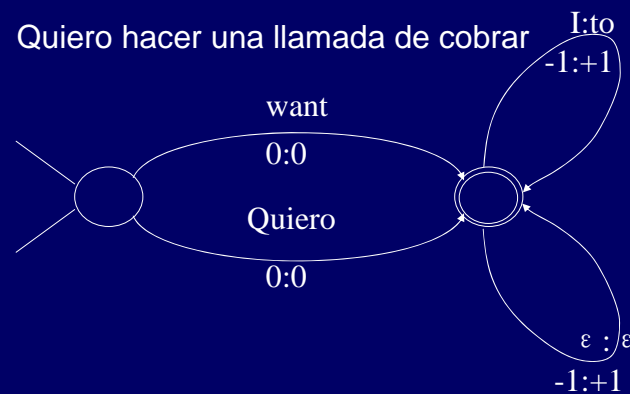
NLP课题组例会报告

2008-7-72002-12-6

## Head Transducer MT(3)

例子：I want to make a collect call →

Quiero hacer una llamada de cobrar



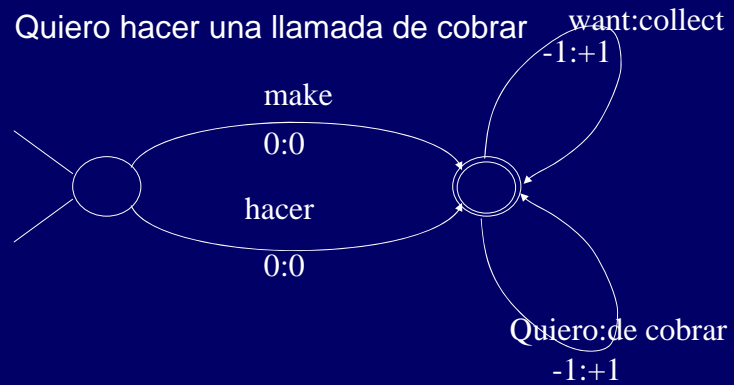
NLP课题组例会报告

2008-7-72002-12-6

## Head Transducer MT(4)

例子: I want to make a collect call →

Quiero hacer una llamada de cobrar



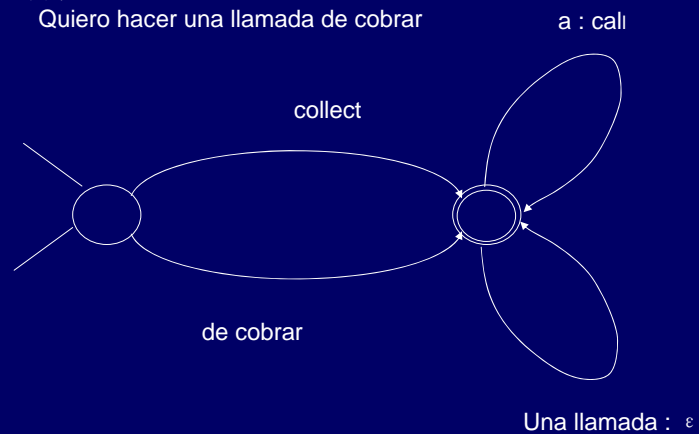
NLP课题组例会报告

2008-7-72002-12-6

## Head Transducer MT(5)

例子: I want to make a collect call →

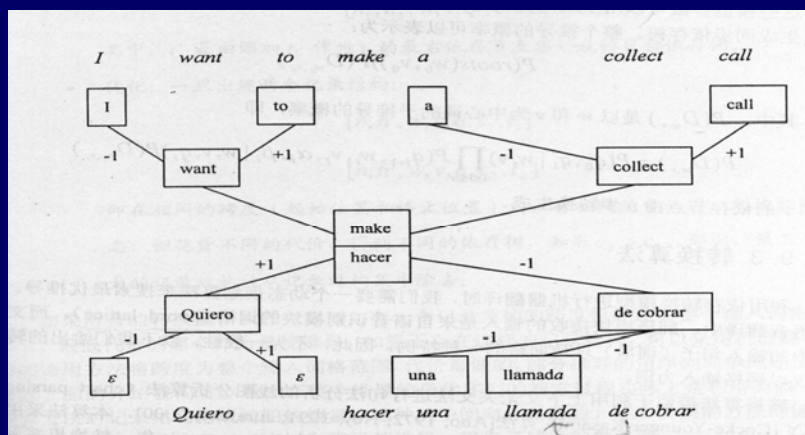
Quiero hacer una llamada de cobrar



NLP课题组例会报告

2008-7-72002-12-6

## Head Transducer MT(6)



NLP课题组例会报告

2008-7-72002-12-6

## Head Transducer MT(7)

- 所有的语言知识（词典、规则）都表现为Head Transducer；
- Head Transducer可以嵌套：一个Head Transducer的边是另一个的识别结果；
- 纯统计的训练方法；
- 对齐的结果是依存树，不使用词性和短语类标记；
- 搜索算法类似于Chart句法分析器。

NLP课题组例会报告

2008-7-72002-12-6

## 基于形式结构(formal structure)的语法

- Synchronous Context-Free Grammar (SCFG, 同步上下文无关语法)
- Inversion Transduction Grammar (ITG, 反向转录语法) proposed by Wu Dekai (1997, 1998)

## SCFG

- 同步上下文无关语法
- Formulated:  $SCFG = (G_1, G_2, \infty)$ 
  - Two CFGs  $G_1, G_2$  and their correspondences
- Or
  - P:  $SCFG = (N, T, P, S)$   
( $A \rightarrow \alpha, B \rightarrow \beta, \infty$ )  
 $X \rightarrow (\alpha, \beta, \infty)$

## SCFG: an example

$S \rightarrow \langle NP_1 VP_2, NP_1 VP_2 \rangle$   
 $VP \rightarrow \langle V_1 NP_2, NP_2 V_1 \rangle$   
 $NP \rightarrow \langle i, watashi \rangle$   
 $NP \rightarrow \langle \text{the box}, \text{wa hako} \rangle$   
 $V \rightarrow \langle \text{open}, \text{akemasu} \rangle$

NLP课题组例会报告

2008-7-72002-12-6

## SCFG: derivation

$\langle S_{[10]}, S_{[10]} \rangle$

$\Rightarrow \langle NP_{[11]} VP_{[12]}, NP_{[11]} VP_{[12]} \rangle$

$\Rightarrow \langle NP_{[11]} V_{[13]} NP_{[14]}, NP_{[11]} NP_{[14]} V_{[13]} \rangle$

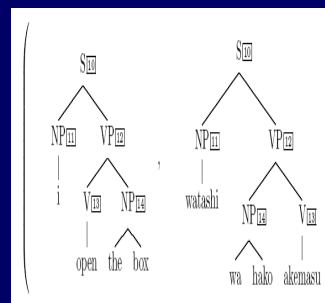
$\Rightarrow \langle i V_{[13]} NP_{[14]}, NP_{[11]} watashi V_{[13]} \rangle$



$\Rightarrow \langle i V_{[13]} NP_{[14]}, watashi NP_{[14]} V_{[13]} \rangle$

$\Rightarrow \langle i \text{ open } NP_{[14]}, watashi NP_{[14]} \text{ akemasu} \rangle$

$\Rightarrow \langle i \text{ open the box}, watashi \text{ wa hako akemasu} \rangle$



NLP课题组例会报告

2008-7-72002-12-6

## 反向转录语法—ITG

- SCFGs in which the links between non-terminals in a production are restricted to two possible configurations:
  - Inverted (反向)
  - Straight (正向)
- Any ITG can be converted into a SCFG of rank two.

$$\begin{aligned}
 S &\rightarrow [NP, VP] \\
 VP &\rightarrow \langle V, NP \rangle \\
 NP &\rightarrow i/watashi \\
 NP &\rightarrow the\text{-}box/wa\text{-}hako \\
 V &\rightarrow open/akemasu
 \end{aligned}$$

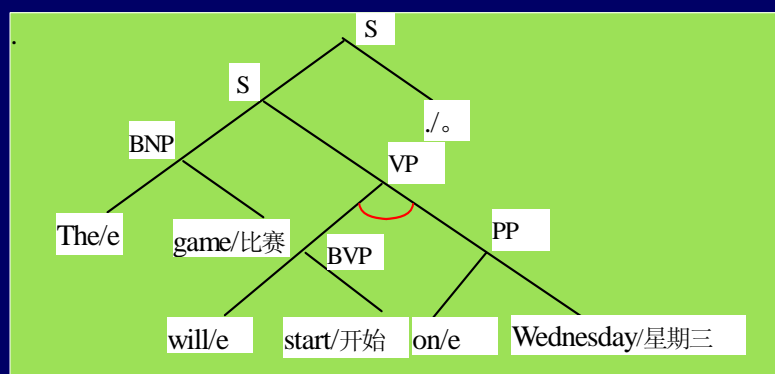
## 括号式转录语法(Bracketing Transduction Grammar – BTG)

$$\begin{array}{lll}
 S \rightarrow \epsilon/\epsilon & A \rightarrow x/\epsilon & A \rightarrow [BC] \\
 A \rightarrow x/y & A \rightarrow \epsilon/y & A \rightarrow \langle BC \rangle
 \end{array}$$

## Inversion Transduction Grammar(1)

The game will start on Wednesday. →

比赛星期三开始。



NLP课题组例会报告

2008-7-72002-12-6

## Inversion Transduction Grammar(2)

### ● 规则形式:

- $A \rightarrow [BC]$
- $A \rightarrow \langle BC \rangle$
- $A \rightarrow x/y$

### ● 产生源语言和目标语言串分别为:

- BC BC: 词序相同, 表示为  $[BC]$
- BC CB: 词序交换, 表示为  $\langle BC \rangle$
- $x/y$ : 词典, 例如: boy/男孩

NLP课题组例会报告

2008-7-72002-12-6

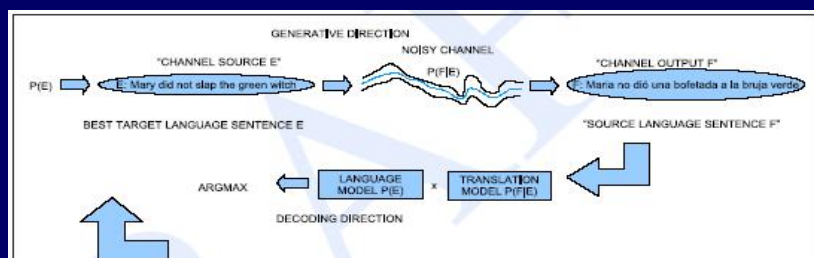
## ITG规则

- $S \rightarrow [S J]$
- $S \rightarrow [BNP VP]$
- $VP \rightarrow \langle BVP PP \rangle$
- $Det \rightarrow the/ \epsilon$
- $N \rightarrow game/比赛$
- $Aux \rightarrow will/ \epsilon$
- $V \rightarrow start/开始$
- $Prep \rightarrow on/ \epsilon$
- $N \rightarrow Wednesday/星期三$

NLP课题组例会报告

2008-7-72002-12-6

## 噪声信道模型



**Figure 24.15** The noisy channel model of statistical MT. If we are translating a source language French to a target language English, we have to think of 'sources' and 'targets' backwards. We build a model of the generation process from an English sentence through a channel to a French sentence. Now given a French sentence to translate, we pretend it is the output of an English sentence going through the noisy channel, and search for the best possible 'source' English sentence.

NLP课题组例会报告

2008-7-72002-12-6

## 噪声信道模型 -cont



- 假设目标语言文本  $T$  是由一段源语言文本  $S$  经过某种奇怪的编码得到的，那么翻译的目标就是要将  $T$  还原成  $S$ ，这也就是就是一个解码的过程。
- 注意，源语言  $S$  是噪声信道的输入语言，目标语言  $T$  是噪声信道的输出语言，与整个机器翻译系统的源语言和目标语言刚好相反。
- 下面，我们仍然按照习惯把机器翻译中的源语言记为  $S$ ，目标语言记为  $T$ 。

## 统计机器翻译基本方程式

$$\hat{T} = \arg \max_T P(T)P(S|T)$$

- P. Brown 称上式为统计机器翻译基本方程式
  - 语言模型:  $P(T)$
  - 翻译模型:  $P(S|T)$
- 语言模型反映“ $T$  像一个句子”的程度: 流利度
- 翻译模型反映“ $T$  像  $S$ ”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，因为后者不考虑流利度，容易导致一些不好的译文。

## 语言模型与翻译模型

- 考虑汉语动词“打”的翻译：有几十种对应的英语词译文：  
打人，打饭，打鱼，打毛衣，打猎，打草稿，.....
- 如果直接采用翻译模型，就需要根据上下文建立复杂的上下文条件概率模型
- 如果采用噪声信道思想，只要建立简单的翻译模型，可以同样达到目标词语选择的效果：
  - 翻译模型：不考虑上下文，只考虑单词之间的翻译概率
  - 语言模型：根据单词之间的同现选择最好的译文词

NLP课题组例会报告

2008-7-72002-12-6

## 统计机器翻译的三个问题

- 三个问题：
  - 语言模型 $P(T)$ 的参数估计
  - 翻译模型 $P(S|T)$ 的参数估计
  - 解码（搜索）算法

NLP课题组例会报告

2008-7-72002-12-6

## 语言模型

- 把一种语言理解成是产生一个句子的随机事件
- 语言模型反映的是一个句子在一种语言中出现的概率
- 语言模型
  - N元语法
 
$$P(T)=p(w_0)*p(w_1/w_0)*...*p(w_n/w_{n-1}...w_{n-N})$$
  - 链语法：可以处理长距离依存关系
  - PCFG：要使用句法标记

NLP课题组例会报告

2008-7-72002-12-6

## 翻译模型与对齐

- 引入隐含变量：对齐  $A$  (*alignment*)

$$P(S|T) = \sum_A P(S, A|T)$$

- $P(S|T)$ 的计算转化为 $P(S,A|T)$ 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

NLP课题组例会报告

2008-7-72002-12-6

## IBM Model

- 对 $P(S,A|T)$ 的估计
- 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（词序）的概率
- IBM Model 3加入了一个词翻译成多个词（多义词）的概率
- IBM Model 4
- IBM Model 5

NLP课题组例会报告

2008-7-72002-12-6

## IBM Model 3

- 对于句子中每一个英语单词 $e$ ，选择一个产出率 $\phi$ ，其概率为 $n(\phi|e)$ ；
- 对于所有单词的产出率求和得到 $m$ -prime；
- 按照下面的方式构造一个新的英语单词串：删除产出率为0的单词，复制产出率为1的单词，复制两遍产出率为2的单词，依此类推，产出率 $>2$ 的英语单词是多义词，这样，含有多义词的单词串就会被多次复制，形成不同的法语串；
- 在这 $m$ -prime个单词的每一个后面，决定是否插入一个空单词NULL，插入和不插入的概率分别为 $p_1$ 和 $p_0$ ；
- $\phi_0$ 为插入的空单词NULL的个数。
- 设 $m$ 为目前的总单词数： $m$ -prime+ $\phi_0$ ；
- 根据概率表 $t(f|e)$ ，将每一个单词 $e$ 替换为法语单词 $f$ ；
- 对于不是由空单词NULL产生的每一个法语单词（它们是单义词），根据概率表 $d(j|i,l,m)$ ，赋予一个位置。这里 $j$ 是法语单词在法语串中的位置， $i$ 是产生当前这个法语单词的对应英语单词在英语句子中的位置， $l$ 是英语串的长度， $m$ 是法语串的长度；
- 如果任何一个目标语言位置被多重登录（含有一个以上单词），则返回失败；
- 给空单词NULL产生的单词（它们是多义词）赋予一个法语单词的位置。这些位置必须是空位置（没有被占用）。任何一个赋值都被认为是等概率的，概率值为 $1/\phi_0$ 。
- 最后，读出后选的各个法语串，法语串的概率为上述每一步概率的乘积，概率最大的法语串为翻译结果。

NLP课题组例会报告

2008-7-72002-12-6

## 翻译模型的参数训练

- Viterbi Training (对比: EM Training)
  1. 给定初始参数;
  2. 用已有的参数求最好 (Viterbi) 的对齐;
  3. 用得到的对齐重新计算参数;
  4. 回到第二步, 直到收敛为止。
- IBM Model 1: 存在全局最优
- IBM Model 2~5: 不存在全局最优, 初始值取上一个模型训练的结果

NLP课题组例会报告

2008-7-72002-12-6

## 统计机器翻译的解码

- 借鉴语音识别的搜索算法: 堆栈搜索
- 参数空间极大, 搜索不能总是保证最优
- 从错误类型看, 只有两种:
  - 模型错误: 概率最大的句子不是正确的句子
  - 搜索错误: 没有找到概率最大的句子
- 后一类错误只占总错误数的5% (IBM)
- 搜索问题不是瓶颈

NLP课题组例会报告

2008-7-72002-12-6

## 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

NLP课题组例会报告

2008-7-72002-12-6

## 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算法
  - 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：最大熵方法

NLP课题组例会报告

2008-7-72002-12-6

## 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		

NLP课题组例会报告

2008-7-72002-12-6

## JHU的1999年夏季研讨班

- 由来
  - IBM的实验引起了广泛的兴趣
  - IBM的实验很难重复：工作量太大
- 目的
  - 构造一个统计机器翻译工具（EGYPT）并使它对于研究者来说是可用的（免费传播）；
  - 在研讨班上用这个工具集构造一个捷克语—英语的机器翻译系统；
  - 进行基准评价：主观和客观；
  - 通过使用形态和句法转录机改进基准测试的结果；
  - 在研讨班最后，在一天之内构造一个新语对的翻译器。
- JHU夏季研讨班大大促进了统计机器翻译的研究

NLP课题组例会报告

2008-7-72002-12-6

## EGYPT工具包

- EGYPT的模块
  1. **GIZA**: 这个模块用于从双语语料库中抽取统计知识（参数训练）
  2. **Decoder**: 解码器，用于执行具体的翻译过程（在信源信道模型中，“翻译”就是“解码”）
  3. **Cairo**: 整个翻译系统的可视化界面，用于管理所有的参数、查看双语语料库对齐的过程和翻译模型的解码过程
  4. **Whittle**: 语料库预处理工具
- EGYPT可在网上免费下载，成为SMT的基准

NLP课题组例会报告

2008-7-72002-12-6

## EGYPT工具包的性能

“当解码器的原形系统在研讨班上完成时，我们很高兴并惊异于其速度和性能。1990年代早期在IBM公司举行的DARPA机器翻译评价时，我们曾经预计只有很短（10个词左右）的句子才可以用统计方法进行解码，即使那样，每个句子的解码时间也可能是几个小时。在早期IBM的工作过去将近10年后，摩尔定律、更好的编译器以及更加充足的内存和硬盘空间帮助我们构造了一个能够在几秒钟之内对25个单词的句子进行解码的系统。为了确保成功，我们在搜索中使用了相当严格的域值和约束，如下所述。但是，解码器相当有效这个事实为这个方向未来的工作预示了很好的前景，并肯定了IBM的工作的初衷，即强调概率模型比效率更重要。”

——引自JHU统计机器翻译研讨班的技术报告

NLP课题组例会报告

2008-7-72002-12-6

## 对IBM方法的改进

- IBM方法的问题
  - 不考虑结构：能否适用于句法结构差别较大的语言？
  - 数据稀疏问题严重
- 后续的改进工作
  - 王野翊的改进—基于结构的对齐模型
  - Yamada和Knight的改进—基于句法的统计翻译模型
  - Och等人的改进—最大熵模型

NLP课题组例会报告

2008-7-72002-12-6

## 基于结构的对齐模型(1)

- 王野翊（CMU）基于结构的对齐模型（structure-based alignment model）
- 背景：德英口语翻译系统
  - 语法结构差异较大
  - 数据稀疏（口语语料的训练数据有限）
- 改进：两个层次的对齐模型
  - 粗对齐(rough alignment)：短语之间的对齐，  

$$E = E_0 E_1 E_2 \dots E_n \Leftrightarrow G = G_0 G_1 G_2 \dots G_m$$
  - 细对齐(detailed alignment)：短语内词的对齐  

$$E_1 = (e_{11} e_{12} \dots) \Leftrightarrow G_1 = (g_{11} g_{12} \dots)$$

NLP课题组例会报告

2008-7-72002-12-6

## 基于结构的对齐模型(2)

### ● 文法推导

- 词语聚类(**clustering**): 基于互信息的方法, 把词语聚类为等价类。
- 短语归并(**phrasing**): 满足一定阈值的等价类序列形成短语, 德语短语译文中单词之间的结合程度必须足够紧密; 如果候选德语译文短语中的单词的跨度过大, 则被过滤。

### ● 优点

- 机器翻译的正确率提高: 错误率降低了11%
- 提高了整个系统的效率: 搜索空间更小
- 缓解了因口语数据缺乏导致的数据稀疏问题

## 基于句法的统计翻译模型(1)

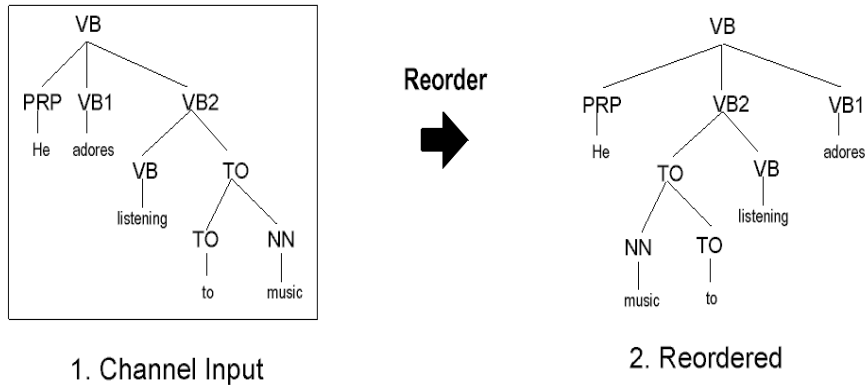
### ● Yamada和Knight提出基于句法的统计翻译模型 (Syntax-based Statistical Translation Model) :

- 输入是源语言句法树
- 输出是目标语言句子

### ● 翻译的过程:

- 每个内部结点的子结点随机地重新排列: 排列概率
- 在每一个结点的左边或右边随机插入一个单词
  - 左、右插入和不插入的概率取决于父结点和当前结点标记
  - 插入哪个词的概率只与被插入词有关, 与位置无关 ⊕
- 对于每一个叶结点进行翻译: 词对词的翻译概率
- 输出译文句子

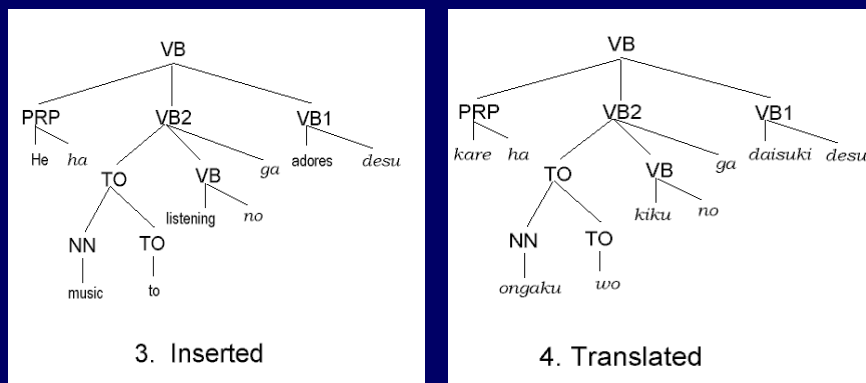
## 基于句法的统计翻译模型 (2)



NLP课题组例会报告

2008-7-72002-12-6

## 基于句法的统计翻译模型 (3)



*kare ha ongaku wo kiku no ga daisuki desu*

5. Channel Output

NLP课题组例会报告

2008-7-72002-12-6

## 基于句法的统计翻译模型(4)

- 调序操作(Reordering operation)
  - $A \rightarrow (A_1 A_2 A_3, A_1 A_3 A_2)$
- 插入操作(Insertion operation)
  - $A \rightarrow (w A_1, A_1)$
- 翻译操作(Translating operation)
  - $A \rightarrow (x, y)$

## 基于句法的统计翻译模型(4)

original order	reordering	P(reorder)
PRP VB1 VB2	PRP VB1 VB2	0.074
	PRP VB2 VB1	0.723
	VB1 PRP VB2	0.061
	VB1 VB2 PRP	0.037
	VB2 PRP VB1	0.083
	VB2 VB1 PRP	0.021
VB TO	VB TO	0.251
	TO VB	0.749
TO NN	TO NN	0.107
	NN TO	0.893
⋮	⋮	⋮

r□table

parent	TOP	VB	VB	VB	TO	TO	⋮
node	VB	VB	PRP	TO	TO	NN	⋮
P(None)	0.735	0.687	0.344	0.709	0.900	0.800	⋮
P(Left)	0.004	0.061	0.004	0.030	0.003	0.096	⋮
P(Right)	0.260	0.252	0.652	0.261	0.007	0.104	⋮

n□table

w	P(ins w)
ha	0.219
ta	0.131
wo	0.099
no	0.094
ni	0.080
te	0.078
ga	0.062
⋮	⋮
desu	0.0007
⋮	⋮

E	adores	he	i	listening	music	to	⋮
J	daisuki 1.000	kare 0.952 NULL 0.016 nani 0.005 da 0.003 shi 0.003	NULL 0.471 watasi 0.111 kare 0.055 shi 0.021 nani 0.020	kiku 0.333 kii 0.333 mi 0.333	ongaku 0.900 naru 0.100	ni 0.216 NULL 0.204 to 0.133 no 0.046 wo 0.038	⋮

t□table

Table 1: Model Parameter Tables

## 调序

### Reordered-table (r-table)

- PRP VB1 VB2 → PRP VB2 VB1:  
P (reorder) = 0.723
- VB TO → TO VB: P (reorder) = 0.749
- TO NN → NN TO: P (reorder) = 0.893
- 调序后整个结构树的概率为:  
 $0.723 \times 0.749 \times 0.893 = 0.484$

## 插入结点

### node-table(1)

- n-table(1):假定在输入句子中有一个看不见的空间，它生成的单词随机地分布在输出句子的任何位置上，插入单词的位置根据句中法树中结点的不同来决定，其概率见表 n-table(1)。
- P( None | Parent=TOP, Node=VB) = 0.735
- P( Right | Parent=VB, Node=PRP) = 0.652
- P( Right | Parent=VB, Node=VB2) = 0.252
- P( Right | Parent=VB, Node=VB1) = 0.252
- P( None | Parent=VB, Node=TO ) = 0.709
- P( Right | Parent=VB2, Node=VB) = 0.252
- P( Right | Parent=TO, Node=NN) = 0.800
- P( Right | Parent=TO, Node=TO) = 0.900

## 插入结点 node-table (2)

- 插入虚词的概率见n-table(2)
- $P(\text{ha}) = 0.219$
- $P(\text{no}) = 0.094$
- $P(\text{ga}) = 0.062$
- $P(\text{desu}) = 0.0007$

## 整个树的插入概率

- 插入结点位置的概率与插入结点概率得乘积，就是整个树的插入概率。
- 例句中整个树的插入概率  

$$= [0.735 \times 0.652 \times 0.252 \times 0.252 \times 0.709$$

$$\times 0.252 \times 0.800 \times 0.900]$$

$$\times [0.219 \times 0.094 \times 0.062 \times 0.0007]$$

$$= 3.498e-9$$

## 翻译概率t-table

- 源语言单词翻译为目标语言单词的概率在t-table中给出。
- 例句的单词翻译概率= $P(\text{kare} | \text{he}) \times P(\text{ongaku} | \text{music}) \times P(\text{wo} | \text{to}) \times P(\text{kiku} | \text{listening}) \times P(\text{daisuki} | \text{adores})$   
 $= 0.952 \times 0.900 \times 0.038 \times 0.333 \times 1.000$   
 $= 0.0108$

## 调序-插入-翻译的联合概率

- 联合概率等于调序-插入-翻译概率的乘积。
- 例句的联合概率= $0.484 \times 3.498e-9 \times 0.0108 = 1.828e-11$

## EM算法概率估计的步骤

- 初始化所有的概率表（概率按均匀分布设置）
- 清除每种操作的计数器
- 对于每一个英语句子和日语句子的偶对执行如下操作：
  - 计算reorder, insert和translation三种操作每种可能的组合方式的概率；
  - 把每种操作的概率分别加到相应的计数器中
- 根据计数器的值，重新计算r-table, n-table和t-table中的概率值，并修改这些概率值
- 重复上述第2步到第4步操作，使概率收敛到一定程度

NLP课题组例会报告

2008-7-72002-12-6

## 统计翻译模型的训练

- 训练
  - 英日词典例句2121对，平均句长日9.7和英6.9
  - 词汇量：英语3463，日语3983，大部分词只出现一次
  - 使用Brill's POS Tagger和Collins' Parser
  - 使用中心词词性标记取得短语标记
  - 压扁句法树：中心词相同的句法子树合并
  - EM训练20遍迭代；IBM Model 5用20遍迭代

NLP课题组例会报告

2008-7-72002-12-6

## 基于句法统计翻译模型的结果

### ● 结果

	Alignment ave. score	Perfect sents
<b>Knight Model</b>	<b>0.582</b>	<b>10</b>
<b>IBM Model 5</b>	<b>0.431</b>	<b>0</b>

### ● 困惑度Perplexity:

knight Model: 15.70

IBM Model: 9.84 (Over-fitting)

## Och等人对IBM方法的改进

- 著名语音翻译系统VerbMobil的一个模块
- 对IBM方法的改进
  - 基于类的模型：词语自动聚类：各400个类
  - 语言模型：基于类的五元语法，回退法平滑
  - 翻译模型：基于对齐模板的方法
    - 短语层次对齐
    - 词语层次对齐
  - 短语划分：使用动态规划的方法

# Och等人提出的对齐模板(1)

对齐模板

T3	.	.	■	■	■
T2	.	■	.	.	.
T1	■	.	.	.	.
	S1	S2	S3	S4	S5

T1: zwei, drei, vier, fünf, ...  
T2: Uhr  
T3: vormittags, nachmittags, abends, ...

S1: two, three, four, five, ...  
S2: o'clock  
S3: in  
S4: the  
S5: morning, evening, afternoon, ...

# Och等人提出的对齐模板(2)

	under	pressure	from	the	Indian	People's	Party
zai	■						
yindu					■		
renmin						■	
dang							■
de			■				
yali		■					
xia	■						

$\langle X \rightarrow \text{zai } X_{\square} \text{ de yali, } X \rightarrow \text{under pressure from } X_{\square} \rangle$

## 基于最大熵的 统计机器翻译模型(1)

- Och等人提出最大熵模型 (Maximum Entropy, ME) 的思想来源于Papineni提出的基于特征的自然语言理解方法
- 不使用噪声信道思想, 而直接使用统计翻译模型, 因此是一种直接翻译模型
- 是一个比噪声信道模型更具一般性的模型, 噪声信道模型是其一个特例
- 与一般最大熵方法的区别: 使用连续量作为特征

NLP课题组例会报告

2008-7-72002-12-6

## 基于最大熵的 统计机器翻译模型(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, 那么直接翻译概率可以用以下公式模拟:

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

NLP课题组例会报告

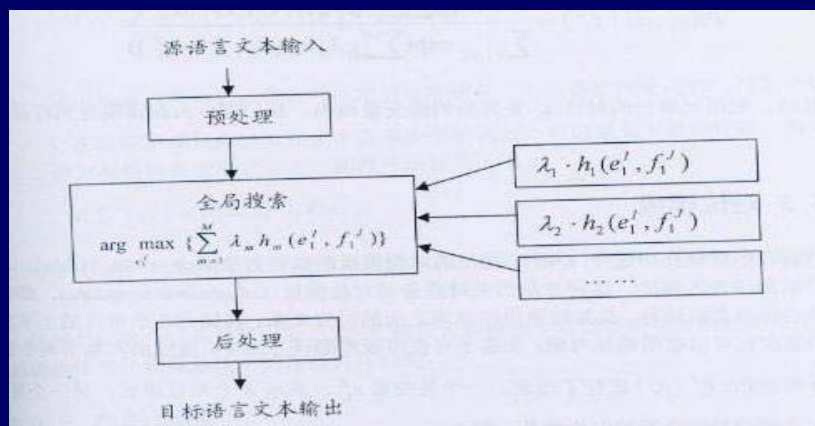
2008-7-72002-12-6

## 基于最大熵的 统计机器翻译模型(3)

对于给定的 $f$ ，其最佳译文 $e$ 可以用以下公式表示：

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

## 基于最大熵的 统计机器翻译模型(4)



## 基于最大熵的 统计机器翻译模型(5)

- 取以下特征和参数时等价于噪声信道模型：
  - 仅使用两个特征
  - $h_1(\mathbf{e}, \mathbf{f}) = p(\mathbf{e})$ : 等价于语言模型
  - $h_2(\mathbf{e}, \mathbf{f}) = p(\mathbf{f}|\mathbf{e})$ : 等价于翻译模型
  - $\lambda_1 = \lambda_2 = 1$

NLP课题组例会报告

2008-7-72002-12-6

## 基于最大熵的 统计机器翻译模型(6)

- 参数训练
 

$$\hat{\lambda}_1^M = \operatorname{argmax}_{\lambda_1^M} \left\{ \sum_{s=1}^S \log p_{\lambda_1^M}(\mathbf{e}_s | \mathbf{f}_s) \right\}$$
- 最优化后验概率准则：区别性训练 (discriminative training)
- 这个判断准则是凸的，存在全局最优
- 考虑多个参考译文：

$$\hat{\lambda}_1^M = \operatorname{argmax}_{\lambda_1^M} \left\{ \sum_{s=1}^S \frac{1}{R_s} \sum_{r=1}^{R_s} \log p_{\lambda_1^M}(\mathbf{e}_{s,r} | \mathbf{f}_s) \right\}$$

NLP课题组例会报告

2008-7-72002-12-6

## 基于最大熵的 统计机器翻译模型(7)

- Och等人的实验(1): 方案
  - 首先将噪声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  $\lambda_1$  和  $\lambda_2$ , 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。

NLP课题组例会报告

2008-7-72002-12-6

## 基于最大熵的 统计机器翻译模型(8)

- Och等人的实验(2): 增加如下特征
  - 句子长度特征(WP): 对于产生的每一个目标语言单词进行词惩罚(word penalty);
  - 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。

NLP课题组例会报告

2008-7-72002-12-6

## 基于最大熵的 统计机器翻译模型(9)

### ● Och等人的实验(2): 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

## 基于最大熵的 统计机器翻译模型(10)

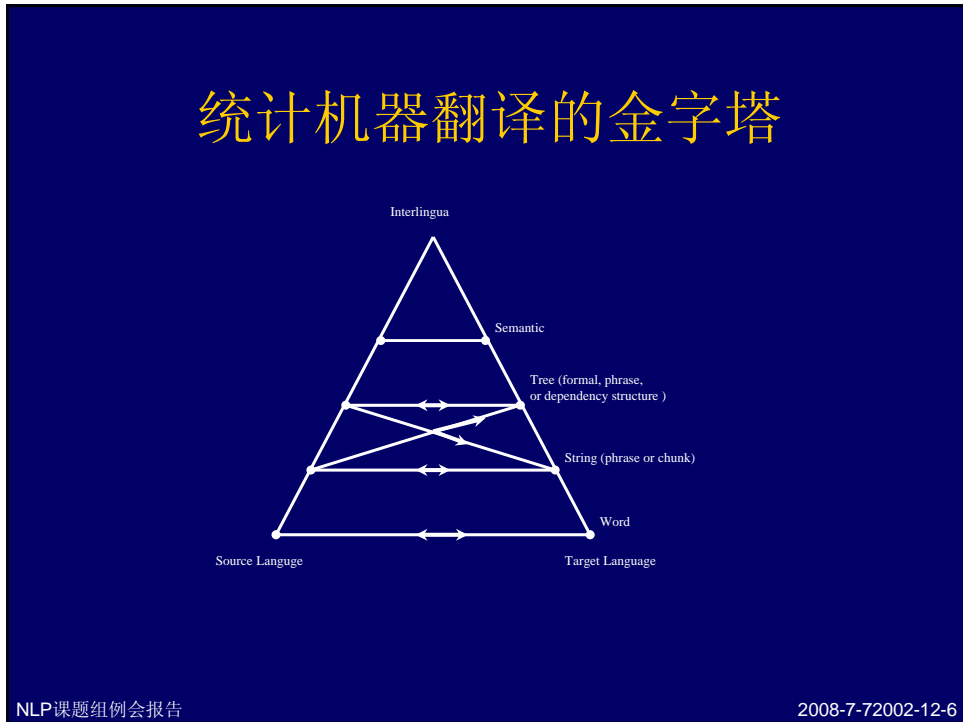
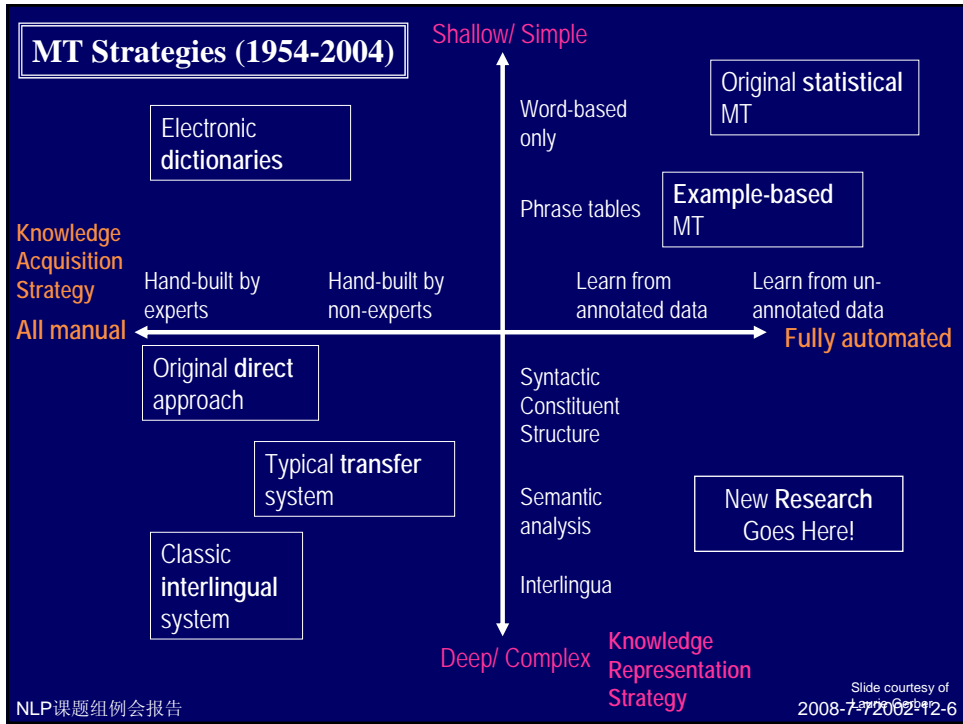
- 经典的噪声信道模型只有在理想的情况下才能达到最优，对于简化的语言模型和翻译模型，取不同的参数值实际效果更好；
- 最大熵方法大大扩充了统计机器翻译的思路；
- 特征的选择更加灵活。

## 统计机器翻译的应用

- 传统机器翻译的应用领域
- 跨语言检索
  - 聂建云使用IBM Model 1进行CLIR
- 机器翻译系统开发的新方向
  - 可以针对未知语言开发机器翻译系统
  - 可以快速开发机器翻译系统

## 总结

- IBM当年的工作是有一定超前性的
- 虽然很多人怀疑统计方法在机器翻译中能否取得成功，但现在这已不再是问题
- 基于平行概率语法的机器翻译方法总体上不成功
- 基于最大熵的方法为统计机器翻译方法开辟了一个新天地



## 参考文献(1)

- [Al-Onaizan 1999] Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith and David Yarowsky (1999). Statistical Machine Translation: Final Report, Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, MD.
- [Alshawi 1998] Alshawi, H., Bangalore, S. and Douglas, S. "Automatic Acquisition of Hierarchical transduction models for machine translation," Proc. 36th Conf. Association of Computational Linguistics, Montreal, Canada, 1998.
- [Berger 1994] Berger, A., P. Brown, S. Della Pietra, V. Della Pietra, J. Gillett, J. Lafferty, R. Mercer, H. Printz, L. Ures, The Candide System for Machine Translation, Proceedings of the DARPA Workshop on Human Language Technology (HLT)
- [Berger 1996] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39-72, March 1996.
- [Brown 1990] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin, A Statistical Approach to Machine Translation, Computational Linguistics, 1990

## 参考文献(2)

- [Brown 1993] Peter. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, Vol 19, No.2, 1993
- [Ker 1997] Sue J. Ker, Jason S. Chang, A Class-based Approach to Word Alignment, Computational Linguistics, Vol. 23, No. 2, Page 313-343, 1997
- [Knight 1999] Kevin Knight, A Statistical Machine Translation Tutorial Workbook. unpublished, prepared in connection with the JHU summer workshop, August 1999. (available at <http://www.clsp.jhu.edu/ws99/projects/mt/wkbk.rtf>).
- [Och 1998] Franz Josef Och and Hans Weber. Improving statistical natural language translation with categories and rules. In Proc. Of the 35th Annual Conf. of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics, pages 985-989, Montreal, Canada, August 1998.
- [Och 1999] F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In Proc. of the Joint SIGDAT Conf. On Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20-28, University of Maryland, College Park, MD, June 1999.

## 参考文献(3)

- [Och 2001] Franz Josef Och, Hermann Ney. What Can Machine Translation Learn from Speech Recognition? In: proceedings of MT 2001 Workshop: Towards a Road Map for MT, pp. 26-31, Santiago de Compostela, Spain, September 2001.
- [Och 2002] Franz Josef Och, Hermann Ney, Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, ACL2002
- [Papineni 1997] K. A. Papineni, S. Roukos, and R. T. Ward. 1997. Feature-based language understanding. In European Conf. on Speech Communication and Technology, pages 1435-1438, Rhodes, Greece, September.
- [Papineni 1998] K. A. Papineni, S. Roukos, and R. T. Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, pages 189-192, Seattle, WA, May.
- [Takeda 1996] Koichi Takeda, Pattern-Based Context-Free Grammars for Machine Translation, Proc. of 34th ACL, pp. 144-- 151, June 1996
- [Wang 1998a] Y. Wang and A. Waibel. Modeling with Structures in Statistical Machine Translation. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Montreal, Canada. August 1998.

## 参考文献(4)

- [Wang 1998b] Ye-Yi Wang, Grammar Inference and Statistical Machine Translation, Ph.D Thesis, Carnegie Mellon University, 1998
- [Wu 1995] Dekai Wu. Stochastic Inversion Transduction Grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora. 14<sup>th</sup> Intl. Joint Conf. On Artificial Intelligence, pp1328-1335, Montreal, Aug, 1995. IJCAI-95
- [Wu 1997] Dekai Wu, Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora, Computational Linguistics Vol.23 No.3 1997.
- [Yamada 2001] K. Yamada and K. Knight, A Syntax-Based Statistical Translation Model, in Proc. of the Conference of the Association for Computational Linguistics (ACL), 2001
- [冯志伟 孙乐 2005] 冯志伟 孙乐 译(D.Jurafsky, J. Martin) 自然语言处理综论,电子工业出版社,2005年.

😊谢谢😊