



# 机器翻译

---

Machine Translation

詹卫东

北京大学中文系

[zwd@pku.edu.cn](mailto:zwd@pku.edu.cn)

<http://ccl.pku.edu.cn/doubtfire/>

# 提纲

---

## □ 1 机器翻译概述

- 分类、基本方法、实现策略

## □ 2 机器翻译技术

- 2.1 基于规则的机器翻译
- 2.2 基于实例的机器翻译
- 2.3 基于统计的机器翻译

## □ 3 机器翻译评测

# 1 机器翻译概述

---

- Machine Translation (MT) 机器翻译

用计算机实现从一种自然语言文本（源语言/source language）到另一种自然语言文本（目标语言/target language）的翻译

## 按需求分类

---

- 传播信息 (dissemination) → 出版/信息发布
- 浏览信息 (assimilation) → 网页翻译
- 交流信息 (interchange) → 实时/多语聊天室
- 查询信息 (information access) → 跨语言信息检索

Note: 对于不同的需求，机器翻译系统的设计应该有针对性，同时对系统的要求也会有所不同

Hutchins, John, 1999, *The Development and use of Machine Translation system and computer-based translation tools*, In Proceeding of International Conference on MT & Computer Language Information Processing, 1999.6.26-28. Beijing

# 按人-机关系分类

---

- Fully Automatic Machine Translation (FAMT) 全自动机译
- Human Assisted Machine Translation (HAMT) 人助机译
- Computer Aided Translation (CAT) 机助人译

# 按实现策略分类

---

- 针对受限语言的机器翻译
- 面向受限领域的机器翻译
- 通用机器翻译系统

Xerox , Boeing 等大公司都使用受限英语（或simplified English）来撰写技术文档，以及进行技术手册的机器翻译

俞士汶，1995，《关于受限的规则汉语的设想》，载王均主编《语言现代化论丛》，山东教育出版社，1995年，pp193-205

张伟，1998，《受限汉语辅助写作系统的构想》，载《计算机世界报》1998年4月13日，第13期D版技术专题

# 按技术方法分类

---

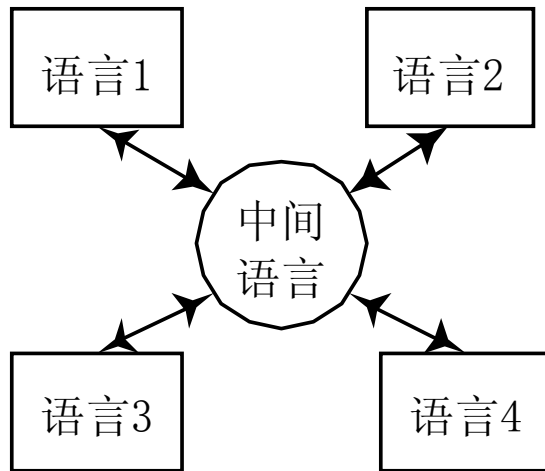
理性主义/基于规则的  
MT方法  
(RBMT)

- 直接翻译法
- 转换法
- 中间语言法

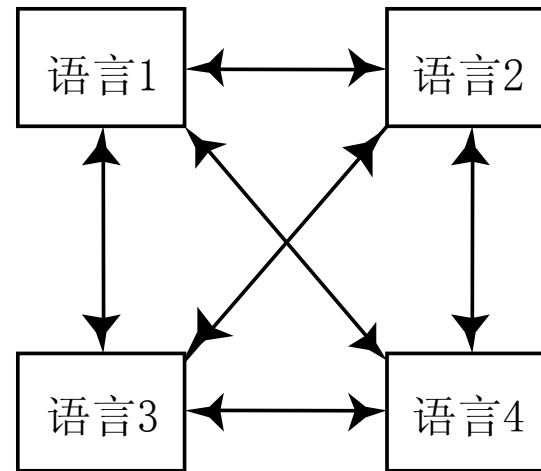
经验主义/基于语料  
库、基于统计的MT  
方法

- EBMT
- Translation Memory
- Pattern-based MT
- Statistical approach to MT

# 中间语言法 ( Interlingua )



(1)

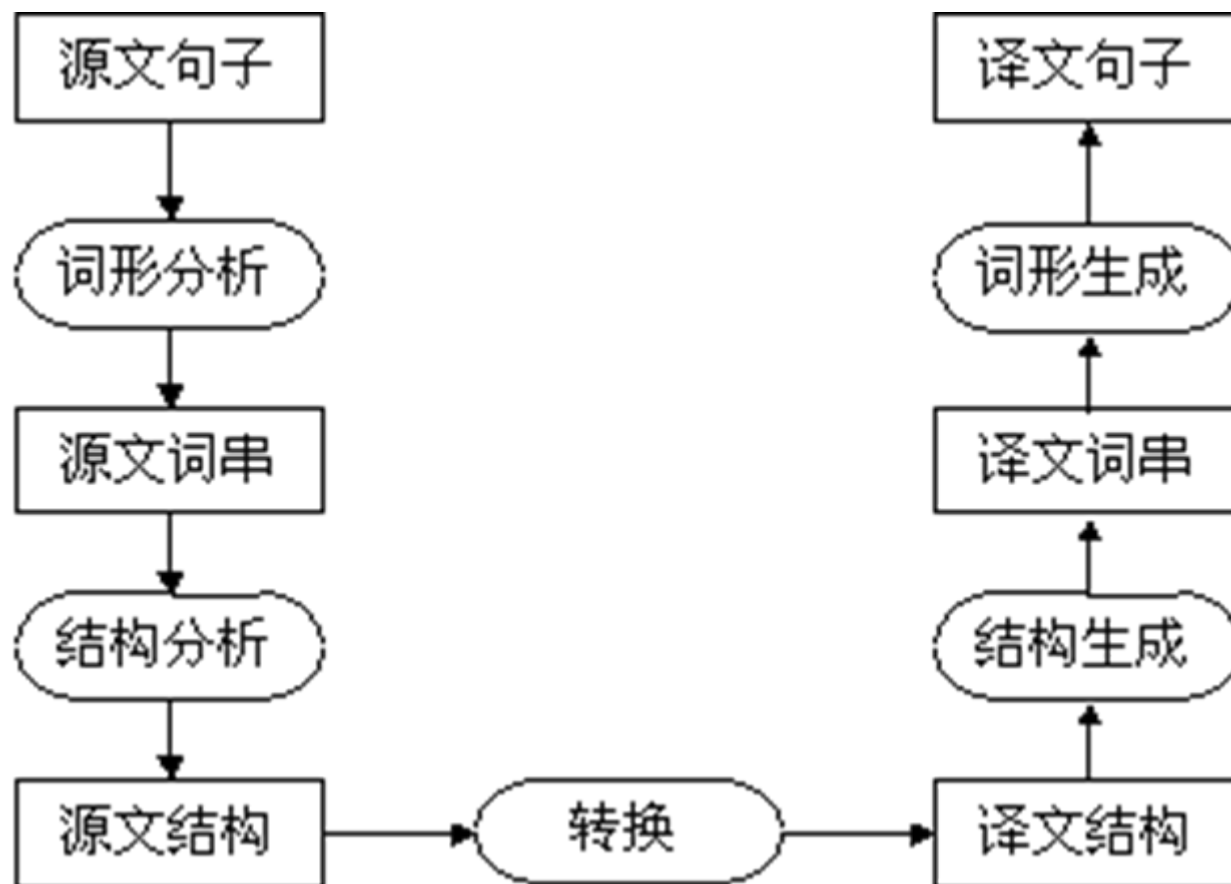


(2)

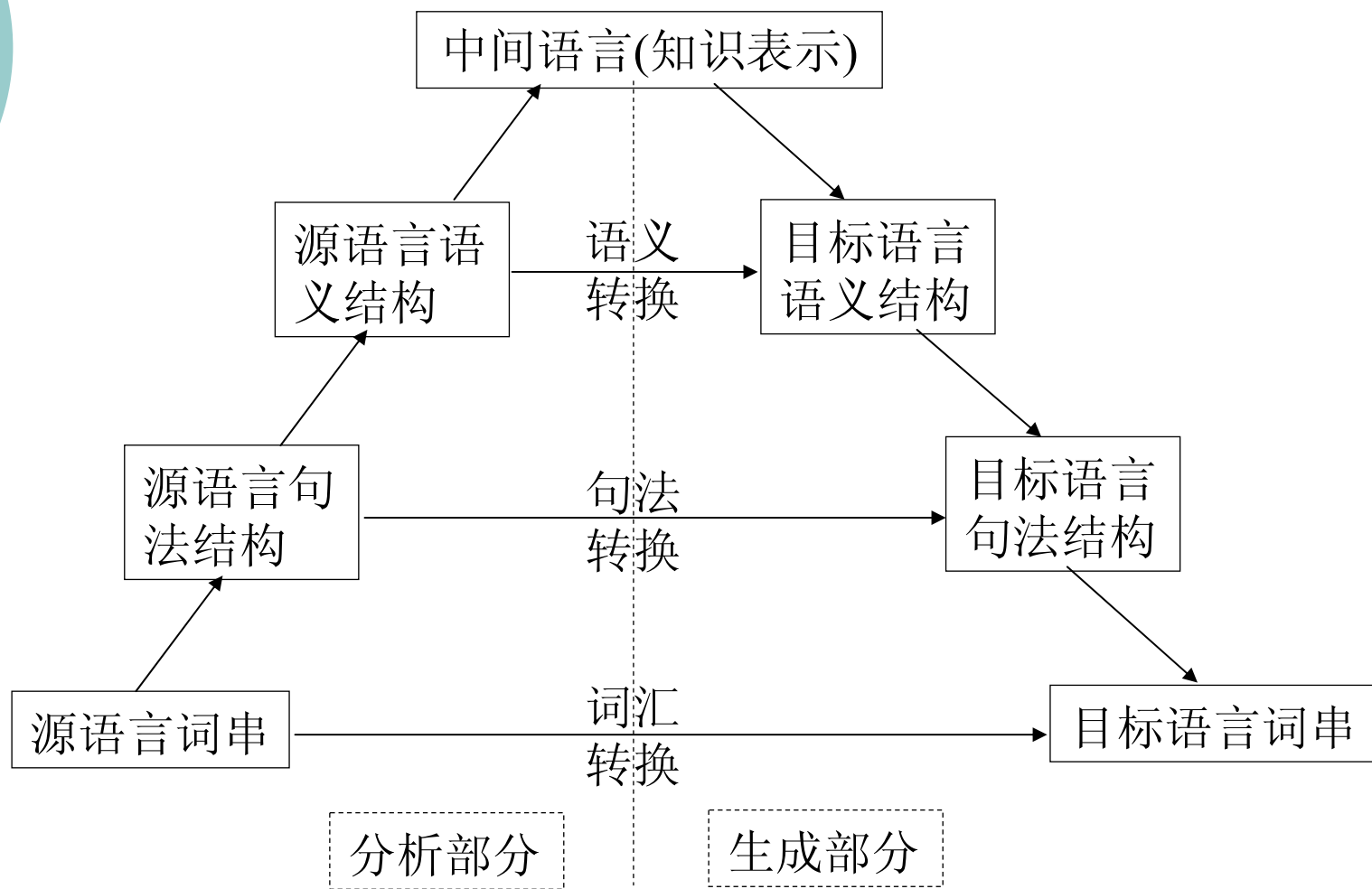
中间语言的例子：世界语，人工定义的语言等

## 转换法 (Transfer)

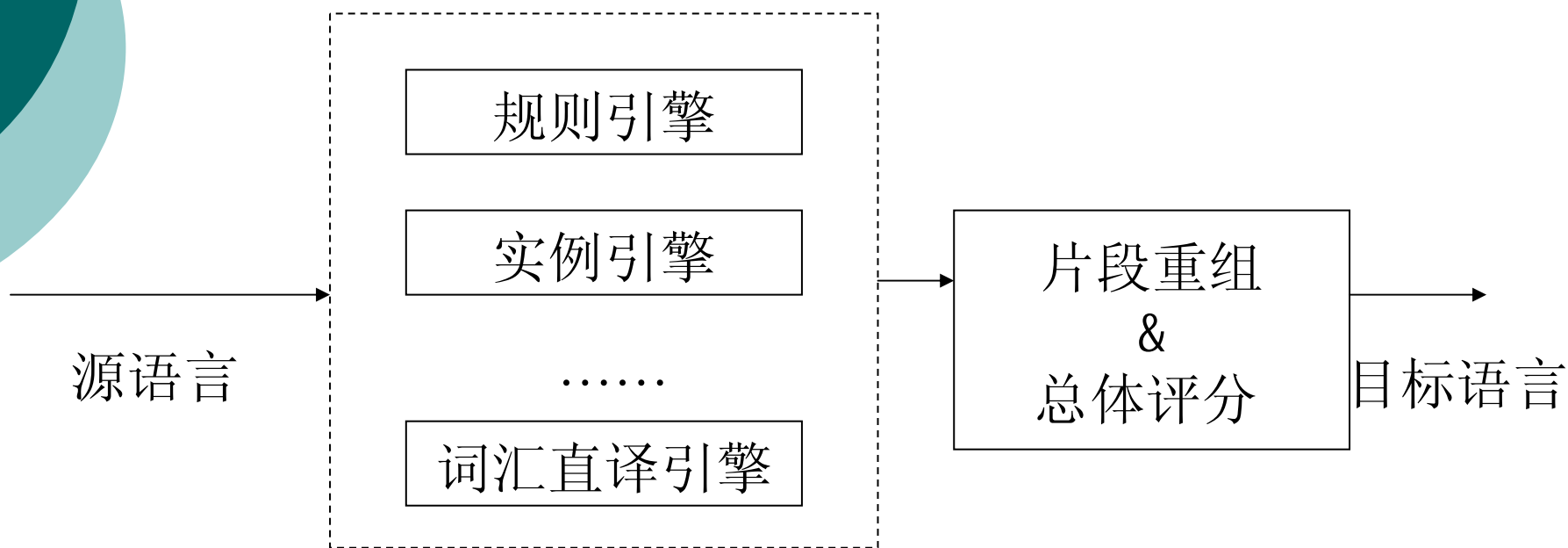
---



# RBMT的一般图示



## 综合型机器翻译系统 (multi-engine MT)



Christopher Hogan, Robert E. Frederking, 1998, *An Evaluation of Multi-engine MT Architecture*, In "Machine Translation and the Information Soup, pages 113-123, Third Conference of the Association for Machine Translation in Americas (AMTA '98), Langhorne, PA. USA, October"

<http://www-2.cs.cmu.edu/afs/cs.cmu.edu/user/chogan/Web/Publications.html>

## 2.1 基于规则的MT

---

她把一束花放在桌上。  $\implies$  She put a bunch of flowers on the table.



Segmenter/POS Tagger



她/r 把/p-q-v-n 一/m-d 束/q 花/n-v-a 放/v 在/p-d-v 桌/n 上/f-v 。 /w



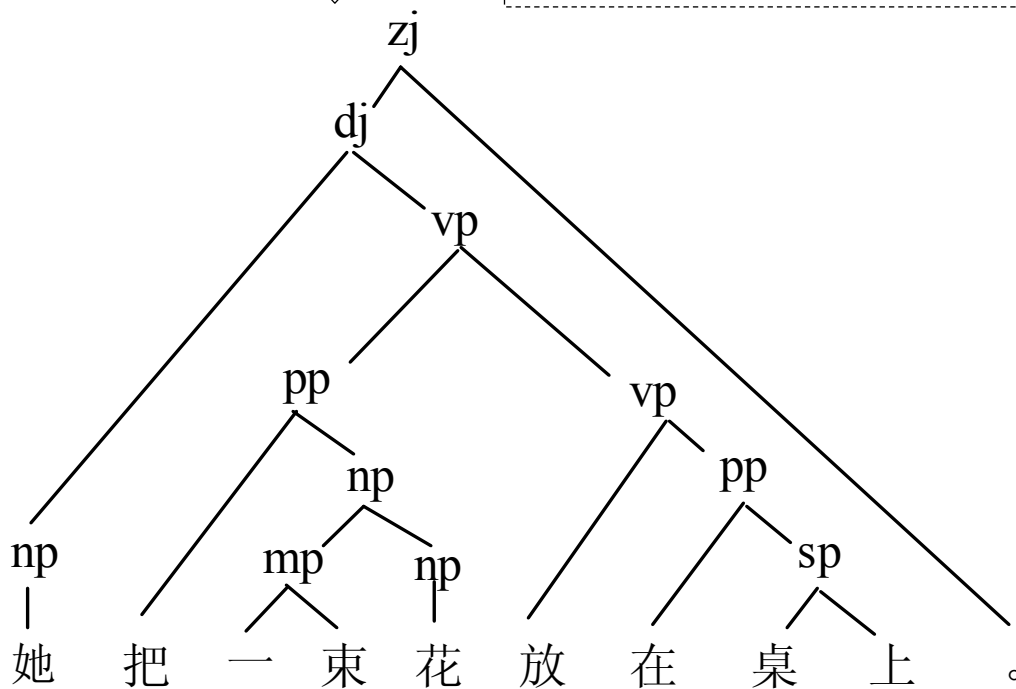
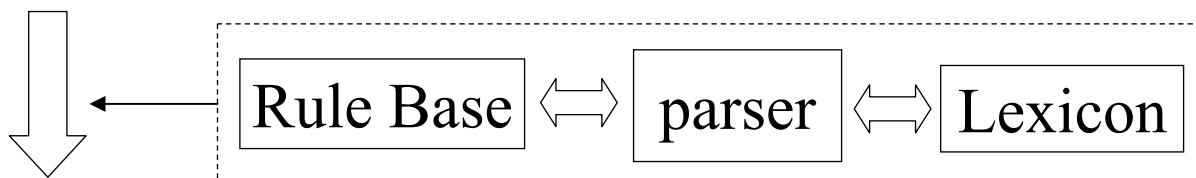
filter



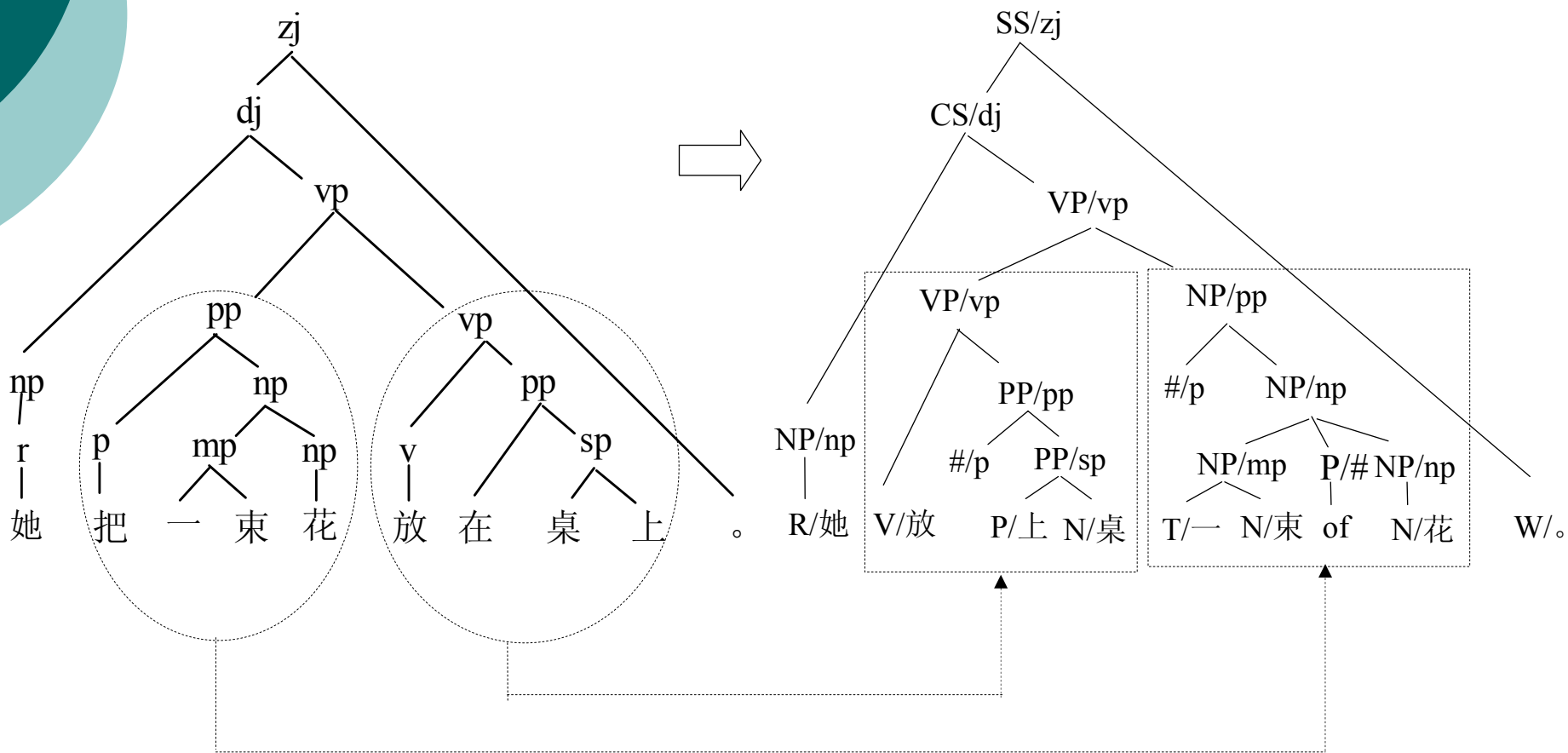
她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。 /w

# 对源语言进行句法分析

她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。 /w

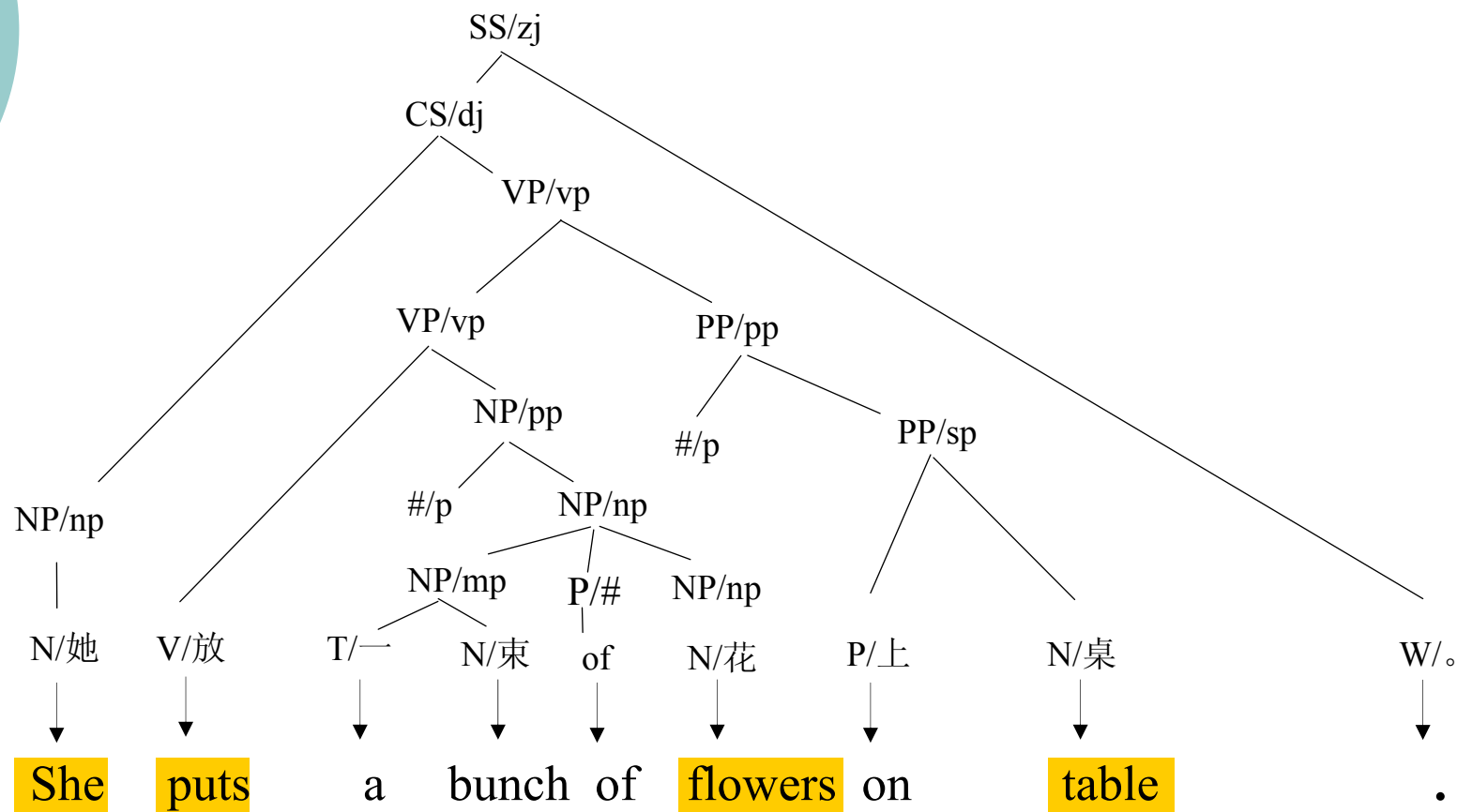


# 从源语言结构树到目标语言结构树





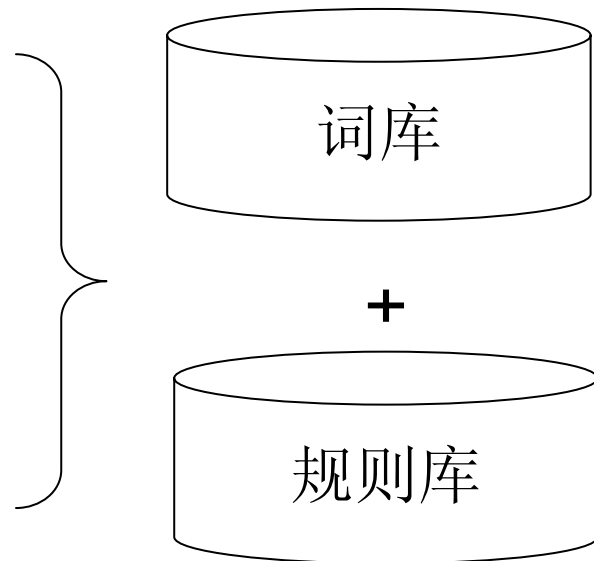
# 对目标语言词语进行变形调整



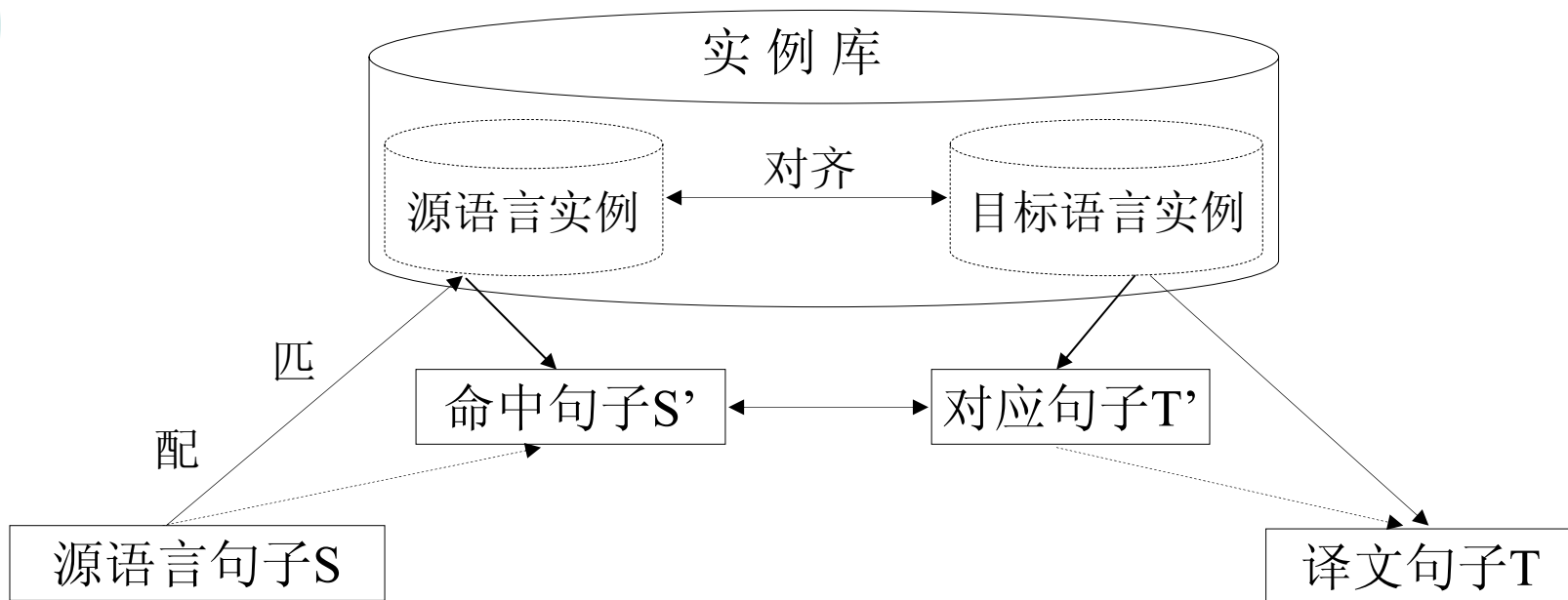
# RBMT系统的知识表示

---

- 1 源语言知识
- 2 目标语言知识
- 3 源语言-目标语言的对译知识
- 4 领域知识
- 5 百科知识



## 2.2 基于实例的MT



*Makoto Nagao (1984)*

## EBMT示例

---

英语实例	汉语实例
He eats vegetable	他吃蔬菜
Acid eats metal	酸腐蚀金属

输入:

I eat potatoes

输出:

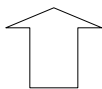
我吃土豆

## EBMT示例（续）

英语实例	汉语实例
<He> <sub>1e</sub> <turned on> <sub>2e</sub> <the radio> <sub>3e</sub>	<他> <sub>1c</sub> <把收音机> <sub>3c</sub> <打开了> <sub>2c</sub>
<The wallet> <sub>4e</sub> <is put> <sub>5e</sub> <on the table> <sub>6e</sub>	<钱包> <sub>4c</sub> <放> <sub>5c</sub> <在桌上> <sub>6c</sub>

输出：

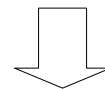
He is put on the table flower



$1e + \text{Replace}(2e, (5e + 6e)) + 3e$

输入：

他把花放在桌上



$1c + 3c + \text{Replace}(2c, (5c + 6c))$

# EBMT需要解决的问题

---

## ○ 如何构建实例库

- 双语对齐：句子级，  
短语级（亚句子级），  
词汇级

## ○ 如何查找翻译实例

- 匹配精度：句子级匹配 } 相似度计算技术  
亚句子级匹配 }

## ○ 如何生成好的译文

## 词汇对齐 (word alignment)

---

- 互为翻译的一对单词要比相互不为翻译的一对单词更有可能出现在同一个对齐的句子对中
- 假设-检验方法
  - 假设阶段  
生成所有候选的对译词对 $\ell$  (translation equivalent)
  - 检验阶段  
根据统计关联度量选择出统计意义上较为可靠的对译词对 $\ell$

Gale, 1991

# 联立表(contingency table)

---

	t	$\neg t$
s	a	b
$\neg s$	c	d

$\langle s, t \rangle$  是候选对译词对<sub>儿</sub>

a : 语料中同时出现s和t的句对<sub>儿</sub>数

b : 语料中出现s不出现t的句对<sub>儿</sub>数

c : 语料中不出现s出现t的句对<sub>儿</sub>数

d : 语料中s,t同时不出现的句对<sub>儿</sub>数

语料规模  $n = a + b + c + d$

# 词汇对齐可能性的度量方法

---

$$(1) \quad MI(st, tt) = \log_2 \frac{n \times a}{(a+b) \times (a+c)}$$

$$(2) \quad DICE(st, tt) = \frac{2a}{(a+b) \times (a+c)}$$

$$(3) \quad LL(st, tt) = 2 \times \left( a \times \log \frac{a \times n}{(a+b) \times (a+c)} + b \times \log \frac{b \times n}{(a+b) \times (b+d)} \right. \\ \left. + c \times \log \frac{c \times n}{(c+d) \times (a+c)} + d \times \log \frac{d \times n}{(c+d) \times (b+d)} \right)$$

$$(4) \quad \chi^2(st, tt) = \frac{n \times (a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)}$$

## 词汇对齐示例（chi-square方法）

	房子	¬房子		家庭	¬家庭
house	619	120	house	174	980
¬house	93	19168	¬house	41	18105

**n = 20,000**

$$\begin{aligned}\chi^2(\text{house}, \text{房子}) &= \frac{20000 \times (619 \times 19168 - 120 \times 93)^2}{(619 + 120) \times (619 + 93) \times (120 + 19168) \times (93 + 19168)} \\ &= 644847.17\end{aligned}$$

$$\begin{aligned}\chi^2(\text{house}, \text{家庭}) &= \frac{20000 \times (174 \times 18105 - 980 \times 41)^2}{(174 + 980) \times (174 + 741) \times (980 + 18105) \times (41 + 18105)} \\ &= 13879.98\end{aligned}$$

$$\chi^2(\text{house}, \text{房子}) \succ \chi^2(\text{house}, \text{家庭})$$

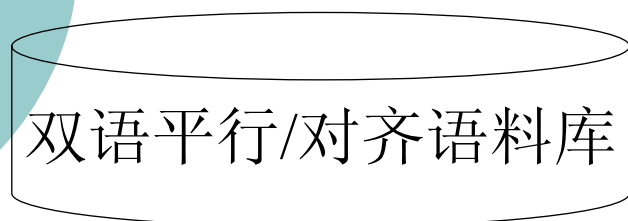
# Multi-word translation equivalents

---

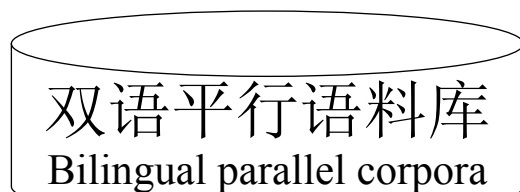
Chinese	English	$\chi^2$ -score
成人_图书馆	adult_library	68620.5
影子_董事	shadow_director	68469.8
幕_墙	curtain_wall	68469.8
卤味_店	lo_mei	68282.1
橡胶_手套	rubber_glove	68041.9
橡胶_围裙	rubber_apron	67723.5
疾病_津贴	sickness_allowance	67433.1
计算机_软件	computer_software	67281.6
软_雪糕	ice_cream	67281.6
污水_隧道	sewage_tunnel	66626.8
工程_原理	<i>engineering_principle</i>	66626.8

# Word-type / Word-token alignment

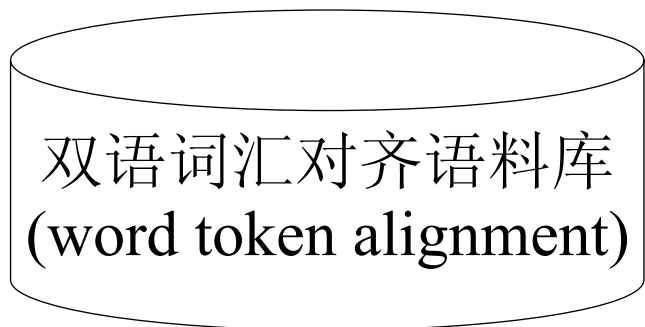
---



extraction



alignment



# EBMT在两个方向上的发展

---

- 基于模板的机器翻译系统 Pattern-based MT

汉语模板: 这个 [np1] 比 [np2] 更 [ap]

英语模板: The [NP1/np1] be more [AP/ap] than [NP2/np]

- 翻译记忆 Translation Memory  
翻译重用/术语翻译

<http://www.trados.com/>

# 模板的获取方法

---

- 利用一对实例进行泛化

Jaime G. Carbonell & Ralf D. Brown, Generalized Example-Based Machine Translation, a four-year project(1997-2001)  
<http://www.cs.cmu.edu/~ralf/ebmt.html>

- 利用两对实例进行比较

H. Altay Guvenir, Ilyas Cicekli(1998), Learning Translation Templates from Examples, Information Systems, vol.23, No.6, pp.353-363

# 利用一对实例进行泛化

---

Karl Marx was born in Trier, Germany on May 5, 1818.

卡尔·马克思于1818年5月5日出生在德国特里尔城。



[Person] was born in [City] in [Date]

[Person] 于 [Date] 出生在 [City]

“实体”泛化

# 利用两对实例进行比较

我给玛丽一支笔  
我给汤姆一本书

I gave Mary a pen.  
I gave Tom a book.



我给  $[np]_1$  一  $[q]$   $[np]_2$

I gave  $[NP]_1$  a  $[NP]_2$



$np_1 \leftrightarrow NP_1$   
 $np_2 \leftrightarrow NP_2$   
 $q \leftrightarrow \Phi$

$np_1$ .语义类 = 人  
 $np_2$ .语义类 = 物品

存“同”变“异”

# 翻译记忆的基本思想

---

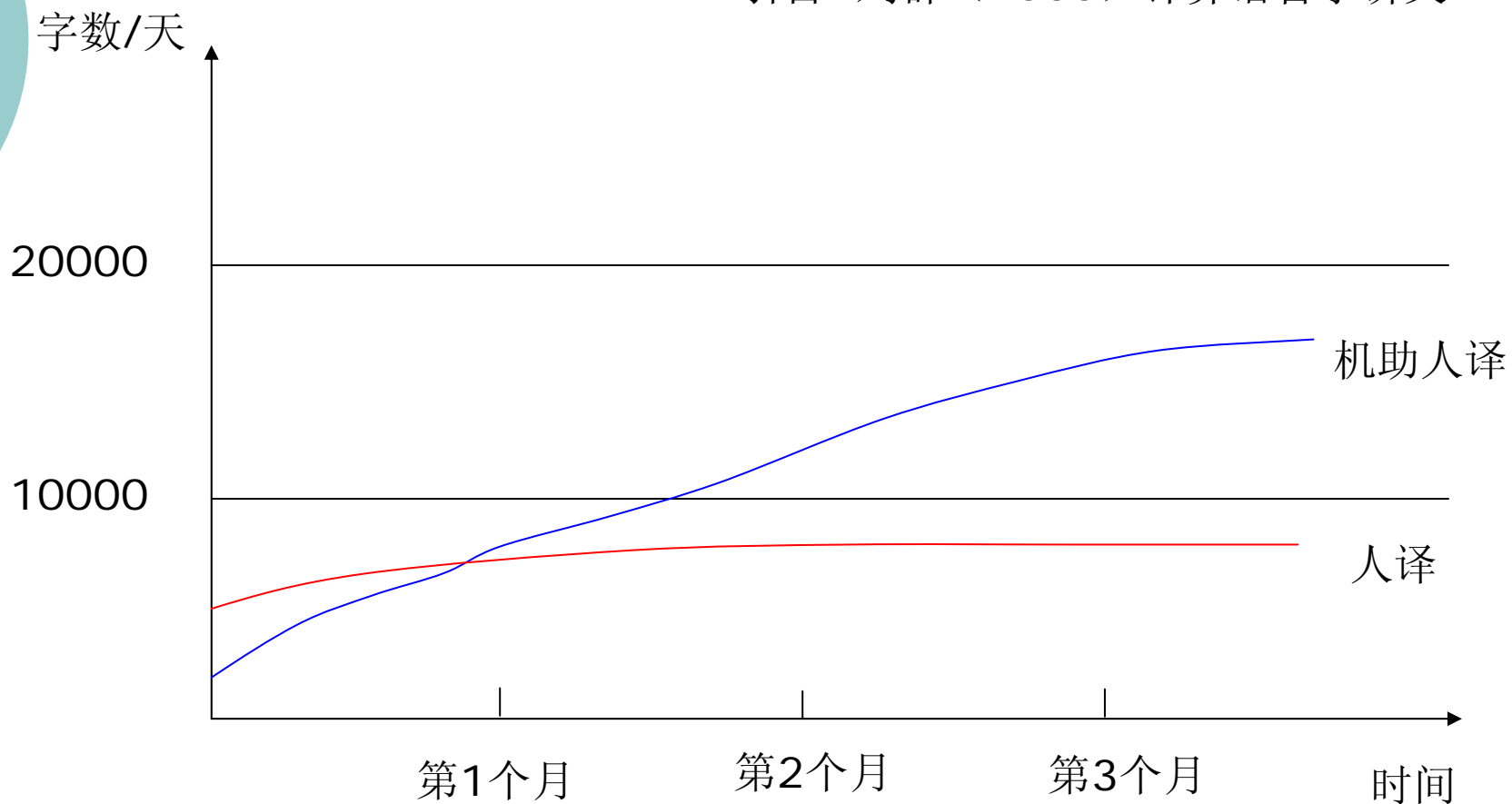
- 存储翻译历史：  
把已经翻译过的句子保存起来
- 翻译一个新句子时，直接到Translation Memory库中去查找
  - 如果发现相同的句子，直接输出译文
  - 否则交给人去翻译，同时提供相似句子的参考译文

TM涉及到的主要技术问题

多种文件格式的分解与合成  
术语库管理功能  
语料库的句子对齐（历史资料的重复利用）  
翻译任务的分解与合并  
翻译工作量的估计  
数据共享和数据交换

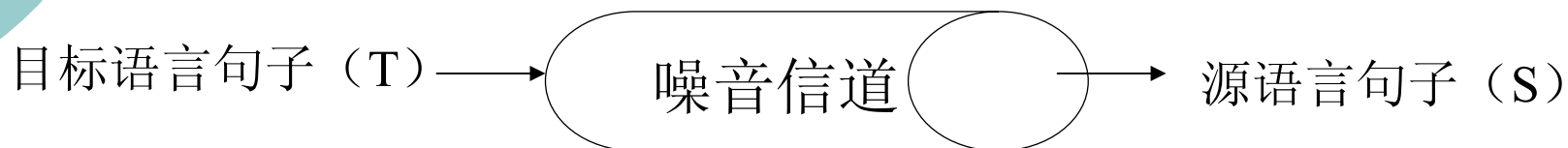
# TM 可提高翻译效率

引自 刘群（2005）计算语言学讲义



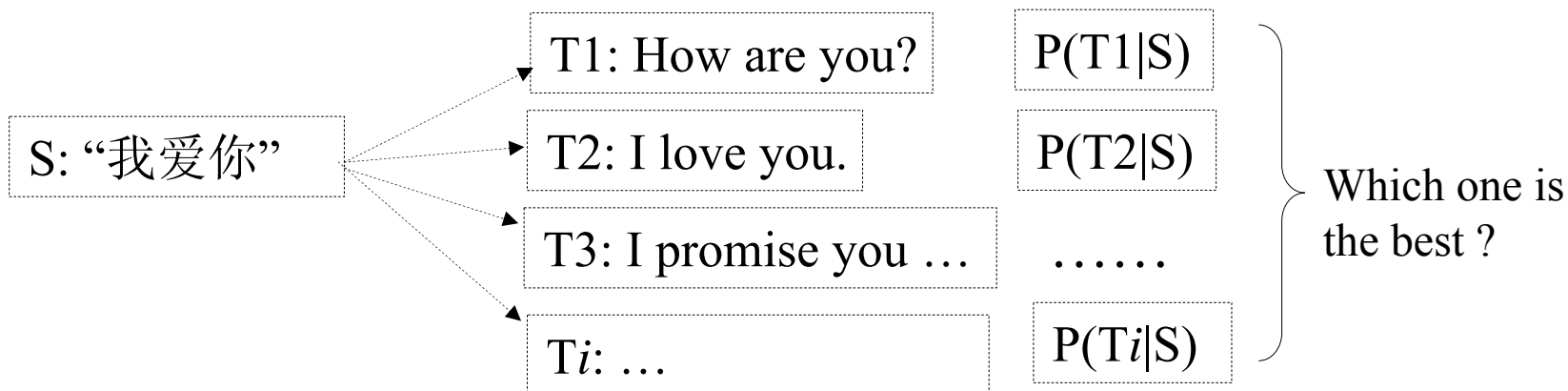
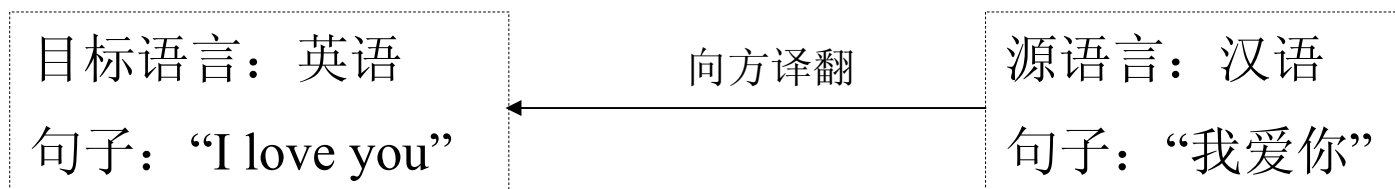
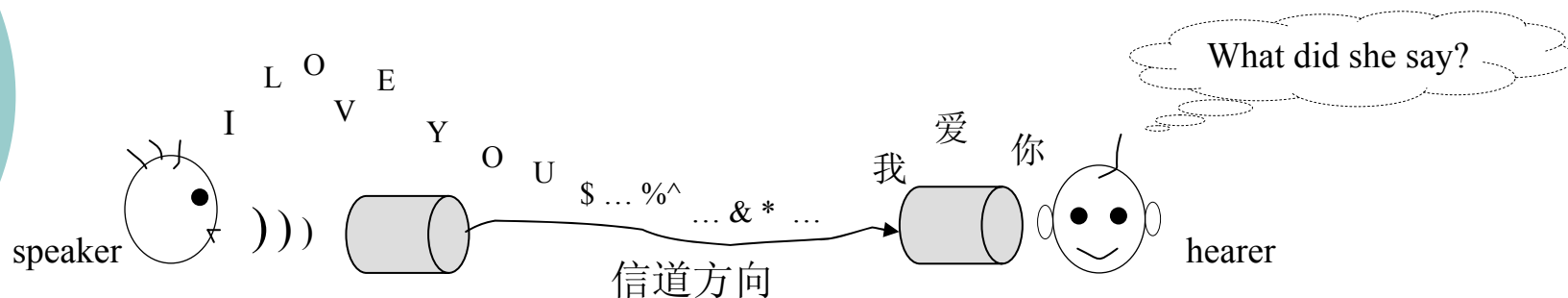
## 2.3 基于统计的MT

- 刘群. 统计机器翻译综述 《中文信息学报》 2003年第4期, 1-12.
- IBM公司Brown et al. (1990), (1993)

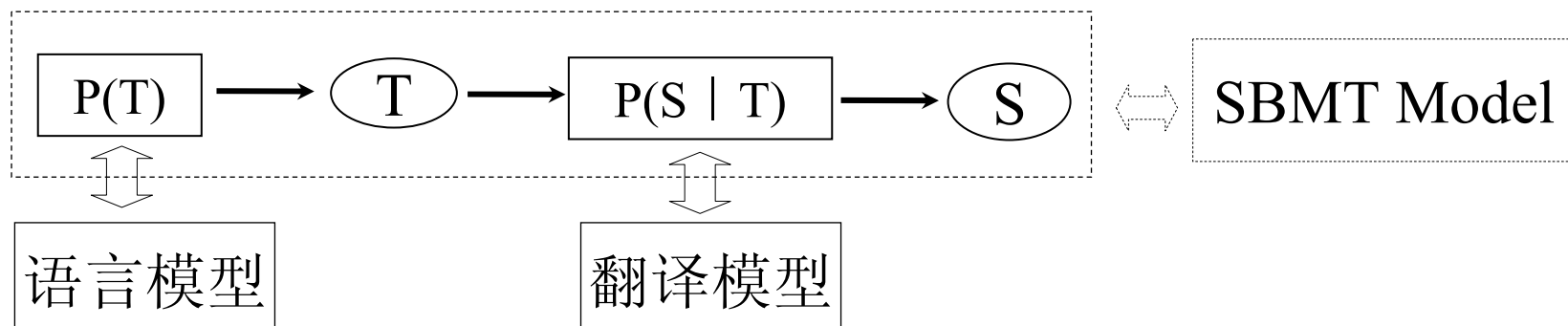


$$\hat{T} = \operatorname{argmax}_T P(T | S)$$
$$P(T | S) = \frac{P(T)P(S | T)}{P(S)}$$
$$\left. \begin{array}{l} \hat{T} = \operatorname{argmax}_T P(T | S) \\ P(T | S) = \frac{P(T)P(S | T)}{P(S)} \end{array} \right\} \hat{T} = \operatorname{argmax}_T P(T)P(S | T)$$

# 汉英翻译的噪音信道模型



# 统计机器翻译模型面对的问题



解决三个问题：

- 1) 语言模型 $P(T)$ 的参数估计
  - 2) 翻译模型 $P(S|T)$ 的参数估计
  - 3) 译文快速搜索（如何快速找到  $\hat{T}$ ）
- 建模问题
- 解码问题

# 语言模型P(T)的参数估计

---

N-gram

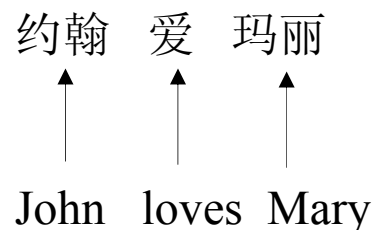
$$\begin{aligned} P(T) &= P(T_1 T_2 \dots T_n) = P(T_1) P(T_2 | T_1) \dots P(T_n | T_1 T_2 \dots T_{n-1}) \\ &\approx P(T_1) P(T_2 | T_1) \dots P(T_n | T_{n-1}) \end{aligned}$$

$$P(T_n | T_{n-1}) = \frac{\text{Number}(T_{n-1}, T_n)}{\text{Number}(T_{n-1})}$$

# 翻译模型P(S|T)的参数估计

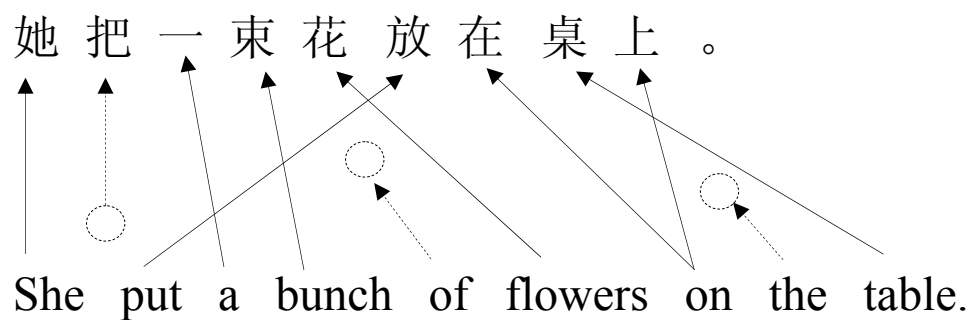
○ 考虑翻译的三种可能性

• 直译 (direct translation)



• 繁殖 (fertility)

• 变形 (distortion)



变形

繁殖

## 翻译模型P(S|T)的参数估计（续）

$$S = s_1 s_2 \dots s_m$$

$$T = t_1 t_2 \dots t_n$$

$$P(S | T) = P(s_1 s_2 \dots s_m | t_1 t_2 \dots t_n)$$

$$P(S | T) \approx \prod_{i=1}^n \left[ P(f_i | t_i) \times \prod_{j=1}^{f_i} P(s_j | t_i) \right] \times \prod_{j=1}^m P(j | i, m)$$

繁殖概率，即一个目标语单词 ( $t_i$ ) 翻译成  $f_i$  个源语言单词的概率

直译概率，即一个目标语单词 ( $t_i$ ) 翻译成一个或若干个源语言单词 ( $s_j$ ) 的概率

变形概率，即第  $i$  个目标语单词 ( $t_i$ ) 对译为第  $j$  个源语言单词 ( $s_j$ ) 的概率

对于长度为9的汉语句子，英  
语句子里第7个词翻译为汉语  
句子中第9个词的概率

## P(S|T)计算示例

源语言词语位置  
序号：对译目标  
语词语位置序号

源语言

1:1 2:0 3:3 4:4 5:6 6:2 7:7 8:9 9:7  
她把一束花放在桌上。

目标语言

She put a bunch of flowers on the table.  
1 2 3 4 5 6 7 8 9

“a”对译1个汉语词的概率

$P(S | T) =$

$[[P(1 | she)P(她 | she)][P(1 | put)P(放 | put)][P(1 | a)P(一 | a)]$

$[P(1 | bunch)P(束 | bunch)][P(0 | of)][P(1 | flowers)P(花 | flowers)]$

$[P(2 | on)P(在 | on)P(上 | on)][P(0 | the)][P(1 | table)P(桌 | table)]$

$[P(1 | 1,9)P(2 | 0,9)P(3 | 3,9)P(4 | 4,9)P(5 | 6,9)$

$P(6 | 2,9)P(7 | 7,9)P(8 | 9,9)P(9 | 7,9)]$

“table”翻译为“桌子”的概率

## 3 机器翻译评测

---

- 人工评测

忠实度 (Fidelity) + 可懂度 (Intelligibility)

- 自动评测

- 基于测试点的自动评价
- 基于编辑距离的自动评价
- 基于N-gram的自动评价

<http://www.863data.org.cn>

<http://www.nist.gov/speech/tests/mt/>

# 人工评测 vs. 自动评测

---

	人工评测	自动评测
成本	高	低
准确性	高	低
重用性	低	高

## 人工评测

---

评分	忠实度	可懂度
0	完全没有译出来	完全不可理解
1	译文只有个别词符合原文	译文晦涩难懂
2	译文有少数内容符合原文	译文很不流畅
3	译文基本表达了原文的意思	译文基本流畅
4	译文表达了原文的绝大部分信息	译文流畅，但是在地道性方面有所不足
5	译文准确完整地表达了原文信息	译文是流畅而且地道的句子

这两个评价标准源自ALPAC报告（1966）

# 自动评测 – 基于测试点

---

## 基本思想:

- 不针对整句进行质量评价，而是设置测试点，简化测试目标（模拟人类标准化考试的办法）
- 采用测试描述语言（TDL）来描述一个句子中的测试点。使测试可以全自动完成
- 构建大规模测试集，通过汇集大量的孤立测试点的测试，能尽可能翻译整句翻译的质量

俞士汶 等（1992） 基于测试集与测试点的机译系统评估，载陈肇雄（1992） 主编《机器翻译研究进展》电子工业出版社，pp.524-537。

# 自动评测 – 基于测试点（举例）

原文： They got up at eleven this morning.

测试点TDL描述

测试开始 → R → (741:2) \* 上午 &(eleven) \$A \*

测试科目 → R → (741:1) \* 上午的 &(eleven) \$A \*

得分 → R → (741:1) \* 早晨 &(eleven) \$A \*

\$A → 点/时/点钟

测试结束 → ##

\* 通配符

&(eleven) 表示查词典中的译文

\$A 是“非终结符”

“上午/点/时/...”是终结符

译文1: ...上午十一点 ... score: 2

译文2: ...上午的11点钟 ... score: 1

译文3: ...早晨十一时 ... score: 1

# 自动评测 – 基于编辑距离

---

编辑前：机器译文

编辑后：参考译文

编辑操作：插入、删除、替换

“编辑距离”定义为  
“编辑操作的次数”

原文：She is a star with the theatre company.

机器译文：她是与剧院公司的一颗星。

参考译文：她是剧团的明星。

编辑距离：6

插入次数（4次）：与 公司 一 颗

替换次数（2次）：剧团 → 剧院 明星 → 星

# 自动评测 – 基于N-gram

---

## 基本思想

- 用机器译文中出现的N元组和参考译文中出现的N元组相比，计算匹配的N元组个数与机器译文的N元组总个数的比例。比例越高，表示机器译文质量越好。
- 允许一个源文有多个参考译文，进行综合评分。

IBM (2001) 技术报告: BLEU 评测方法

# Bi-Lingual Evaluation Understudy

---

$$BLEU = BP \cdot e^{\left( \sum_{n=1}^N \frac{1}{N} \log p_n \right)}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

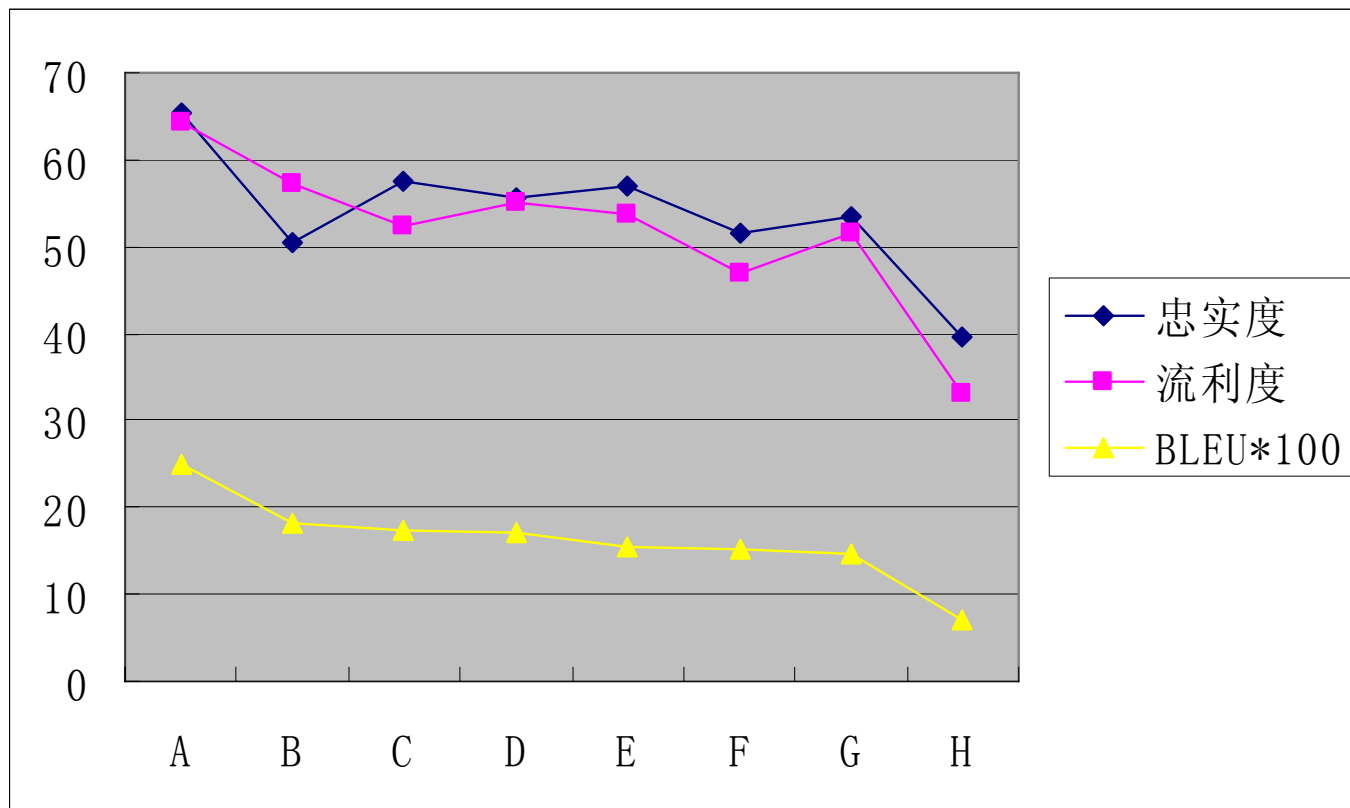
$P_n$  机器译文与参考译文的n元组重合的个数  
占机器译文中n元组总数的比例；

$N$  为最大的n元语法阶数（ $N$ 一般取4）  $1 \leq n \leq N$ ；

$c$  为机器译文中单词的个数；

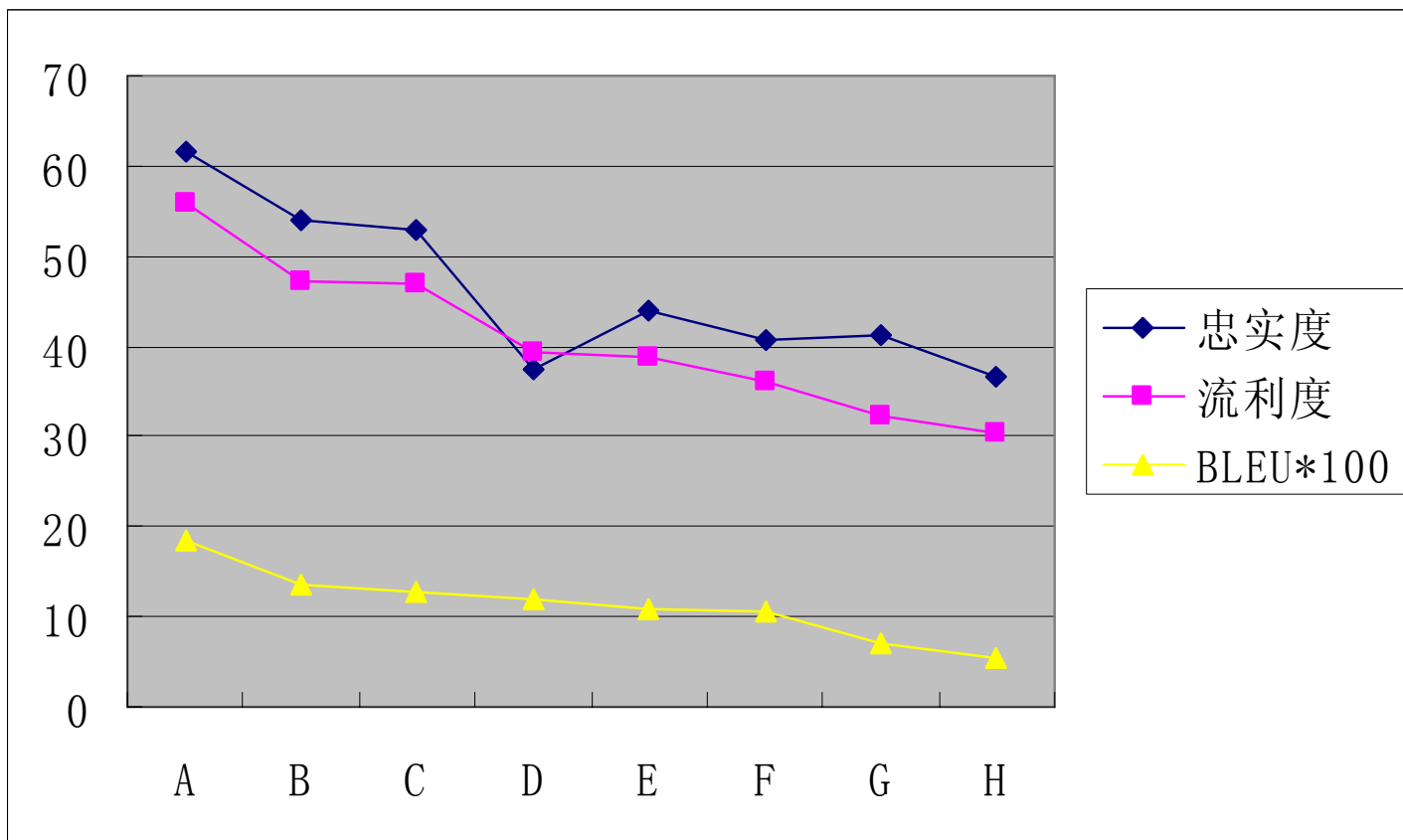
$r$  为参考译文中与 $c$ 最接近的译文单词个数；

# 2005年度863机器翻译评测



汉英机器翻译 – 对话语料 (A-H八个系统参评)

# 2005年度863机器翻译评测



汉英机器翻译 – 篇章语料 (A-H八个系统参评)

## 进一步阅读文献

---

- 冯志伟 (2004) 《机器翻译研究》，中国对外翻译出版公司2004年版。
- 冯志伟 (1995) 《自然语言机器翻译新论》，语文出版社1995年版。
- 赵铁军 (2000) 《机器翻译原理》，第10章，哈尔滨工业大学出版社。
- 翁富良、王野翊 (1998) 《计算语言学导论》，第8章，中国社会科学出版社。
- 刘群、俞士汶 (1998)，《汉英机器翻译的难点分析》，载黄昌宁主编《1998中文信息处理国际会议论文集》，清华大学出版社。pp507-514
- M.Nagao, 1984, *A framework of a mechanical translation between Japanese and English by Analogy Principle*, In *Artificial and Human Intelligence*, A.Elithorn et al eds., NATO Publication.
- W. Gale, 1991, *Identifying words correspondences in parallel Texts*, DARPA speech and Natural language workshop. Asilomar, CA., 1991
- P. Brown et al, 1990, *A statistical approach to machine translation*, Computational linguistics, Vol.16, No. 2, 1990
- P. Brown et al, 1993, *The mathematics of statistical machine translation: parameter estimation*, Computational Linguistics, Vol. 19, No.2, 1993.
- Yu Shiwen, 1993, *Automatic evaluation of output quality for machine translation systems*, Machine Translation, Vol. 8, Kluwer Academic Publisher, pp117-126.
- Papineni, Roukos, Ward, Zhu (2001). *Bleu: a Method for Automatic Evaluation of Machine Translation*, (IBM Technical Report, Keyword. RC22176- W0109-022)

## 附录1：机器翻译发展小史

---

1946 – 1954	第1个MT系统在美国Georgetown大学问世，6条规则，250个词，俄语 → 英语 (50个句子/化学文本)
1966	ALPAC报告，MT陷入低谷
1970 – 1980	反思，计算语言学理论的发展，人工智能的发展
1980 – 1990	基于规则的系统日益成熟； 与此同时，人们开始探索更多其他的MT方法
1990 –	MT应用需求呈上升趋势，技术日益靠拢实用目标，与语音技术、互联网应用的结合趋势日渐明显

赵铁军（2000）第1章；冯志伟（1995）第1—3章

## 附录2：机器翻译系统

---

Systran公司 <http://www.systransoft.com/>

Google [http://www.google.com/language\\_tools?hl=zh-CN](http://www.google.com/language_tools?hl=zh-CN)

AltaVista <http://babelfish.altavista.com/>

SDL国际 <http://www.freetranslation.com/>

WorldLingo <http://www.worldlingo.com/>

中软译星（汉英 双向 简繁） <http://www.transtar.com.cn/gb/default.asp>

金桥译港（汉英 双向） <http://yg.gb.com.cn/index.htm>

看世界（英汉 简繁） <http://www.readworld.com/tran/index.html>

华建MT系统 <http://www.hjtek.com/newnew/platform/second/demodown.htm>