自然语言理解

第十章　统计机器翻译

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

No.95, Zhongguancun East Road
Beijing 100080, China

http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

第十章 统计机器翻译

10.1 机器翻译的产生与发展

❑ **概念：机器翻译** (machine translation) **是使用计算机把一种语言 (源语言，** source language) **翻译成另一种语言 (目标语言，** target language) **的一门学科。**

机器翻译是一门学科也是一门技术。

10.1 机器翻译的产生与发展

- **起始与发展**

1）古希腊

2）17世纪：笛卡儿(Descartes)莱布尼兹(Leibniz)试图用统一的数字代码编写词典；17世纪中页贝克(Cave Beck)等人出版类似的词典。

3) 1930s：亚美尼亚法国工程师阿尔楚尼(G. B. Arsouni)提出了用机器来进行语言翻译的想法，并在1933年7月22日获得了一项“翻译机”的专利，叫做机器脑 (mechanical brain)

1933年，前苏联发明家特洛扬斯基设计了用机械方法把一种语言翻译成为另一种语言的机器。

10.1 机器翻译的产生与发展

4) 1946年英国工程师布斯 (A. D. Booth)和美国洛克菲勒基金会副总裁韦弗 (W. Weaver) 提出了用计算机进行语言翻译的想法。

5) 1947年，韦弗发表了以《翻译》为题目的备忘录，正式提出机器翻译问题：

a) 翻译类似于解读密码的过程：“当我阅读一篇用俄语写的文章时，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，当我阅读时，我是在进行解码。”

10.1 机器翻译的产生与发展

b) **原文和译文“说的是同样的事情”因此，当把语言A翻译为语言B时，就意味着，从语言A出发，经过某一“通用语言** (universal language)”**或“中间语言** (interlingual language)”**，然后转换为语言B，这种“通用语言”或“中间语言”可以假定是全人类共同的。**

1954**年，美国乔治敦大学在国际商用机器公司**(IBM)**的协助下，用**IBM-701**计算机，进行了世界上第一次机器翻译实验，把几个简单的俄语句子翻译成英语。**

苏联、英国、日本也进行了MT**实验，从此**MT**出现热潮。**

10.1 机器翻译的产生与发展

1964年，美国科学院成立语言自动处理咨询委员会(Automatic Language Processing Advisory Committee, ALPAC)，调查机器翻译的研究情况，并于1966年11月公布了一个题为“语言与机器”的报告，简称 ALPAC报告，宣称：

“在目前给机器翻译以大力支持还没有多少理由”

“机器翻译遇到了难以克服的语义障碍(semantic barrier)”

1954 ～ 1970 (ALPAC)：草创时期；

1970 ～ 1976：复苏阶段；

1976 ～ 现在：繁荣时期。

10.2 机器翻译基本方法

1) 初期：人们一般采用直接翻译的方法，从源语言句子的表层出发，将单词或者词组、短语甚至句子直接置换成目标语言译文，必要时进行一些词序的调整，便可以生成译文的句子。对原文句子的分析仅仅满足于特定译文生成的需要。这类翻译系统一般真对某一个特定的语言对，将分析与生成、语言数据、文法和规则与程序等都融合在一起。

10.2 机器翻译基本方法

2) 1957年美国学者V. Yingve在《句法翻译框架》(Framework for Syntactic Translation)一文中提出了对源语言和目标语言均进行适当描述、把翻译机制与语法分开、用规则描述语法的实现思想，这就是基于规则的(Rule-based)转换翻译方法。其翻译过程分成六个步骤：

a) 对源语言句子进行词法分析

b) 对源语言句子进行句法分析

c) 原语言到译文词汇转换

d) 源语言到译文结构转换

e) 译文句法生成

c) 译文词法生成

10.2 机器翻译基本方法

“独立分析－独立生成－相关转换”的翻译方法

其代表系统是法国格勒诺布尔（Grenoble）医科大学（现为格勒诺布尔大学）信息、数学与应用研究院（IMAG）的机器翻译研究所（GETA）开发的ARIANE翻译系统。

1976年加拿大蒙特利尔大学与加拿大联邦翻译局联合开发的实用性机器翻译系统 TAU-METEO：天气预报信息服务。

10.2 机器翻译基本方法

基于规则转换的翻译方法的评价：

主要优点：可以较好地保持原文的结构，产生的译文结构与源文的结构关系密切，尤其对于语言现象已知的或句法结构规范的源语言语句具有较强的处理能力和较好的翻译效果。

主要不足：分析规则都要由人工编写，工作量大，规则的主观性强，规则的一致性难以保障，不利于系统扩充，尤其对非规范的语言现象缺乏相应的处理能力 。

10.2 机器翻译基本方法

3）基于中间语言的翻译方法（Interlingua-based）

- **方法：输入语句→ 中间语言→ 翻译结果**
- **代表系统：**JANUS (CMU) **早期版本**

☆ **源语言解析器**

☆ **比较准确的中间语义描述语言**(Interlingua)

☆ **目标语言生成器**(Target Language Generator)

10.2 机器翻译基本方法

对基于中间语言的翻译方法评价：

主要优点：中间语言的设计可以不考虑具体的翻译语言对，因此，该方法尤其适合多语言之间的互译。

主要弱点：如何定义和设计中间语言的表达方式并不是一件容易的事情，中间语言在语义表达的准确性、完整性等很多方面都面临若干困难。

－国际先进语音翻译研究联盟(C-STAR)**定义的中间转换格式**(Interchange Format, IF)

－日本东京联合国大学 (United Nations University) **提出的通用网络语言**(Universal Networking Language, UNL)

10.2 机器翻译基本方法

4) 基于语料库的翻译方法

(a) 基于事例的翻译方法（Example-based）

- 方法：输入语句→ 与事例相似度比较→ 翻译结果
- 代表系统：ATR-MATRIX（ATR, Japan)
- 事例库：源语言语句或成份S1 ==> 目标语言表达T1

  … …

  源语言语句或成份Sn ==> 目标语言表达Tn

输入语句S’ :: { S1, S2, … … Sn } ==> 翻译结果T’

10.2 机器翻译基本方法

(b) **基于记忆的翻译方法**（memory-based MT）

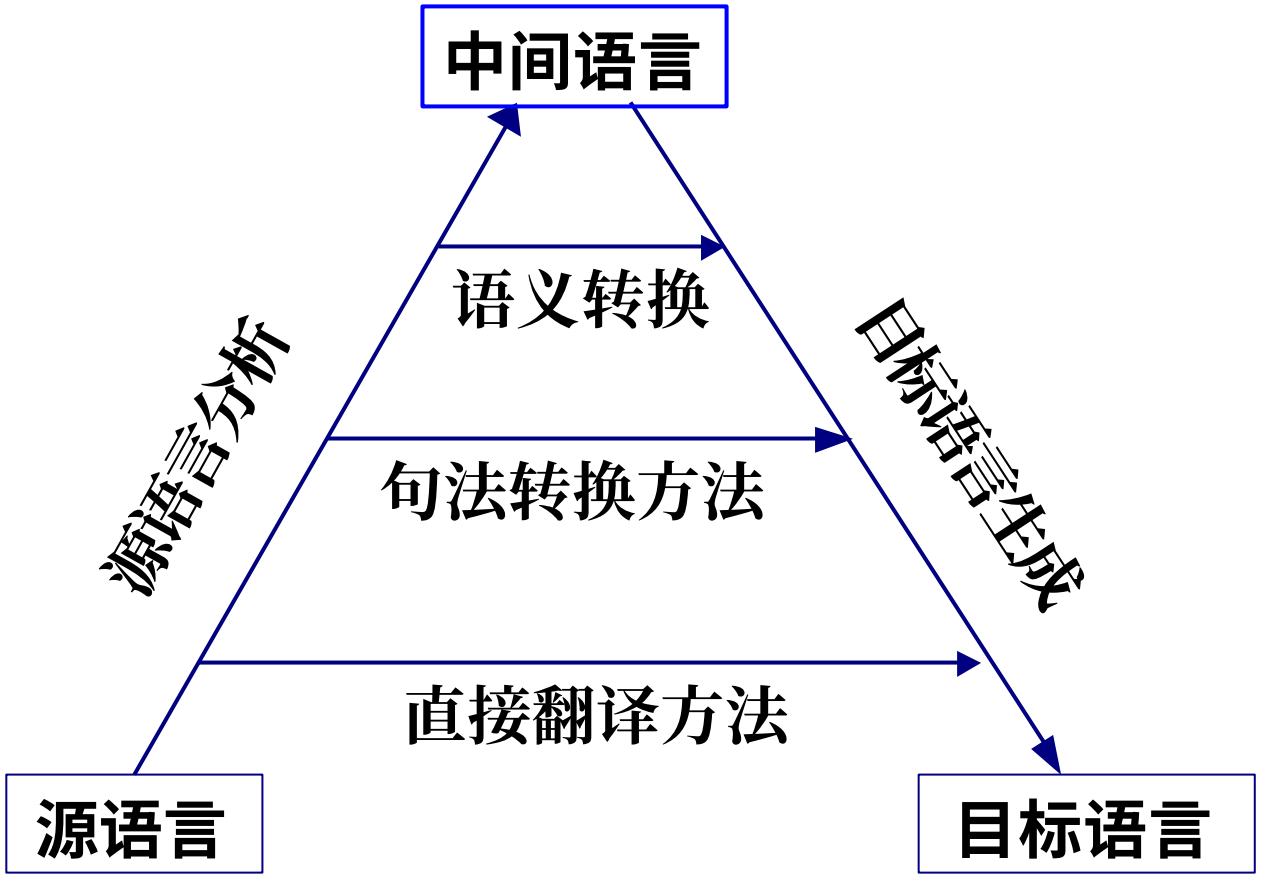
(c) **基于统计模型的翻译方法**（statistical method）

(d) **基于神经网络的翻译方法**（neural network MT）

(e) **基于模板的翻译方法** (template-based)

……

10.2 机器翻译基本方法



10.2 机器翻译基本方法

10.3 机器翻译研究现状

－若干系统已实用化，如：

. Systran (http://www.systransoft.com)

. TAUM-METEO

. Google

. 华建英汉翻译系统等

－仍面临若干问题

. 科学问题　　. 用户问题

10.2 机器翻译基本方法

一 应尽快消除对机器翻译的误解

黛玉自在枕上感念宝钗……又听见窗外竹梢焦叶之上，雨声淅沥，消寒透幕，不觉又滴下泪来。

As she lay there alone, Dai-Yu's thoughts turned to Bao-chai …Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she bad begun to cry.

10.3 关于分析方法与统计方法的思考

1、日本 ATR对两种口语翻译方法的比较*

依据的语料：

	日语	英语
句子数	204,108	204,108
词汇个数	1,689,449	1,235,747
词汇量	19,640	15,374
平均长度（单词/句子）	8.3	6.1

* Eiichiro Sumita, Corpus-Centered Computation, in *Proc. ACL Workshop: Speech-to-speech Translation*. Philadelphia, USA. July 11, 2002. pp. 1-8.

10.3 关于分析方法与统计方法的思考

实验系统测试结果：

	A	A+B	A+B+C
日英统计翻译系统[1]	25%	46%	64%
英日统计翻译系统	41%	48%	57%
基于事例的日英翻译系统－1（D[3]）[2]	47%	66%	77%
基于实例的日英翻译系统－2（HPAT）[3]	50%	61%	71%

[1] 基于词的统计翻译模型

[2] Dp-match Driven transDucer [Sumita, 2001, Example-based Machine Translation Using DP-matching between Word Sequences. *Proc. of DDMT(ACL)*, pp. 1-8.]

[3] Hierarchical Phrase Alignment (HPA) based Translation (HPAT) [Imamura, 2002, Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment, *Proc. of TMI*.]

10.3 关于分析方法与统计方法的思考

2、Verbmobil 口语翻译系统对两种方法的比较*

（1）Verbmobil 系统概况:

- **受德国联邦教育研究部**（German Ministry for Education and Research，BMBF）**资助**
  - **－第一阶段**（1993－1996）：$33M
  - **－第二阶段**（1997－2000）：$28M
  - **工业界**：$17M
  - **其它**： $11M
  - **共计**： $89M
- 23**个参加单位，**900**多位全职研究人员和学生**

** Wolfgang Wahlster, Verbmobil Multilingual Processing of Spontaneous Speech, www.dfki.de/~wahlster/VM-final*

10.3 关于分析方法与统计方法的思考

（2）Verbmobil 口语翻译系统构成:

- 10,175 **德语单词，**6871 **英语词汇**
- **统计翻译引擎：**58,332 **德英语句对训练，**8.9 **德词**/Sent.**，**9.4**英词**/Sent.
- **基于格的翻译（**Case-based**）引擎：**30,000**模板**(Template)
- **基于转换的翻译（**Transfer-based**）引擎：**

  22,783**格转换规则；**13,640**个微观规划规则**

- **基于对话意图（**Dialogue-Act**）的翻译器：**334**个多语言** FST (Finite State Transducers)
- **基于子串的（**Substring-based**）翻译器：**24,680**德英语句对训练**

10.3 关于分析方法与统计方法的思考

（3）系统评测方法：

- 基于网络的大规模测试：43,180 Translations
  （25,345(德)+17,835(英)）
- 65评估人员
- 句子长度：1－60单词

10.3 关于分析方法与统计方法的思考

实验系统的测试结果：

翻译引擎与最终系统	翻译词正确率 ³ 50%	翻译词正确率 ³ 75%	翻译词正确率 ³ 80%
Case-based Translation	37%	44%	46%
Statistical Translation	69%	79%	81%
Dialogue-act based Trans.	40%	45%	46%
Semantic Transfer (SeT)	40%	47%	49%
Substring-based Translation	65%	75%	79%
Automatic Selection	57% / 78%*	66% / 83%*	68% / 85%*

** After training with instance-based learning algorithm*

10.3 关于分析方法与统计方法的思考

3、问题与思考

- ATR的D^3与Verbmobil系统中的 SeT 比较孰优孰劣？意义？
- 两个系统翻译的语言对差异较大：日 ↔ 英，德 ↔ 英
- 两个统计翻译引擎都是基于词对位，基于短语、块或其它将如何？
- 训练语料的规模对统计翻译系统的影响？（D^3: 20.4 万; SeT: 5.8万）
- 统计翻译方法和基于分析的翻译方法都是孤军奋战？

10.3 关于分析方法与统计方法的思考

4、统计翻译方法的“石头汤 (Stone soup)”之说

(http://spanky.triumf.ca/www/fractint/stone_soup.html)

- in Eastern Europe, there was a great famine
- a peddler drove his wagon into a village
- iron cauldron, water, stone, fire
- most of the villagers had come to the square or watched
- cabbage, salt beef, potatoes, onions, carrots, mushrooms, … …
- a delicious meal

10.3 关于分析方法与统计方法的思考

5、观点与认识

1) 统计方法的崛起打破了分析方法在整个自然语言处理领域一统天下的僵局，但并不意味着分析方法的结束

2) 在很多方面，分析方法与统计方法并不是完全对立的

- ❖ 绝对化与模糊化
- ❖ 唯一化与多选化
- ❖ 主观性与客观性
- ❖ 定性与定量

10.3 关于分析方法与统计方法的思考

3) 统计方法与分析方法的结合是必然的结局

- 知识准备与资源建设
  - 词对 / 规则 / 模板 / 实例 的自动抽取与学习
- 处理方法实现过程
  - 句法分析与语义提取、Chunk / 句子边界检测、远距离相关分析

4) 多翻译引擎协同竞争机制是口语翻译系统首选的策略

- 基于模板（Template）或模式（Pattern）的方法
- 基于规则或中间语言（语义表示）的方法
- 基于统计的方法
- 基于实例的方法

10.4 统计翻译基本原理

❑ 基本思想

1990年IBM 的Peter F. Brown 等人在计算语言学杂志（Computational Linguistics）发表的论文“统计机器翻译方法”[Brown, 1990]和1993年他们发表在该杂志的“统计机器翻译的数学：参数估计”一文[Brown, 1993]奠定了统计机器翻译的基础。

– 噪声信道模型

10.4 统计翻译基本原理

源语言句子： $S = s_1^m \equiv s_1 s_2 \cdots s_m$

目标语言句子： $T = t_1^l \equiv t_1 t_2 \cdots t_l$

$$P(T \mid S) = \frac{P(T)P(S \mid T)}{P(S)} \qquad (10\text{-}1)$$

$$\hat{T} = \arg\max_{T} P(T)P(S \mid T) \qquad (10\text{-}2)$$

语言模型
Language model, LM

翻译模型
Translation Model

10.4 统计翻译基本原理

通常将实现这个搜索过程的模块称为解码器（Decoder）。

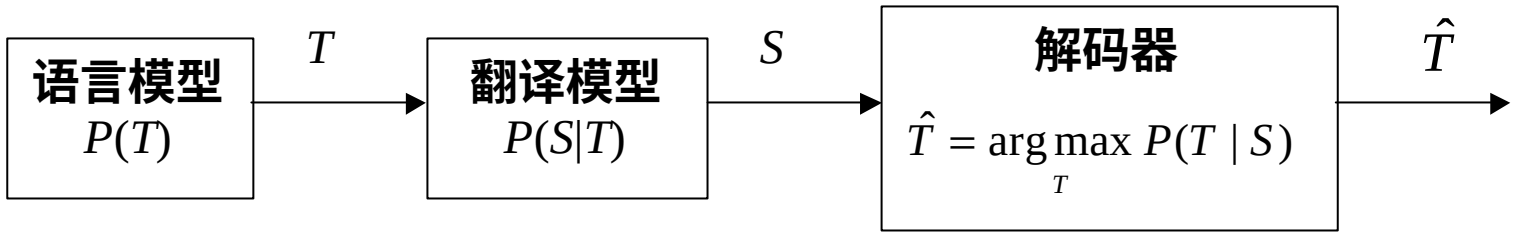


图10-1. 统计机器翻译中的噪声信道模型

统计翻译必须解决三个关键的问题：

（1）估计语言模型概率$P(T)$；

（2）估计翻译概率$P(S \mid T)$；

（3）设计有效快速的搜索算法以求解使得$P(T) \times P(S \mid T)$最大。

10.4 统计翻译基本原理

- **估计语言模型概率** $P(T)$

给定句子： $t_1^l = t_1 t_2 \cdots t_l$

概率： $P(t_1^l) = P(t_1)P(t_2 \mid t_1)\cdots P(t_l \mid t_1 t_2 \cdots t_{l-1})$

N-gram **问题，不再赘述。**

10.4 统计翻译基本原理

❑ 翻译概率$P(S|T)$的计算

一个很关键的问题是怎样定义目标语言句子中的词与源语言句子中的词之间的对应关系。

假设英语与法语的翻译对：

(Le programme a été mis en application | And the (1) program (2) has (3) been(4) implemented (5, 6, 7))

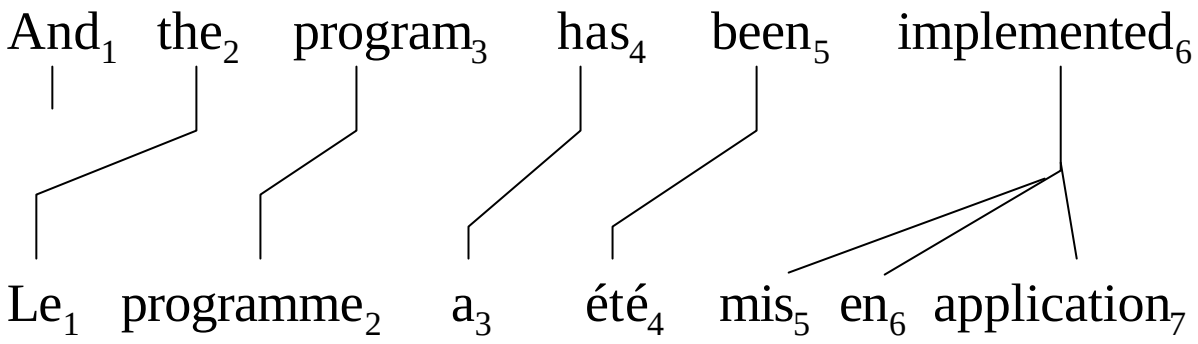


图10-2

10.4 统计翻译基本原理

不妨，我们用$A(S, T)$表示源语言句子S与目标语言句子T之间所有对位关系的集合。在目标语言句子T的长度（单词的个数）为l，源语言句子S的长度为m的情况下，T和S的单词之间有$l \times m$种不同的对应关系。由于一个对位是由词之间的对应关系决定的，并且不同的对应方式应该是$2^{l \times m}$的子集，因此，$A(S, T)$中共用$2^{l \times m}$种对位。

对于一个给定的句对(S|T)，我们可以假定所有的单词对(s_j, t_i)之间存在着对应关系。那么，用来刻画这些对应关系的模型叫做对位模型（alignment model）

10.4 统计翻译基本原理

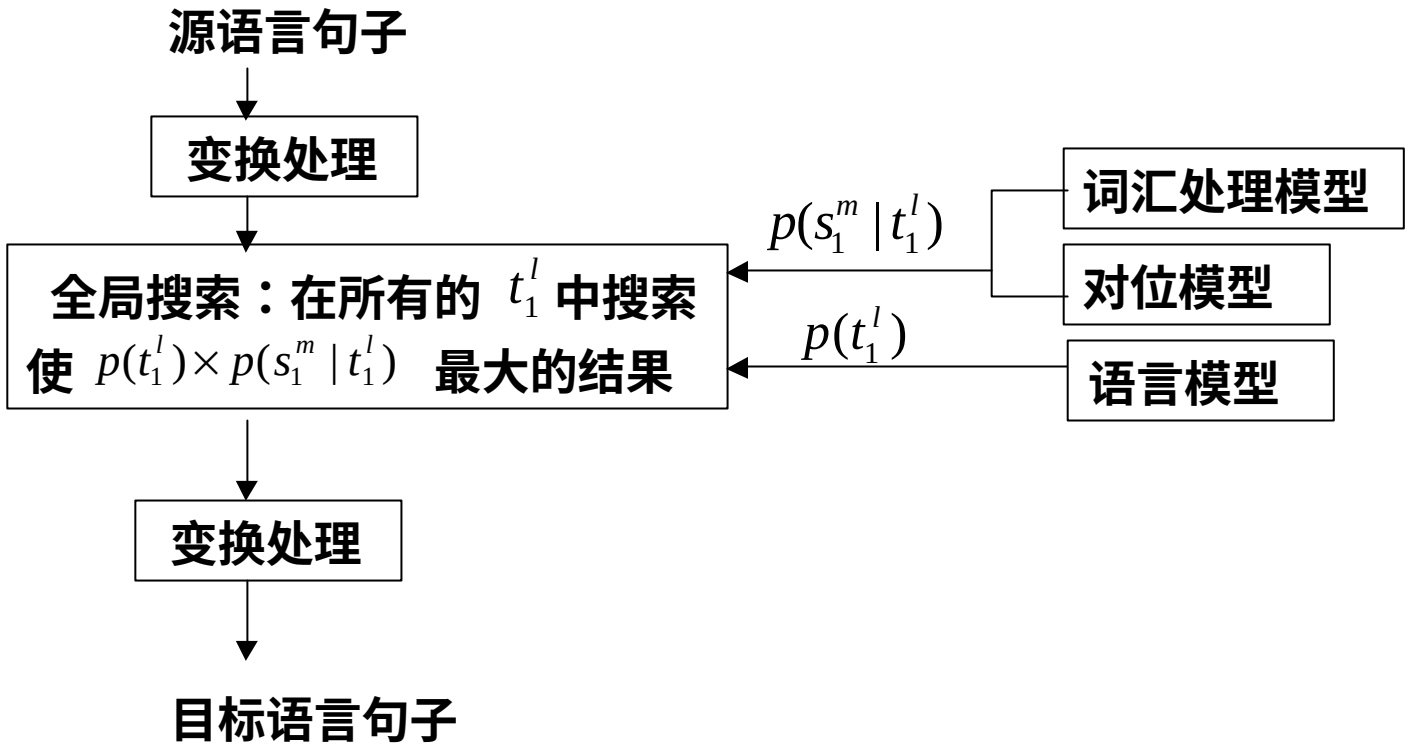


图10-3. 统计翻译系统框架

10.4 统计翻译基本原理

❑ 基本的数学问题

求解联合概率分布$P(S=S, A=A, T=T)$，其中，S,T 分别表示翻译中的源语言和目标语言字符串随机变量，A 为S与T 之间的对位关系的随机变量。S, A, T 分别表示随机变量S，A，T 的一个具体取值。

约定用l, m 分别表示目标语言句子的长度和源语言句子的长度，现在我们约定L 和M分别表示长度l 和m 的随机变量。在不引起混淆的情况下，我们一般用$P(S, A, T)$ 替代$P(S=S, A=A, T=T)$。

10.4 统计翻译基本原理

翻译句对$(S \mid T)$ 的似然率可以通过条件概率$P(S, A \mid T)$ 获得：

$$P(S \mid T) = \sum_{A} P(S, A \mid T) \qquad (10\text{-}3)$$

按照前面的约定，源语言句子 $S = s_1^m \equiv s_1 s_2 \cdots s_m$ 有m个单词，目标语言句子 $T = t_1^l \equiv t_1 t_2 \cdots t_l$ 有l个单词，对位序列表示成：

$$A = a_1^m = a_1 a_2 \cdots a_m$$

其中，a_j $(j = 1 \,..\, m)$的取值范围为0到l之间的整数，如果源语言中的第j个词与目标语言中的第i个词对齐，那么，$a_j = i$，如果没有词与它对齐，则$a_j = 0$。

10.4 统计翻译基本原理

不失一般性，

$$P(S,A|T)=P(m|T)\prod_{j=1}^{m}P(a_j|a_1^{j-1},s_1^{j-1},m,T)P(s_j|a_1^{j},s_1^{j-1},m,T) \qquad (10\text{-}4)$$

实际上，$P(S,A|T)$ 可以写成多种形式的条件概率的乘积，（11-4）式只是其中的一种。

10.5 IBM-1 翻译模型

❑IBM-1 模型

在上面的（11-4）式中，由于等号右边有太多的参数，因此，我们不能保证这些参数之间总是互相独立的。因此，在遵循如下假设的情况下，我们得到翻译模型－1。

1)假定$P(m|T)$ 与目标语言T和源语言的句子长度m无关，那么，

$e \equiv P(m \mid T)$ 是一个比较小的常量；

2)假定 $P(a_j \mid a_1^{j-1}, s_1^{j-1}, m, T)$ 只依赖于目标语言的长度l，那么，

$$P(a_j \mid a_1^{j-1}, s_1^{j-1}, m, T) = \frac{1}{l+1}$$

10.5 IBM-1 翻译模型

3) **假定** $P(s_j \mid a_1^j, s_1^{j-1}, m, T)$ **仅依赖于**s_j**和** t_{a_j} **，那么，我们称**

$$p(s_j \mid t_{a_j}) \equiv P(s_j \mid a_1^j, s_1^{j-1}, m, T)$$

为给定 t_{a_j} **的情况下单词**s_j **的翻译概率**（translation probability）。

10.5 IBM-1 翻译模型

如何估计模型－1的翻译概率？

根据上面的假设，在给定目标语言句子的情况下，源语言句子和对位关系的联合似然率为：

$$P(S,A|T)=\frac{e}{(l+1)^m}\prod_{j=1}^{m}p(s_j|t_{a_j}) \qquad (10\text{-}5)$$

10.5 IBM-1 翻译模型

(10-5)式的理解：

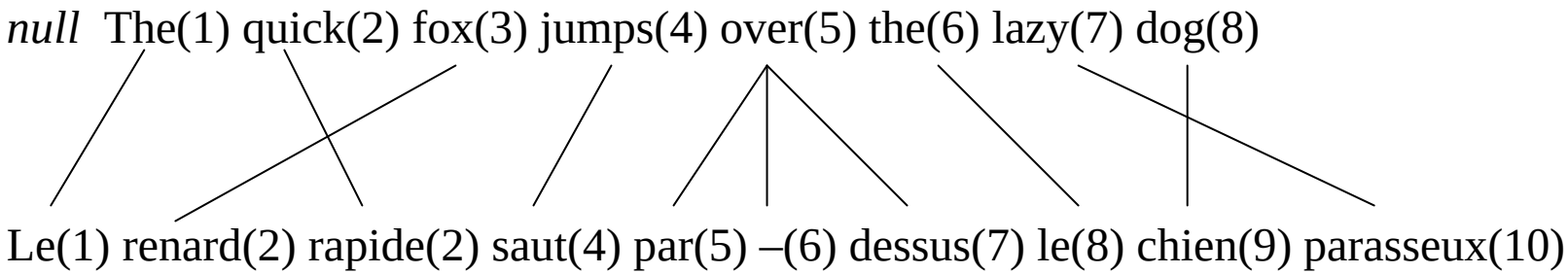


图10-4

$$P(S,A|T)=\frac{e}{(8+1)^{10}}\times[\underbrace{p(Le|The)\times p(renard|\ fox)\times\cdots\times p(parasseux|\ lazy)}_{\text{共10项}}]$$

10.5 IBM-1 翻译模型

由于对位关系由1到m个a_j 的具体值所决定，而每个a_j 的取值可以是0到l之间的任意数，因此，

$$P(S|T)=\frac{e}{(l+1)^m}\sum_{a_1=0}^{l}\cdots\sum_{a_m=0}^{l}\prod_{j=1}^{m}p(s_j|t_{a_j}) \qquad (10\text{-}6)$$

我们需要知道所有词对$(s|t)$ 的对应概率p使得翻译概率$P(S|T)$最大，并且对于每一个给定的单词t满足以下约束条件：

$$\sum_{s}p(s|t)=1 \qquad (10\text{-}7)$$

10.5 IBM-1 翻译模型

为了求限定条件下概率$P(S \mid T)$达到最大值，我们引入拉格朗日乘法因子l_t，然后，求下列辅助函数的无限定条件的极大值：

$$h(p, l) \equiv \frac{e}{(l+1)^m} \sum_{a_1}^{l} \cdots \sum_{a_m}^{l} \prod_{j=1}^{m} p(s_j \mid t_{a_j}) - \sum_t l_t \left(\sum_s p(s \mid t) - 1 \right) \qquad (10\text{-}8)$$

根据求极大值的条件，h函数关于p和λ的偏导数应等于零。而该函数关于λ的偏导等于零，实际上只是简单地重申约束条件：翻译概率之和等于1，没有其它的意义。函数h关于$p(s|t)$的偏导数为：

10.5 IBM-1 翻译模型

$$\frac{\partial h}{\partial p(s|t)} = \frac{e}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \sum_{j=1}^{m} d(s,s_j) d(t,t_{a_j}) p(s|t)^{-1} \prod_{k=1}^{m} p(s_k|t_{a_k}) - l_t \tag{10-9}$$

其中，δ 是Kronecker函数，当它的两个参数相同时，$\delta=1$，否则，$\delta=0$。这样，在偏导数等于0情况下，可以求得

$$p(s|t) = l_t^{-1} \frac{e}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \sum_{j=1}^{m} d(s,s_j) d(t,t_{a_j}) \prod_{k=1}^{m} p(s_k|t_{a_k}) \tag{10-10}$$

10.5 IBM-1 翻译模型

（11-10）式给我们提供了一种利用迭代过程求解极大值的思路：给翻译概率一个任意的初始估计值，我们可以计算出等式右边的值，并可以利用这个值作为新的$p(s|t)$ 的估计值。

我们称这个重复进行的迭代过程为期望最大化（Expectation Maximization, EM）算法，简称EM算法。

借助于(11-6)式，我们可以将(11-10)式写成如下形式：

$$p(s|t) = l_t^{-1} \sum_A P(S,A|T) \sum_{j=1}^{m} d(s,s_j) d(t,t_{a_j}) \qquad (10\text{-}11)$$

10.5 IBM-1 翻译模型

…… ……

忽略详细的数学推导，IBM-1**模型表示为：**

$$P(S \mid T) = \frac{e}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} p(s_j \mid t_i) \qquad (10\text{-}16)$$

10.5 IBM-2 翻译模型

在模型－2中，我们除了假定概率 $P(a_j \mid a_1^{j-1}, s_1^{j-1}, m, T)$ 依赖于位置j、对位关系a_j 和源语言句子长度m以及目标语言句子长度l以外，另外两个假设与模型－1中的假设一样。引入了对位概率（alignment probabilities）的概念：

$$a(a_j \mid j, m, l) \equiv P(a_j \mid a_1^{j-1}, s_1^{j-1}, m, l)$$

对于每一个三元组(j, m, l)，对位概率满足如下约束条件：

$$\sum_{i=0}^{l} a(i \mid j, m, l) = 1$$

10.5 IBM-2 翻译模型

类似于IBM-1的推导，得到IBM-2模型：

$$P(S \mid T) = e \prod_{j=1}^{m} \sum_{i=0}^{l} p(s_j \mid t_i) a(i \mid j, m, l) \qquad (11\text{-}26)$$

如果对位概率设为常数，IBM-2模型退化为IBM-1模型，即模型1是模型2的特例。

10.5 IBM-3 翻译模型

定义：在随机选择对位关系的情况下，与目标语言句子中的单词t对应的源语言句子中的单词数目是一个随机变量，不妨记做Φ_t，该变量我们称之为单词t的繁衍能力或产出率(fertility)。一个具体的取值记做：f_t

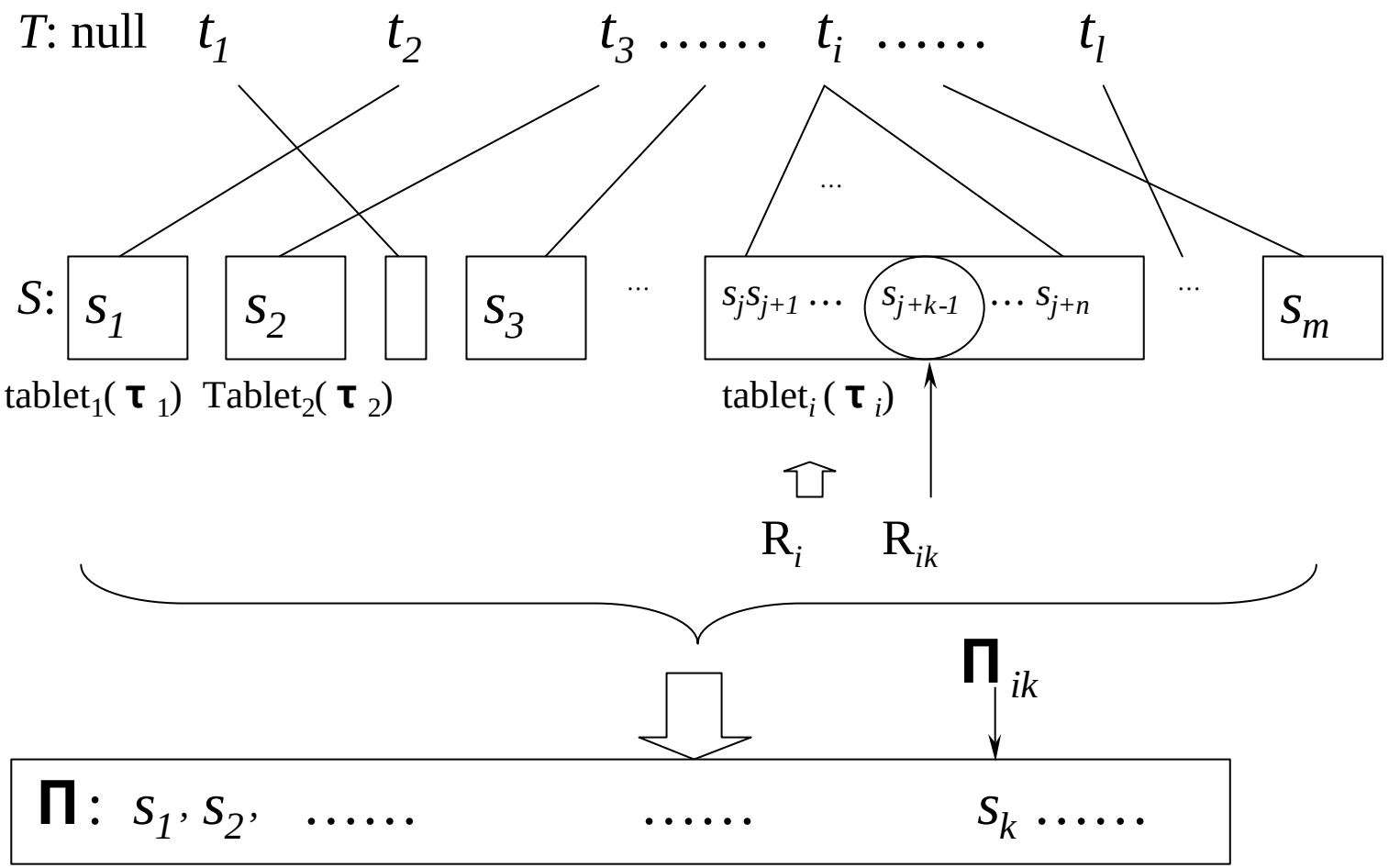
实际上，所谓的繁衍能力就是目标语言单词与源语言单词之间一对多的关系。

10.5 IBM-3 翻译模型

定义：假设给定一个目标语言句子T，T中的每一个单词t在源语言句子中可能有若干个词与之对应，源语言句子中所有与t对位的单词列表我们称之为t的一个便笺或写字板(tablet)，当然这个便笺可能为空。一个目标语言句子T的所有便笺的集合是一个随机变量，我们称之为T的便笺集或标释集(tableau)，记做符号R。

T的第i个单词的便笺也是一个随机变量，不妨记做R_i，那么，T的第i个单词的便笺中第k个源语言单词也是一个随机变量，我们记做R_{ik}。

10.5 IBM-3 翻译模型



10.5 IBM-3 翻译模型

标释集 τ (R的一个具体取值) 和单词排列 π(Π 的一个具体取值，即τ中单词的一种排列方式) 的联合似然率为：

$$P(t,p\mid T)=\prod_{i=1}^{l}P(f_i\mid f_1^{i-1},T)P(f_0\mid f_1^{l},T)\times$$

繁衍概率
(fertility prob.)

$$\prod_{i=0}^{l}\prod_{k=1}^{f_i}P(t_{ik}\mid t_{i1}^{k-1},t_0^{i-1},f_0^{l},T)\times$$

翻译概率
(tran. prob.)

$$\prod_{i=1}^{l}\prod_{k=1}^{f_i}P(p_{ik}\mid p_{i1}^{k-1},p_1^{i-1},t_0^{l},f_0^{l},T)\times$$

$$\prod_{k=1}^{f_0}P(p_{0k}\mid p_{01}^{k-1},p_1^{l},t_0^{l},f_0^{l},T)$$

位置概率
(distortion prob.)

10.5 IBM-3 翻译模型

假设：

1) **对于1到l中的每一个i，概率** $P(f_i \mid f_1^{i-1}, T)$ **仅依赖于** f_i **和** t_i**，记做，** $n(f \mid t_i) \equiv P(f \mid f_1^{i-1}, T)$

2）**对于所有的i，概率** $P(t_{ik} \mid t_{i1}^{k-1}, t_0^{i-1}, f_0^l, T)$ **只依赖于τ_{ik}和t_i，记做：** $p(s \mid t_i) \equiv P(R_{ik} = s \mid t_{i1}^{k-1}, t_0^{k-1}, f_0^l, T)$

3）**对于1到l中的每一个i，概率** $P(p_{ik} \mid p_{i1}^{k-1}, p_1^{i-1}, t_0^l, f_0^l, T)$ **只依赖于π_{ik}，i, m和l。位置概率记作：**

$$d(j \mid i, m, l) \equiv P(\Pi_{ik} = j \mid p_{i1}^{k-1}, p_1^{i-1}, t_0^l, f_0^l, T)$$

10.5 IBM-3 翻译模型

设想 t_1^l 中每个写字板中的一组词都存在一个额外词（即这个词在对齐时对空），假设这个额外词出现的概率为p_1。另外，我们知道：$f_0 + f_1 + \cdots + f_l = m$

因此，

$$P(S|T) = \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} P(S, A|T)$$

$$= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \binom{m-f_0}{f_0} (1-p_1)_0^{m-2f_0} p_1^{f_0} \prod_{i=1}^{l} f_i! n(f_i|t_i) \times \prod_{j=1}^{m} p(s_j|t_{a_j}) d(j|a_j, m, l)$$

10.5 IBM-3 翻译模型

其中，$\sum_s p(s \mid t) = 1$　　$\sum_j d(j \mid i, m, l) = 1$　　$\sum_f n(f \mid t) = 1$

估计这些参数和 p_1。

模型－4考虑标释集的中心词的概率和其它单词的位置概率。

模型－5源语言句子单词间的相对位置。

请参见文献 [P. F. Brown, 1993]。

10.6 搜索算法

搜索空间随源语言句子的长度增长呈指数增长。

常用的搜索算法：

A* **搜索**;

堆栈搜索(stack search);

柱搜索 (beam search)。

… …

10.7 统计翻译方法的改进

- **基于结构的翻译模型** [Wang Yeyi, 1998]
  - **粗对齐**（rough alignment）：

    源语言和目标语言的短语之间；

    在训练语料库上，通过基于互信息的双语词语聚类和短语归并迭代，得到基于词语聚类的短语规则，利用这些规则进行短语分析。
  - **细对齐**（detailed alignment）：

    短语内部的单词细对齐。

10.7 统计翻译方法的改进

- **基于句法结构的翻译模型** [Yamada, Knight, 2001]

  – 输入：源语言句子的句法树，输出：目标语言句子。

  – 变换：1）将句法树扁平化处理 (相同中心词的多层结点压缩到一层)；

  2) **句法树上每个结点的子结点进行随机重新排列，每种排列方式都有一个概率；**

  3) **对于句法树任何一个位置随机地插入一个新的目标语言单词，每个位置、每个被插入的单词都有不同的概率；**

  4) **对句法树上每个叶子结点的源语言单词翻译成目标语言，每个不同的译文词都有不同的概率；**

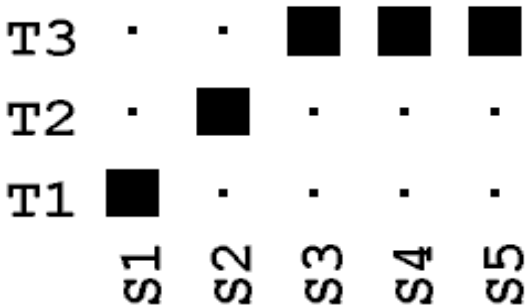
  5) **根据句子概率（上述概率乘积）输出句子。**

10.7 统计翻译方法的改进

- RWTH **的工作**[Hermann, 2000; Och, 2002**等**]
  - **运用词聚类技术，采用基于类的模型，解决数据稀疏问题；**
  - **在语言模型上，采用基于类的**5-gram**；**
  - **在翻译模型上，提出对齐模板**(alignment template)**方法，实现两种层次的对齐：短语层次对齐、词语层次对齐；**

    对齐模板采用基于类的对齐矩阵形式表示。

10.7 统计翻译方法的改进



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, ...

10.7 统计翻译方法的改进

- **基于最大熵的统计翻译方法** [Och, 2002等]

选择一组特征，使得统计模型在这一组特征上与样例中的分布完全一致，同时保证这个模型尽可能的“均匀”（使模型的熵值达到最大）。

假设T、S分别是目标语言句子和源语言句子， $h_1(T, S), \ldots, h_M(T, S)$ **分别是T、S上的M个特征，** $\lambda_1, \ldots, \lambda_M$ **是这些特征分别对应的M个权值。**

10.7 统计翻译方法的改进

对于给定的源语言句子S, 其最佳译文T可以用以下公式表示：

$$\hat{e} = \arg\max_{T} \{P(T \mid S)\}$$

$$= \arg\max_{T} \{\sum_{m=1}^{M} l_m h_m(T,S)\}$$

如果我们将两个特征分别取为log P(T) **和** log P(S|T)**，并取**$\lambda_1 = \lambda_2 = 1$**，那么，这个模型就等价于噪声信道模型。最大熵方法中最常用的做法就是采用二值特征。**

10.7 统计翻译方法的改进

- Stephan Vogel **等人**(CMU) **对统计翻译方法的改进**

  1) **从大规模双语对齐语料中获得对齐短语或词汇，比如，通过**HMM**方法；**

  2) **根据对齐短语的不同组合，进行搜索。**

10.8 译文评估方法

- **常用的评测指标**

1) **句子错误率：译文与参考答案不完全相同的句子为错误句子。错误句子占全部译文的比率。**

2) **单词错误率：译文与参考答案之间的编辑距离的比率。**

3) **与位置无关的单词错误率**

4) **流畅度**

5) **忠实度**

10.8 译文评估方法

- IBM的BLEU评价方法

  －BiLingual Evaluation Understudy

基本思想：

统计同时出现在系统译文和参考译文中的N 元词的个数，最后把匹配到的N元词的数目除以系统译文的单词数目，得到评测结果。

10.8 译文评估方法

- BLEU评价方法

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \log p_n)$$

最大语法的阶数，实际取4。

长度过短句子的惩罚因子

w_n=1/N

出现在答案译文中的n元词语接续组占候选译文中n元词语接续组总数的比例。

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

c为候选译文中单词的个数，r 为答案译文中与c 最接近的译文单词个数。

10.8 译文评估方法

- NIST **评估方法**

- National Institute of Standards and Technology, USA

基本思想：

在BLEU方法的基础上提出来的，如果一个N元词在参考译文中出现的次数越少，表明它所包含的信息量越大，那么，它对于该N元词就赋予更高的权重。

10.8 译文评估方法

- **统计翻译中的译文错误**

1) **模型错误：概率最高的译文不是正确的。**

2) **搜索错误：概率最高的译文是正确的，但搜索算法找不到。这类错误大约占5%。**

10.9 统计翻译的相关活动

- John Hopukins University (JHU) **的夏季研讨班**

1) 1999**年夏天**

2) **开发了原代码公开的统计翻译根据包**－EGYPT

- GIZA：**从双语语料库中抽取统计模型知识(参数训练)；**

- Decoder：**执行翻译过程；**

- Cairo：**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。**

- Whittle：**语料预处理工具。**

10.9 统计翻译的相关活动

- IWSLT: International Workshop on Spoken Language Translation
  - 从2004年开始，每年一次；
  - 旅游领域的口语翻译和新闻领域文本翻译。

- DARPA的机器翻译评测活动
  - 2002年6月, NIST 首次正式的机器翻译评估。
  - 每年一次。

10.9 统计翻译的相关活动

主要研究机构

- RWTH: Aachen University of Technology, Germany
- JHU, USA
- CMU, USA
- ATR, Japan

本章小结

- 机器翻译的发展
- 机器翻译基本方法
- 统计机器翻译的基本原理
- IBM统计机器翻译模型
- 统计机器翻译的改进
- 关于译文评估

习题

1. 认真查阅关于 IBM 统计翻译模型及其相关技术的文献，学习并掌握统计翻译方法。
2. 下载有关的算法代码，实现一个小型的基于统计翻译模型的汉英翻译系统，掌握实现技术。
3. 查阅有关口语理解方法论文，了解口语解析技术的基本思想。
4. 思考如何将基于规则的翻译方法和基于统计的翻译方法相结合，实现高质量的口语翻译系统？

Thanks

谢谢!