

# 自然语言处理

*Natural Language Processing*

权小军 教授

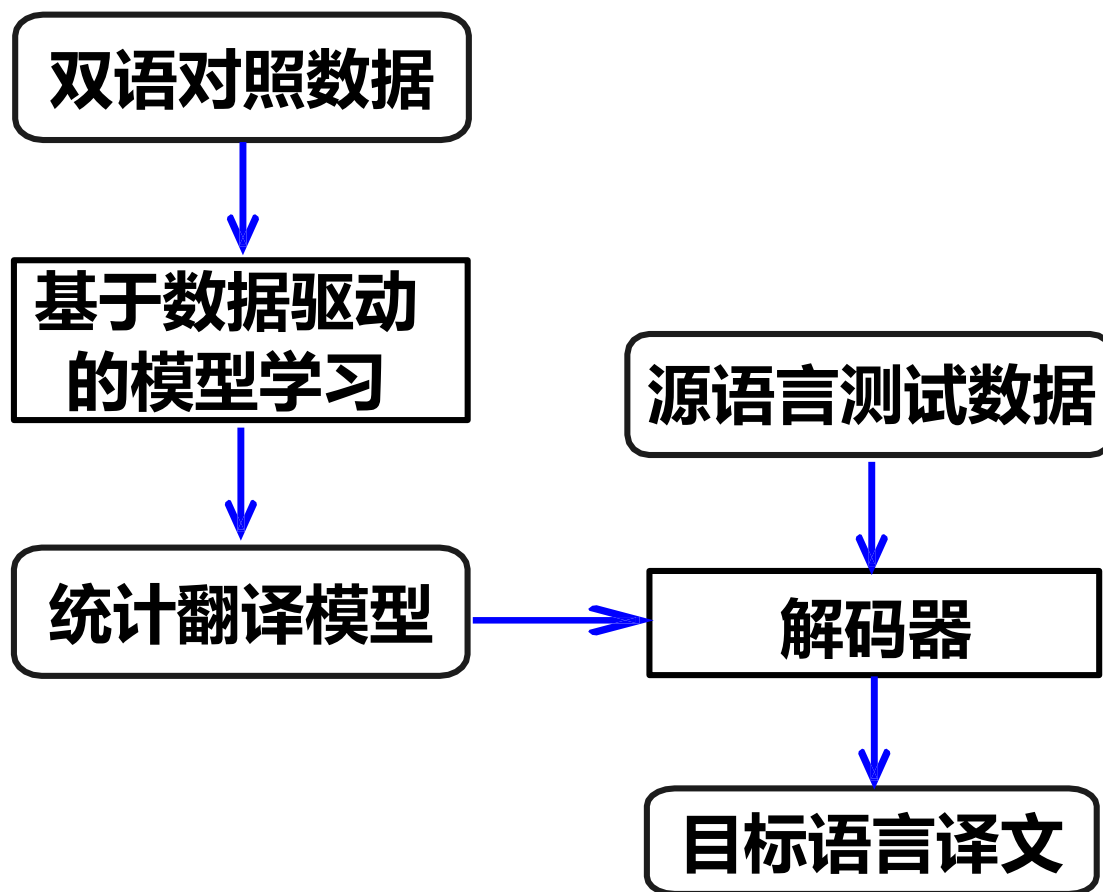
中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

# Lecture 13: 统计机器翻译

# 12.1: 基本思想

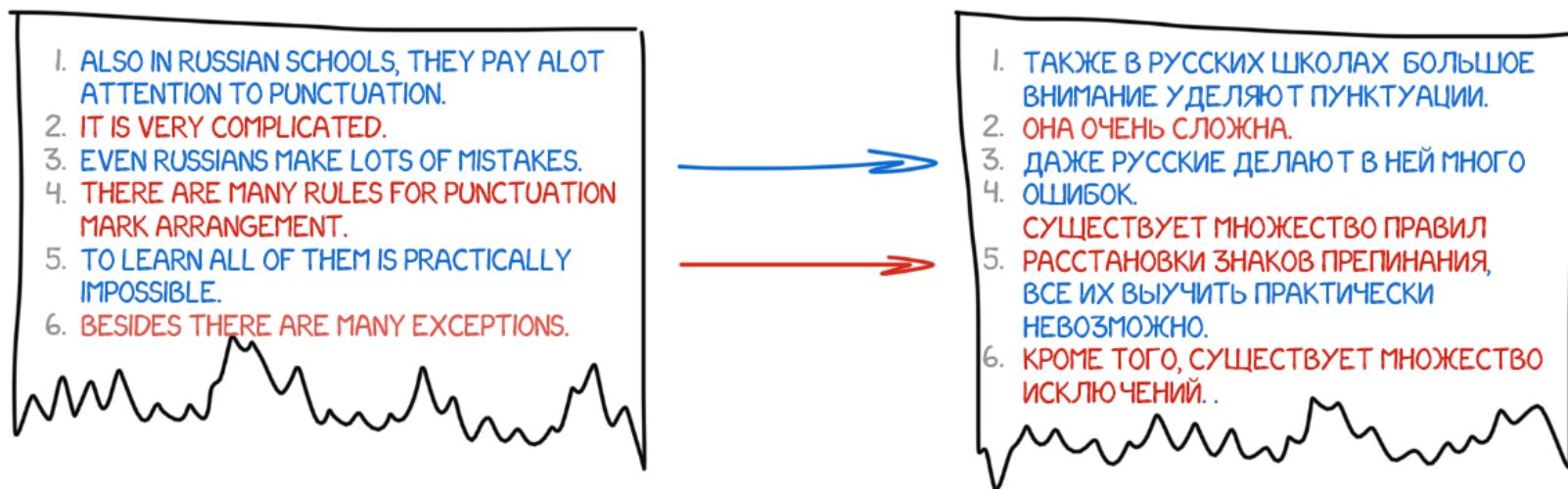
# 统计翻译的思想



# 统计翻译的思想

## □ 平行语料

### PARALLEL CORPUS



宗成庆：《自

# 统计翻译的思想

## □ 平行语料

	facing with the swelling flow of through traffic zooming past their doors .		recunda de inconvenientes que más y más gente tiene que soportar por el tráfico que pasa por delante de sus casas , que aumenta a diario .
5	#77501757 Weekend traffic bans and traffic jams are a curse to road transport .	#74765580	Las prohibiciones de conducir los fines de semana y los embotellamientos asolan el transporte por carretera .
6	#79500725 Some people also want to recoup the cost of traffic jams from those who get stuck in them , according to the 'polluter pays ' principle .	#76764676	Algunos son partidarios de que incluso los costes ocasionados por los atascos se carguen a el ciudadano que se encuentra atrapado en ellos , de conformidad con el principio de que " quien contamina paga " .
7	#79500765 I think this is an excellent principle and I would like to see it applied in full , but not to traffic jams .	#76764713	Me parece un principio acertado y estoy dispuesta a aplicarlo íntegramente , pero no sobre los atascos , ya que éstos son un claro indicio de el fracaso de la política gubernamental en materia de infraestructuras .
8	#79500768 Traffic jams are indicative of failed government policy on the infrastructure front , which is why the government itself , certainly in the Netherlands , must be regarded as the polluter .	#76764747	Por eso es preciso subrayar que en estos casos quien contamina es el propio Gobierno , a el menos en los Países Bajos .
9	#81309716 This would increase traffic jams , weaken road safety and increase costs .	#78586130	Esto aumentaría los atascos , mermaría la seguridad vial e incrementaría los costes .
10	#81997391 In the previous legislature , Parliament gave its opinion on the Commission 's proposals on the simplification of vertical directives on sugar , honey , fruit juices , milk and jams .	#79281114	En efecto , durante la precedente legislatura , el Parlamento se manifestó sobre las propuestas de la Comisión relativas a la simplificación de directivas verticales sobre el azúcar , la miel , los zumos de frutas , la leche y las confituras .
11	#81998167 For jams , I personally reintroduced an amendment that was not accepted by the Committee on the Environment , Public Health and Consumer Policy , but which I hold to .	#79281936	Para las confituras , yo personalmente volví a introducir una enmienda que no fue aceptada por la Comisión de Medio Ambiente , Salud Pública y Política de el Consumidor , pero que es importante para mí .
12	#81998209 It concerns not accepting the general use of a chemical flavouring in jams and marmalades , that is vanillin .	#79281966	Se trata de no aceptar la utilización generalizada de un aroma químico en las confituras y " marmalades " , a saber , la vainillina .
13	#82800065 This is highlighted particularly in towns where it is necessary to find ways of solving environmental problems and the difficulties caused by traffic jams .	#80085988	Esto se pone de relieve aún más en las ciudades , en las que hay que encontrar medios para eliminar los inconvenientes derivados de los problemas medioambientales y de la congestión de el tráfico .

宗成庆：《自

# 统计翻译的诞生

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

# 统计翻译基本原理

## □ 噪声信道模型

一种语言  $T$  由于经过一个噪声信道而发生变形，从而在信道的另一端呈现为另一种语言  $S$ （信道意义上的输出，翻译意义上的源语言）。翻译问题实际上就是如何根据观察到的  $S$ ，恢复最为可能的  $T$  问题。





# 统计翻译基本原理

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

→ 目标语言句子:  $T = t_1^n = t_1 t_2 \cdots t_n$

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

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

语言模型  
Language model, LM

翻译模型  
Translation model, TM

# 统计翻译基本原理

## 统计翻译中的三个关键问题：

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

# 统计翻译基本原理

## □ 估计语言模型概率 $P(T)$

给定句子:  $T = t_1^l = t_1 t_2 \cdots t_n$

怎么估算T的概率? ? ?

句子概率:  $P(T) = P(t_1)P(t_2|t_1) \cdots P(t_n | t_1 t_2 \cdots t_{n-1})$

**基于  $n$ -gram 来计算!**

# 统计翻译基本原理

例句：我们一定要不忘初心

*Candidate 1:* We forget must our heart

*Candidate 2:* We forget must not our original heart

*Candidate 3:* We not must not forget our original heart

*Candidate 4:* We must forget our original intention

*Candidate 5:* We must not forget our original intention

*Candidate 6:* We must remember our original intention

# 统计翻译基本原理

例句：我们一定要不忘初心

3-gram

$P$ (We forget must our heart)	0.001
--------------------------------	-------

$P$ (We forget must not our original heart)	0.012
---	-------

$P$ (We not must not forget our original heart)	0.002
---	-------

$P$ (We must forget our original intention)	0.049
---	-------

$P$ (We must not forget our original intention)	0.051
---	-------

$P$ (We must remember our original intention)	0.045
---	-------

# 统计翻译基本原理

例句：我们一定要不忘初心

3-gram

$P(\text{We forget must our heart})$

~~0.001~~

$P(\text{We forget must not our original heart})$

~~0.012~~

$P(\text{We not must not forget our original heart})$

~~0.002~~

$P(\text{We must forget our original intention})$

0.049

$P(\text{We must not forget our original intention})$

0.051

$P(\text{We must remember our original intention})$

0.045

# 统计翻译基本原理

例句：我们一定要不忘初心

3-gram

$P(\text{We forget must our heart})$	0.001
--------------------------------------	-------

$P(\text{We forget must not our original heart})$	0.012
---	-------

$P(\text{We not must not forget our original heart})$	0.002
---	-------

$P(\text{We must forget our original intention})$	0.049
---	-------

$P(\text{We must forget our original intention})$	0.051
---	-------

$P(\text{We must remember our original intention})$	0.045
---	-------

翻译模型登场！

# 统计翻译基本原理

## □ 翻译概率 $p(S|T)$ 的计算

关键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：

<b><i>T</i></b>	And <sub>1</sub> the <sub>2</sub> program <sub>3</sub> has <sub>4</sub> been <sub>5</sub> implemented <sub>6</sub>
<b><i>S</i></b>	Le <sub>1</sub> programme <sub>2</sub> a <sub>3</sub> été <sub>4</sub> mis <sub>5</sub> en <sub>6</sub> application <sub>7</sub>

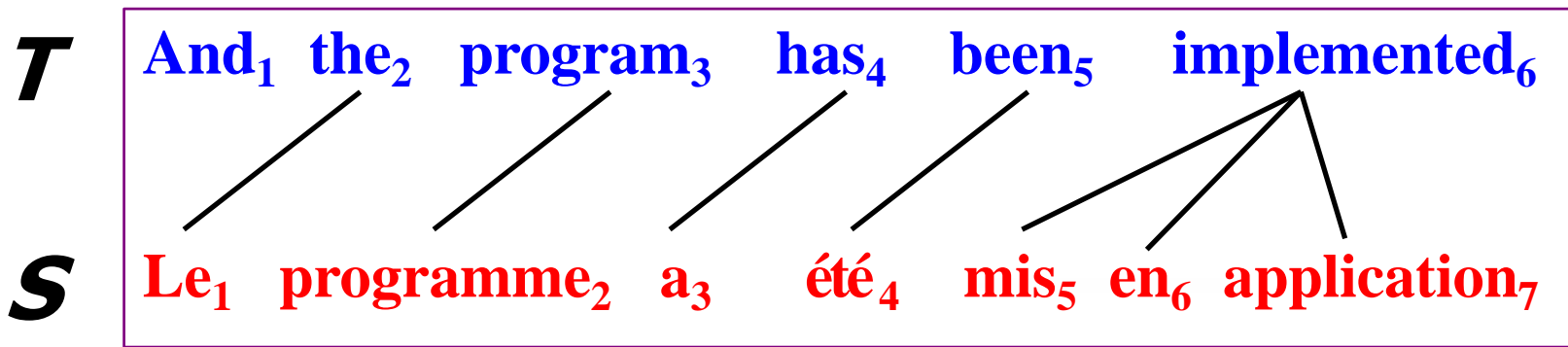


# 统计翻译基本原理

## □ 翻译概率 $p(S|T)$ 的计算

关键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：



# 统计翻译基本原理

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。目标语言句子 $T$ 的长度为 $n$ ，源语言句子 $S$ 的长度为 $m$ ， $T$ 和 $S$ 的单词之间有 $2^{n \times m}$ 种不同的对应关系：

$$|\mathcal{A}(S, T)| = 2^{n \times m} \quad A(S, T) \in \mathcal{A}(S, T)$$

$A(S, T)$ 的模型叫做**对位模型 (alignment model)**

# 统计翻译基本原理

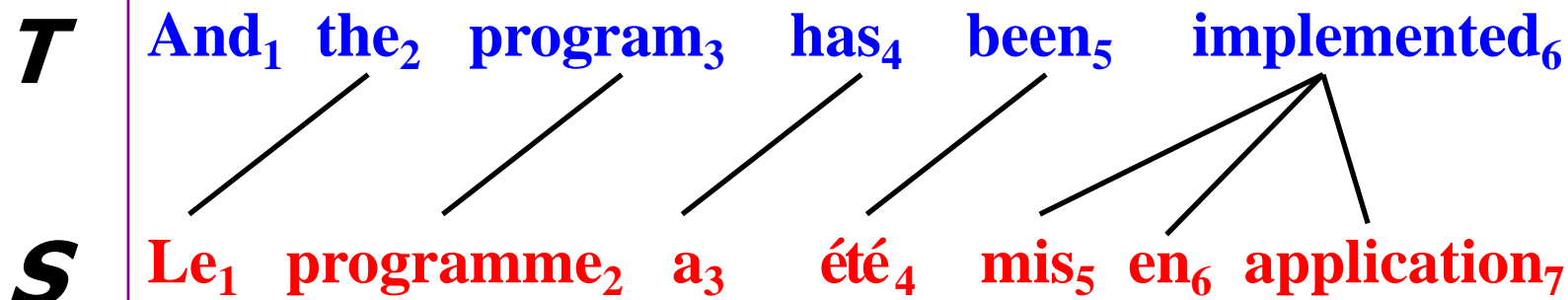
将对位模型  $A$  视为隐含变量，则：

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

- 源语言句子  $S = s_1^m = s_1 s_2 \cdots s_m$  有  $m$  个单词
- 目标语言句子  $T = t_1^n = t_1 t_2 \cdots t_n$  有  $n$  个单词
- 每一种对位序列表示成：

$$A = a_1^m = a_1 a_2 \cdots a_m \quad a_j \in [0, 1, \cdots, n]$$

# 统计翻译基本原理



$$m = 7 \quad n = 6$$

$$a_1 = 2, a_2 = 3, a_3 = 4, a_4 = 5, a_5 = 6, a_6 = 6, a_7 = 6$$

$$A = a_1^m = (2, 3, 4, 5, 6, 6, 6)$$

# 统计翻译基本原理

## □ 翻译概率 $p(S|T)$ 的计算

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

→  $P(S, A|T)$  ???

$$P(S, A|T) = p(m|T) \times P(A|T, m) \times P(S|T, A, m)$$

对位模型

词汇翻译模型

# 统计翻译基本原理

例句：我们一定要不忘初心


语言模型    翻译模型

$P$ (We forget must our heart)	0.001	0.041
$P$ (We forget must not our original heart)	0.012	0.052
$P$ (We not must not forget our original heart)	0.002	0.051
$P$ (We must forget our original intention)	0.049	0.059
$P$ (We must not forget our original intention)	0.051	0.071
$P$ (We must remember our original intention)	0.045	0.069

# 统计翻译基本原理

$$\begin{aligned} P(S, A|T) &= p(m|T) \times P(A|T, m) \times P(S|T, A, m) \\ &= p(m|T) \prod_{j=1}^m p(a_j | \alpha_1^{j-1}, s_1^{j-1}, m, T) \times p(s_j | \alpha_1^j, s_1^{j-1}, m, T) \end{aligned}$$

# 统计翻译基本原理

$$\begin{aligned} P(S, A|T) &= p(m|T) \times P(A|T, m) \times P(S|T, A, m) \\ &= p(m|T) \prod_{j=1}^m p(a_j | \alpha_1^{j-1}, s_1^{j-1}, m, T) \times p(s_j | \alpha_1^j, s_1^{j-1}, m, T) \end{aligned}$$
A red box highlights the term  $P(A|T, m)$  in the first line. An arrow points from this box to another red box in the second line, which highlights the product term  $p(a_j | \alpha_1^{j-1}, s_1^{j-1}, m, T)$ . This illustrates how the probability of the entire alignment  $A$  is decomposed into the product of probabilities for each word in the alignment.



# 统计翻译基本原理

$$\begin{aligned} P(S, A|T) &= p(m|T) \times P(A|T, m) \times P(S|T, A, m) \\ &= p(m|T) \prod_{j=1}^m p(a_j | \alpha_1^{j-1}, s_1^{j-1}, m, T) \times p(s_j | \alpha_1^j, s_1^{j-1}, m, T) \end{aligned}$$

# 统计翻译基本原理

$$\begin{aligned} P(S, A|T) &= p(m|T) \times P(A|T, m) \times P(S|T, A, m) \\ &= p(m|T) \prod_{j=1}^m p(a_j | a_1^{j-1}, s_1^{j-1}, m, T) \times p(s_j | a_1^j, s_1^{j-1}, m, T) \end{aligned}$$

基于上式，IBM 的研究人员通过采用不同的假设条件得到了5个翻译模型，分别称作 IBM 翻译模型1、2、3、4 和 5

# 12.2: IBM翻译模型1

# IBM 翻译模型1

$$p(m|T) \prod_{j=1}^m p(a_j | a_1^{j-1}, s_1^{j-1}, m, T) \times p(s_j | a_1^j, s_1^{j-1}, m, T)$$

## 翻译模型1：

(1) 假设  $\varepsilon \equiv p(m|T)$  是一个较小的常量；

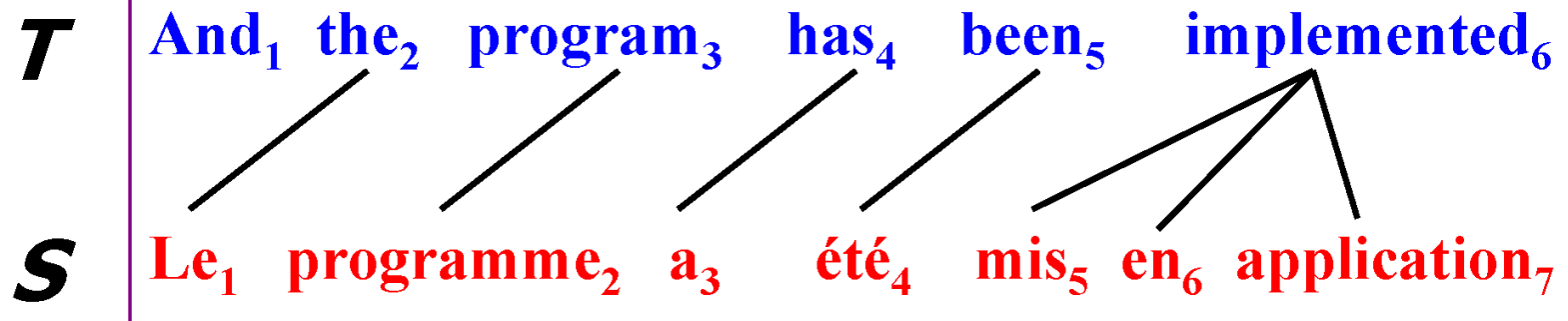
(2) 假设  $a_j \sim \text{uniform}(0, 1, 2, \dots, n)$   $p(a_j | a_1^{j-1}, s_1^{j-1}, m, T) = \frac{1}{n+1}$

(3) 假设  $s_j \sim \text{Categorical}(\theta_{t_{a_j}})$   $p(s_j | a_1^j, s_1^{j-1}, m, T) = p(s_j | t_{a_j})$

# IBM 翻译模型1

$$\begin{aligned} P(S, A|T) &= p(m|T) \prod_{j=1}^m p(a_j | a_1^{j-1}, s_1^{j-1}, m, T) \times p(s_j | a_1^j, s_1^{j-1}, m, T) \\ &= \varepsilon \prod_{j=1}^m \frac{1}{n+1} \times p(s_j | t_{a_j}) \\ &= \frac{\varepsilon}{(n+1)^m} \prod_{j=1}^m p(s_j | t_{a_j}) \end{aligned}$$

# IBM 翻译模型1



$$P(S, A|T) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^m p(s_j | t_{a_j})$$

$$\frac{\varepsilon}{(6+1)^7} \times [p(Le|the) \times \dots \times p(application|implemented)]$$

# IBM 翻译模型1

$$P(S, A|T) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^m p(s_j|t_{a_j})$$

$$\begin{aligned} P(S|T) &= \sum_A P(S, A|T) = \sum_A \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^m p(s_j|t_{a_j}) \\ &= \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^n \cdots \sum_{a_m=0}^n \prod_{j=1}^m p(s_j|t_{a_j}) \end{aligned}$$

如何训练？

# IBM 翻译模型1

$$\operatorname{argmax} P(S|T)$$

$$\text{w.r.t. } \sum_s p(s|t) = 1$$



$$h(p, \lambda) = P(S|T) - \sum_t \lambda_t \left( \sum_s p(s|t) - 1 \right)$$

$$= \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^n \cdots \sum_{a_m=0}^n \prod_{j=1}^m p(s_j | t_{a_j}) - \sum_t \lambda_t \left( \sum_s p(s|t) - 1 \right)$$



# IBM 翻译模型1

$$h(p, \lambda) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^n \cdots \sum_{a_m=0}^n \prod_{j=1}^m p(s_j | t_{a_j}) - \sum_t \lambda_t \left( \sum_s p(s|t) - 1 \right)$$

$$\frac{\partial h(p, \lambda)}{\partial p(s|t)} = 0$$



$$\frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^n \cdots \sum_{a_m=0}^n \sum_{j=1}^m \delta(s = s_j) \delta(t = t_{a_j}) \frac{1}{p(s|t)} \prod_{k=1}^m p(s_k | t_{a_k}) - \lambda_t = 0$$

# IBM 翻译模型1

$$\frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^n \cdots \sum_{a_m=0}^n \sum_{j=1}^m \delta(s = s_j) \delta(t = t_{a_j}) \frac{1}{p(s|t)} \prod_{k=1}^m p(s_k | t_{a_k})$$

$-\lambda_t = 0$



$$p(s|t) =$$

$$\frac{1}{\lambda_t} \times \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^n \cdots \sum_{a_m=0}^n \sum_{j=1}^m \delta(s = s_j) \delta(t = t_{a_j}) \prod_{k=1}^m p(s_k | t_{a_k})$$

$\delta$  : 克罗耐克系数

# IBM 翻译模型1

忽略详细的数学推导，IBM 翻译模型1表示为如下等式：

$$\begin{aligned} P(S|T) &= \frac{\varepsilon}{(n+1)^m} \sum_{a_1=0}^n \cdots \sum_{a_m=0}^n \prod_{j=1}^m p(s_j | t_{a_j}) \\ &= \frac{\varepsilon}{(n+1)^m} \prod_{j=1}^m \sum_{a_j=0}^n p(s_j | t_{a_j}) \end{aligned}$$

# IBM 翻译模型1

**e** And<sub>1</sub> the<sub>2</sub> program<sub>3</sub> has<sub>4</sub> been<sub>5</sub> implemented<sub>6</sub>

**f** f<sub>1</sub> f<sub>2</sub> f<sub>3</sub> f<sub>4</sub> f<sub>5</sub> f<sub>6</sub> f<sub>7</sub>

$\varepsilon \equiv p(m|T)$

**e** And<sub>1</sub> the<sub>2</sub> program<sub>3</sub> has<sub>4</sub> been<sub>5</sub> implemented<sub>6</sub>

**f** f<sub>1</sub> f<sub>2</sub> f<sub>3</sub> f<sub>4</sub> f<sub>5</sub> f<sub>6</sub> f<sub>7</sub>

$\frac{1}{n+1}$

**e** And<sub>1</sub> the<sub>2</sub> program<sub>3</sub> has<sub>4</sub> been<sub>5</sub> implemented<sub>6</sub>

**f** Le<sub>1</sub> programme<sub>2</sub> a<sub>3</sub> été<sub>4</sub> mis<sub>5</sub> en<sub>6</sub> application<sub>7</sub>

$p(s_j|t_{a_j})$

# 12.3: IBM翻译模型2

# IBM 翻译模型2

在IBM 模型2中，除了假定概率  $P(a_j | a_1^{j-1}, s_1^{j-1}, m, T)$  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  $n$  以外，另外两个假设与IBM模型1一样。

**引入了对位概率(alignment probabilities)的概念：**

$$a(a_j | j, m, n) = P(a_j | a_1^{j-1}, s_1^{j-1}, m, n)$$

# IBM 翻译模型2

对于每一个三元组  $(j, m, n)$ ，对位概率满足如下约束条件：

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

类似于IBM模型1的推导，得到模型2：

$$p(S | T) = \varepsilon \prod_{j=1}^m \sum_{i=0}^n p(s_j | t_i) \times a(i | j, m, n)$$

# IBM 翻译模型2

对于每一个三元组  $(j, m, n)$ ，对位概率满足如下约束条件：

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

类似于IBM模型1的推导，得到模型2：

$$p(S | T) = \varepsilon \prod_{j=1}^m \sum_{i=0}^n p(s_j | t_i) \times a(i | j, m, n)$$

$$\frac{1}{n+1}$$

如果对位概率设为常数，IBM 模型2退化为模型1，即模型1是模型2的特例。

$$p(S | T) = \frac{\varepsilon}{(n+1)^m} \prod_{j=1}^m \sum_{i=0}^n p(s_j | t_i)$$



# IBM 翻译模型2

**e**

And<sub>1</sub> the<sub>2</sub> program<sub>3</sub> has<sub>4</sub> been<sub>5</sub> implemented<sub>6</sub>

**f**

f<sub>1</sub> f<sub>2</sub> f<sub>3</sub> f<sub>4</sub> f<sub>5</sub> f<sub>6</sub> f<sub>7</sub>

$$\varepsilon \equiv p(m|T)$$

**e**

And<sub>1</sub> the<sub>2</sub> program<sub>3</sub> has<sub>4</sub> been<sub>5</sub> implemented<sub>6</sub>

**f**

f<sub>1</sub> f<sub>2</sub> f<sub>3</sub> f<sub>4</sub> f<sub>5</sub> f<sub>6</sub> f<sub>7</sub>

$$p(a_j|j, m, n)$$

**e**

And<sub>1</sub> the<sub>2</sub> program<sub>3</sub> has<sub>4</sub> been<sub>5</sub> implemented<sub>6</sub>

**f**

Le<sub>1</sub> programme<sub>2</sub> a<sub>3</sub> été<sub>4</sub> mis<sub>5</sub> en<sub>6</sub> application<sub>7</sub>

$$p(s_j|t_{a_j})$$

# 12.3: 基于短语的翻译模型

# 基于单词有哪些问题

## □ 基于单词的翻译模型有哪些优点？

比较符合人的思维、简单直接、易于实现

单词翻译表	P
我 → I	0.6
绿 → green	0.3
喜欢 → like	0.9
茶 → tea	0.8

**s** = 我    喜欢    绿    茶

# 基于单词有哪些问题

## □ 基于单词的翻译模型有哪些优点？

比较符合人的思维、简单直接、易于实现

单词翻译表	P
我 → I	0.6
绿 → green	0.3
喜欢 → like	0.9
茶 → tea	0.8

**s** = 我      喜欢      绿      茶  
         |      /      /      |  
**t** = I      like      green      tea

# 基于单词有哪些问题

## □ 基于单词的翻译模型有哪些不足？

- 需要定义词是什么
- 独立性假设：单词之间相对独立，没有考虑搭配
- 调序：较弱的调序建模
- ...

# 基于单词有哪些问题

## □ 基于单词的翻译模型有哪些不足？

- 需要定义词是什么
- 独立性假设：单词之间相对独立，没有考虑搭配
- 调序：较弱的调序建模
- ...

单词翻译表	P
我 → I	0.6
喜欢 → like	0.3
红 → red	0.9
红 → black	0.8
茶 → tea	0.6

**s** = 我      喜欢      红      茶

# 基于单词有哪些问题

## □ 基于单词的翻译模型有哪些不足？

- 需要定义词是什么
- 独立性假设：单词之间相对独立，没有考虑搭配
- 调序：较弱的调序建模
- ...

单词翻译表	P
我 → I	0.6
喜欢 → like	0.3
红 → red	0.9
红 → black	0.8
茶 → tea	0.6

**s** = 我      喜欢      红      茶  
         |      /      /      |  
**t** = I      like      red      tea

# 基于单词有哪些问题

## □ 基于单词的翻译模型有哪些不足？

- 需要定义词是什么
- 独立性假设：单词之间相对独立，没有考虑搭配
- 调序：较弱的调序建模
- ...

单词翻译表	P
我 → I	0.6
喜欢 → like	0.3
红 → red	0.9
红 → black	0.8
茶 → tea	0.6

**s =** 我      喜欢      红      茶  
         |      /      /      |  
**t =** I      like      red      tea

“红茶” 为一种搭配，  
应该翻译为 “black tea”



# 引入更大的翻译单元

## □ 简单的单词翻译似乎不行？

单词翻译表	P
我 → I	0.6
喜欢 → like	0.3
红 → red	0.9
红 → black	0.8
茶 → tea	0.6

**s** = 我      喜欢      红      茶  
         |      /      /      |  
**t** = I      like      red      tea



# 引入更大的翻译单元

## □ 简单的单词翻译似乎不行-引入更大的翻译单元

单词翻译表	P
我 → I	0.6
喜欢 → like	0.3
红 → red	0.8
红 → black	0.1
茶 → tea	0.8
我 喜 欢 → I like	0.3
我 喜 欢 → I liked	0.2
绿 茶 → green tea	0.5
绿 茶 → the green tea	0.1
红 茶 → black tea	0.6
...	

**s** = 我      喜 欢      红      茶  
         |      /      /      |  
**t** = I      like      red      tea



# 引入更大的翻译单元

## □ 简单的单词翻译似乎不行-引入更大的翻译单元

单词翻译表	P
我 → I	0.6
喜欢 → like	0.3
红 → red	0.8
红 → black	0.1
茶 → tea	0.8
我 喜 欢 → I like	0.3
我 喜 欢 → I liked	0.2
绿 茶 → green tea	0.5
绿 茶 → the green tea	0.1
红 茶 → black tea	0.6
...	

**s =** 我      喜欢      红      茶  
         |      /      /      |  
**t =**    I    like    red    tea

No

**s =** 我      喜欢      红      茶  
         |      /      /      /  
**t =**    I    like    black    tea

Yes

# 引入更大的翻译单元

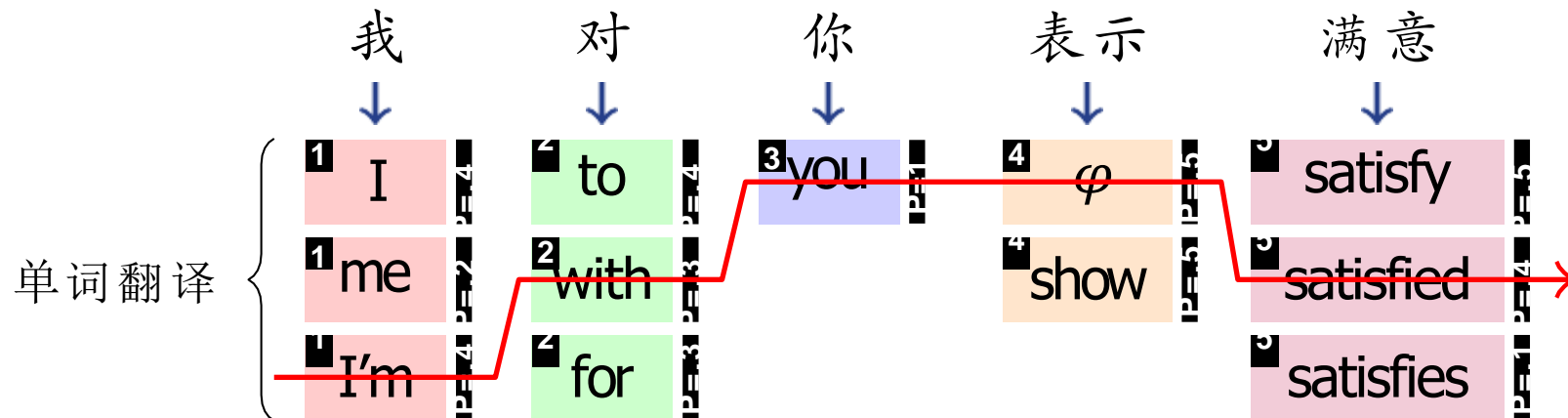
## □优点:

- 翻译时候可以考虑更大范围的上下文信息 比如:  
“红茶”中的“红”如果和“茶”搭配 ...
- 更好的局部调序, 比如: 短语中有“的”字 → ... of ... 结构
- 更大范围的目标语连续词串, 有利于  $n$ -gram 语言模型选择译文

# 基于“短语”的机器翻译

- 对每个单词及连续词串的翻译进行（任意）组合

待翻译句子(已经分词):

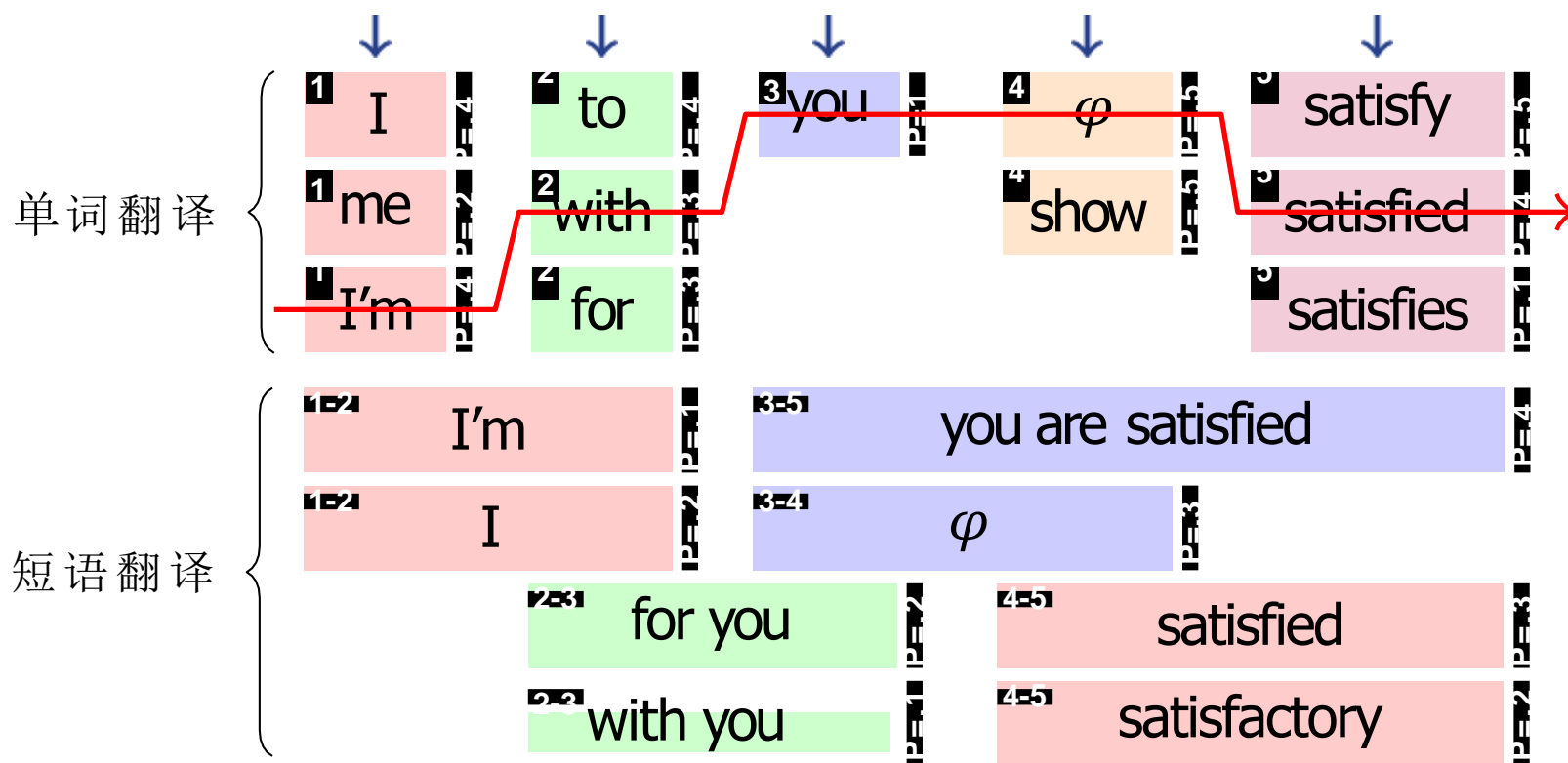


# 基于“短语”的机器翻译

- 对每个单词及连续词串的翻译进行（任意）组合

待翻译句子(已经分词):

我                  对                  你                  表示                  满意

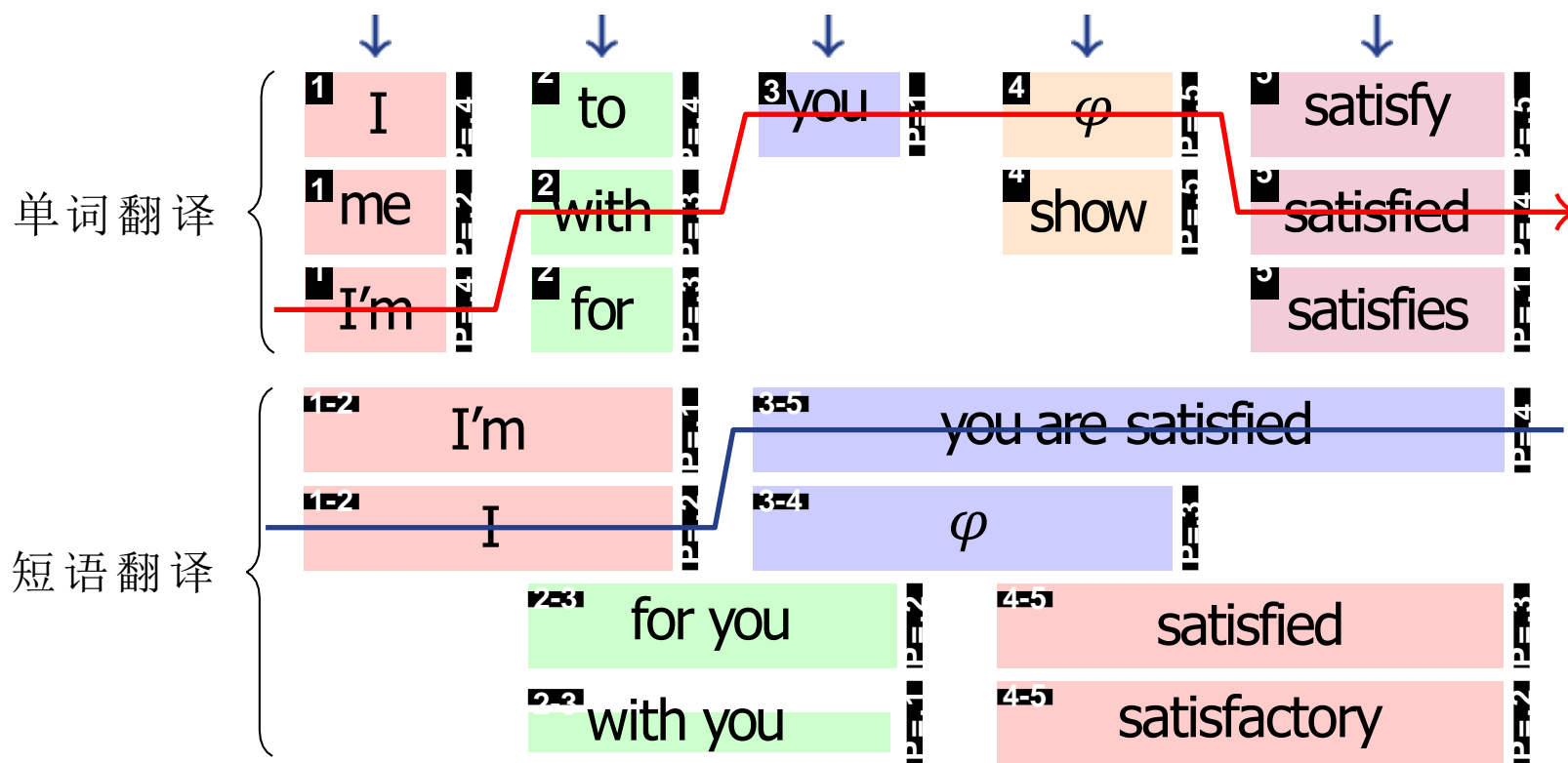


# 基于“短语”的机器翻译

- 对每个单词及连续词串的翻译进行（任意）组合

待翻译句子(已经分词):

我                  对                  你                  表示                  满意



翻译路径（仅含有单词）：



翻译路径（含有短语）：



# 基于“短语”的机器翻译

- 对每个单词及连续词串的翻译进行（任意）组合

待翻译句子(已经分词):

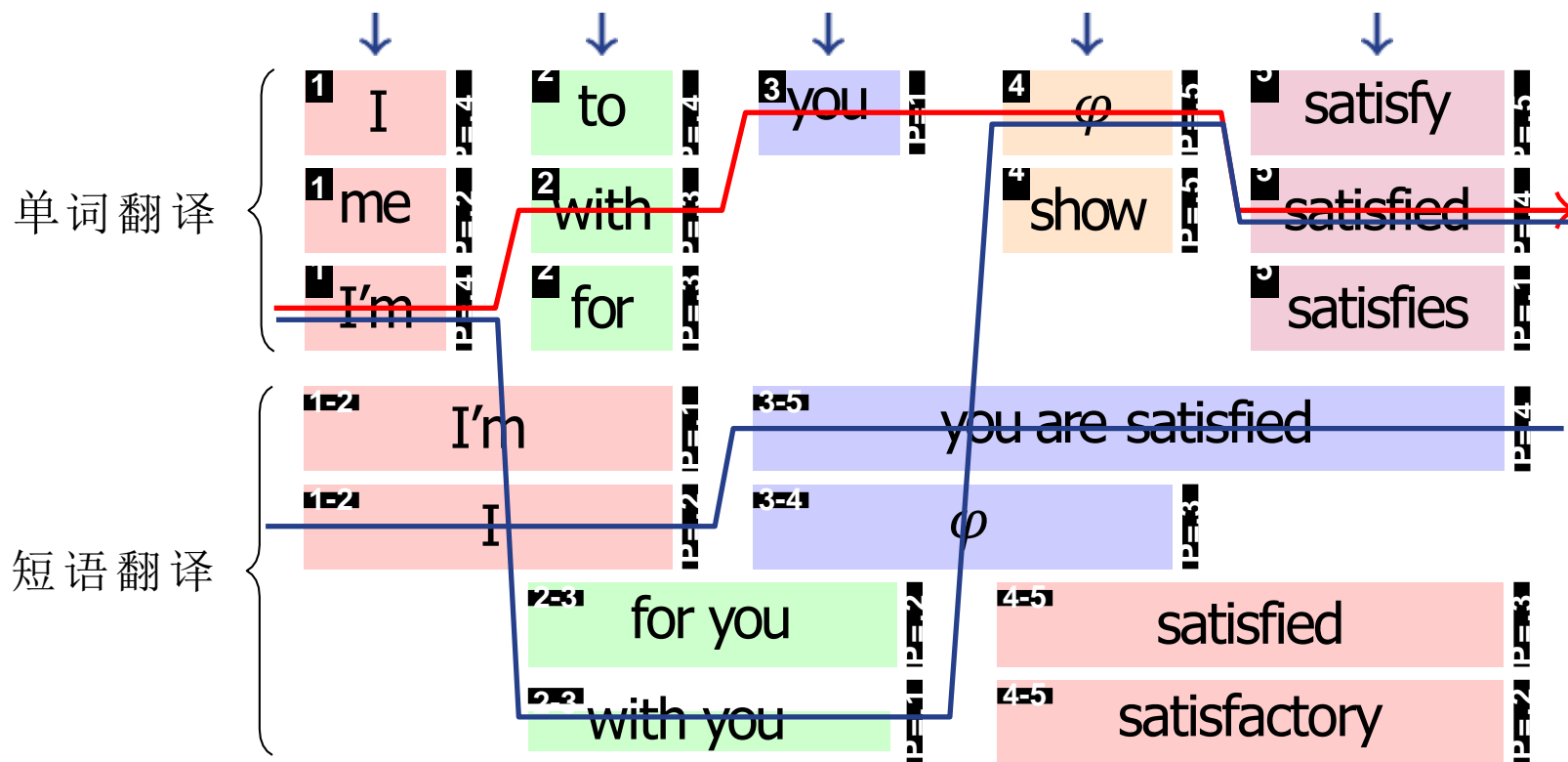
我

对

你

表示

满意



翻译路径（仅含有单词）：



翻译路径（含有短语）：





# 使用短语就够了？

- 短语是具有完整意思的连续词串，可以捕捉更多的上下文信息
  - 不过过大的短语会造成数据稀疏、长距离依赖等问题
  - 而且单纯的词串也缺乏句法功能表示能力

进口 在过去的五到十年间 有了大幅度下降

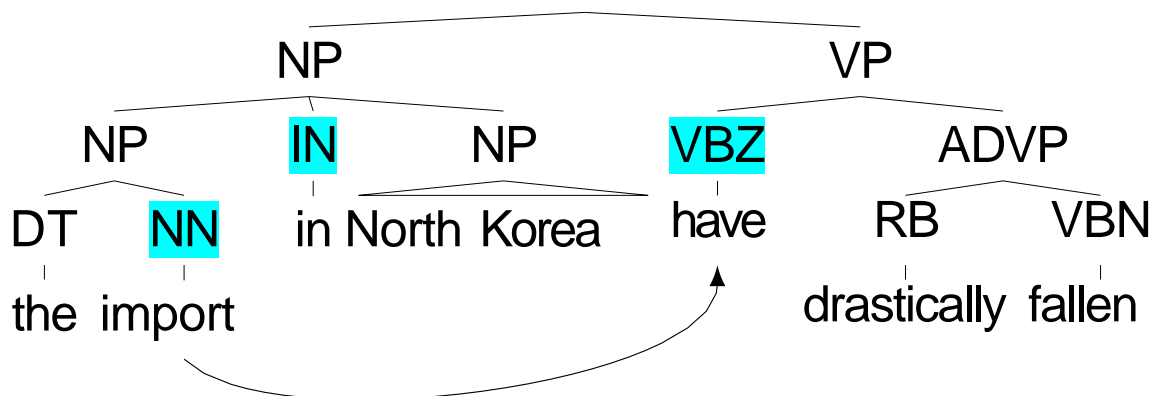


the imports drastically fell in the past five to ten years



# 使用短语就够了？

- 另一种方式是考虑句子的句法结构，这样更容易描述句子的层次结构和长距离依赖关系



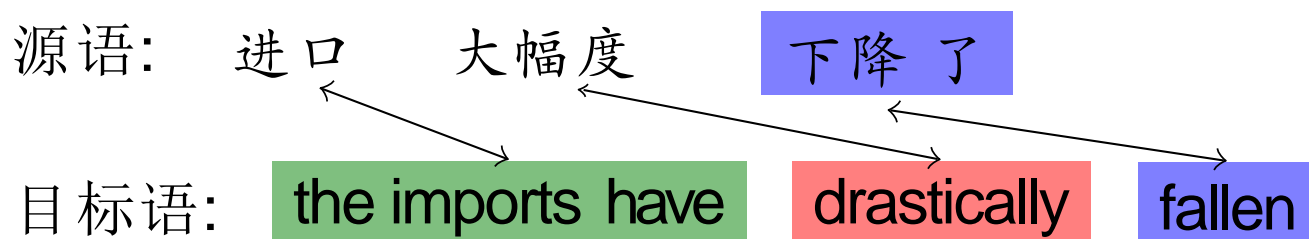
# 引入句法信息

- 依存句法树是句子的更高层次的抽象，相比短语句法树具有更加丰富的句法功能标记，对语言结构的转换很有帮助
  - 更容易捕捉翻译中的远距离调序
  - 使用句法更容易对大范围的上下文建模

# 何为短语？

□ 句对可以用短语对的组合进行表示，比如下图的例子包含三个短语翻译：

- 进口 ↔ the imports have
- 大幅度 ↔ drastically
- 下降了 ↔ fallen



# 何为短语？

□ 显然上图中的短语并不是语言学上的短语这里有：

## 定义 - 短语

对于一个句子  $\mathbf{w} = w_1 \dots w_n$  任意子串  $w_i \dots w_j (i \leq j, 0 \leq i, j \leq n)$  都是句子  $\mathbf{w}$  的一个 **短语**

○  $n$  个词构成的句子可以有  $\frac{n(n+1)}{2}$  个短语

# 何为短语？

□ 进一步，可以定义

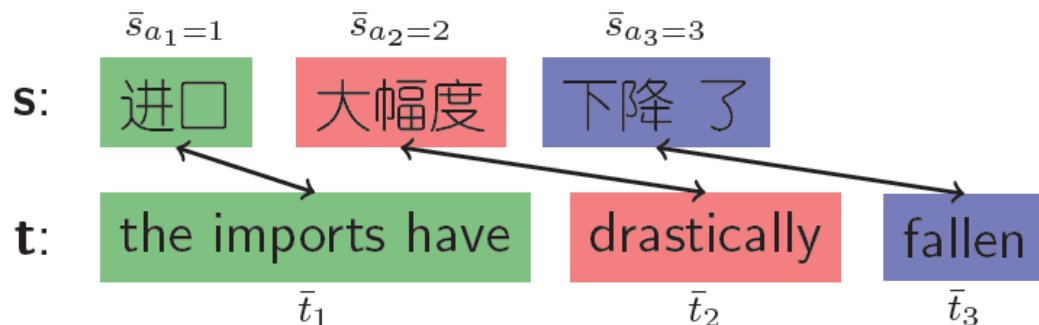
## 定义 - 句子的短语切分

对于一个句子  $\mathbf{W} = w_1 \dots w_n$ ，可以被切分为  $m$  个子串，则称  $\mathbf{W}$  由  $m$  个短语组成，记为  $\mathbf{W} = p_1 \dots p_m$ ，其中  $p_i$  是  $\mathbf{W}$  的一个短语， $p_1 \dots p_m$  也被称作句子  $\mathbf{W}$  的一个短语切分。

# 基于短语的翻译推导

## 定义 - 基于短语的翻译推导

对于源语和目标语句对 $(\mathbf{s}, \mathbf{t})$ ，分别有短语切分 $\{\bar{s}_i\}$ 和 $\{\bar{t}_j\}$ ，且 $\{\bar{s}_i\}$ 和 $\{\bar{t}_j\}$ 之间存在一一对应的关系。令 $\{\bar{a}_j\}$ 表示 $\{\bar{t}_j\}$ 中每个短语对应到源语言短语的编号，则称短语对 $\{(\bar{s}_{\bar{a}_j}, \bar{t}_j)\}$ 构成了 $\mathbf{s}$ 到 $\mathbf{t}$ 的基于短语的翻译推导(简称推导)，记为 $d(\{(\bar{s}_{\bar{a}_j}, \bar{t}_j)\}, \mathbf{s}, \mathbf{t})$ (简记为 $d(\{(\bar{s}_{\bar{a}_j}, \bar{t}_j)\})$ 或 $d$ )。



- $\{(\bar{s}_{\bar{a}_j}, \bar{t}_j)\}$ 构成了 $(\mathbf{s}, \mathbf{t})$ 的一个基于短语的翻译推导
- 需要在建模中描述的两个问题：
  - ▶  $\bar{s}_{\bar{a}_j}$ 是如何被翻译成 $\bar{t}_j$ 的?
  - ▶ 翻译的顺序是如何决定的，即如何得到 $\{\bar{a}_j\}$ ?

# 12.3: 译文评估方法



# 译文评估方法

## ◆常用的评测指标

➤ 主观评测：(1)流畅度；(2)充分性；(3) 语义保持性。

流畅性	
5	完美的英语表达 (Flawless English)
4	较好的英语表达 (Good English)
3	非母语的英语表达 (Non-native English)
2	不流畅的英语表达 (Disfluent English)
1	无法理解的英语表达 (Incomprehensible)

# 译文评估方法

## 充分性

充分性	
5	全部信息都已充分表达了出来 (All information)
4	绝大部分信息已经表达了出来 (Most information)
3	很多信息被表达了出来 (Much information)
2	表达了少量信息 (Little information)
1	没有表达任何信息 (None)

# 译文评估方法

语义保持性 (meaning maintenance)	
0	意思完全相反 (Total different meaning)
1	部分语义相同，但引入了误导信息 (Partially the same meaning but misleading information is introduced)
2	部分语义相同，没有引入新的信息 (Partially the same meaning and no new information)
3	意思几乎相同 (Almost the same meaning)
4	意思完全相同 (Exactly the same meaning)

# 译文评估方法

源语言句子: **This boy is very lovely.**

译文1: 这个小孩很可爱。

译文2: 这个桌子很可爱。

译文3: 这个小孩不可爱。

# 译文评估方法

## □ BLEU评价方法 [Papineni, 2002]

— BiLingual Evaluation Understudy, IBM

- 基本思想：将机器翻译产生的候选译文与人翻译的参考译文相比较，越接近，候选译文的正确率越高；
- 实现方法：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 $n$ 元词数目，得到评测结果。

# 译文评估方法

例如：

- ❑ 系统译文： the the the the the the the.
- ❑ 参考译文1： The cat is on the mat.
- ❑ 参考译文2： There is a cat on the mat.

按照上述计算方法，如果 $n$ 取1的话，该候选译文可以得到7/7的打分，但显然这种翻译结果几乎没有任何意义。

# 译文评估方法

修正的计算一元语法精确度的方法：针对某个待评测的系统译文句子，首先统计每个单词在所有参考译文中出现次数的最大值  $Max\_Ref\_Count$ ，然后，统计该单词在系统译文中出现的总次数  $Count$ ，取  $Count$  和  $Max\_Ref\_Count$  两者中小的一个，即

$$Count_{clip} = \min(Count, Max\_Ref\_Count)$$

这样保证了每个系统译文中的单词计数不会超过该词在某个参考译文中出现次数的最大值。

# 译文评估方法

把系统译文中所有单词的 $Count_{clip}$ 值累加起来，得到 $Total\_Count_{clip}$ ，即待评测的系统译文中出现在参考译文中的单词个数，最后，用 $Total\_Count_{clip}$ 除以系统译文中全部单词的个数。



# 译文评估方法

在上面的例子中，系统译文中的单词the在参考译文1中出现的次数最多， $Max\_Ref\_Count=2$ ，而the在系统译文中出现的次数为7，即 $Count=7$ ，因此， $Count_{clip}=\min(7, 2)=2$ 。候选译文中全部单词的个数等于7，因此，该例中修正后的一元语法精确度为 $2/7$ 。请再看下面的例子：

# 译文评估方法

- 对于含多个句子的测试文本，以句子为单位分别计算  $n$ -gram 的匹配情况，然后，累计所有翻译句子修正后的  $n$ -gram 计数，及测试集的所有  $n$ -gram 计数，二者相除，得到修正后的精确度记分：

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-gram} \in C} Count_{clip}(n\text{-gram})}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-gram}' \in C'} Count(n\text{-gram}')}$$

# 译文评估方法

- 考虑到在修正的  $n$  元语法精度计算中，随着  $n$  值的增大精度值几乎成指数级下降，因此，BLEU方法中采用了修正的  $n$  元语法精度的对数加权平均值，相当于对修正的精度值进行几何平均， $n$  值最大为4。
- 另外，考虑到句子的长度对上述BLEU评分也有一定的影响，例如，如果一个机器翻译系统只翻译最可靠的词汇，译文句子就可能比较短，按上述方法计算出的精度值就会较高。因此，需要进一步考虑候选译文的句子长度对计算评分的影响。

# 译文评估方法

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

长度过短句子的  
惩罚因子

$$w_n = 1/N$$

最大语法的阶  
数，实际取4。

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。

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

$c$ 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。

BLEU分值越高表示译文质量越好，分值越小译文质量越差

## More Resources

<http://statmt.org/>

# 思考

- 统计机器翻译的难点在什么地方？
- 翻译模型的关键是什么？

# 致谢

- 中科院自动化所《自然语言处理》，宗成庆；
- 东北大学《机器翻译：统计建模与深度学习方法》，肖桐，朱靖波；

# Thank you!

权小军 中山大学数据科学与计算机学院