最小编辑距离算法
Minimum Edit Distance

詹卫东

北京大学中文系

编辑距离

编辑前字符串 s

编辑后字符串 t

编辑操作p：插入、删除、替换

“编辑距离”定义为
“编辑操作的次数”

源文：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。

计算机器译文
跟正确答案之
间的距离

编辑距离：6
删除次数（4次）： ~~与~~ ~~公司~~ ~~一~~ ~~颗~~
替换次数（2次）： 剧院 →剧团 星→ 明星

最小编辑距离

原始串　sot

目标串　stop

- 插入操作

  （insertCost权值）　：　1

- 删除操作

  （deleteCost权值）　：　1

- 替换操作

  （substituteCost）　：　2

sot　编辑操作1

→ stot　(1.插入，1分，累计1分)

→ stop　(2.替换，2分，累计3分)

编辑距离：3

sot　编辑操作2

→ stt　(1.替换，2分，累计2分)

→ sto　(2.替换，2分，累计4分)

→ stop　(3.插入，1分，累计5分)

编辑距离：5

最小编辑距离计算：动态规划

(1) D(0, 0) = 0 /* 空串变换成空串的编辑距离 */

(2) D(i, 0) = insertCost * i /* 空串变换成长度为 *i* 的串的编辑距离 */

(3) D(0, j) = deleteCost * j /* 长度为 *j* 的串变换成空串的编辑距离 */

(4)
$$D(i, j) = \min\begin{cases} D(i-1, j) + \text{insertCost}(\text{target}_i) \\ D(i-1, j-1) + \text{substituteCost}(\text{source}_j, \text{target}_i) \\ D(i, j-1) + \text{deleteCost}(\text{source}_j) \end{cases}$$

insertCost = 1

deleteCost = 1

$$\text{substituteCost}\begin{cases} = 0 & \text{if target[i] = source[j]} \\ = 2 & \text{otherwise} \end{cases}$$

i： 目标串字符位置序号

j： 原始串字符位置序号

D(*i*,*j*)： 从 *j* 变化到 *i* 的距离值

source_j： j 位置的字符

target_i： i 位置的字符

最小编辑距离算法描述

```
function Min-Edit_Distance (target, source)
 n = length(target);
 m = length(source);
 create distance matrix d[n, m];
 d[0,0]=0;
 d[0,1]=1,… d[0,m]=m;
 d[1,0]=1,…d[n,0]=n;
 for each i from 1 to n do
     for each j from 1 to m do
             d[i, j] = min( d[i-1, j]    +  insertCost($target_i$),
                            d[i-1, j-1] +  substituteCost($source_j$, $target_i$),
                            d[i, j-1]    +  deleteCost($source_j$));
return d[n, m];
```

最小编辑距离计算示例

source :   s o t

target :   s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t				
2	o				
1	s				
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=0 j=0

d[0,0] = 0;
d[0,1] = 1; … ; d[0,m] = m;
d[1,0] = 1; … ; d[n,0] = n;

最小编辑距离计算示例

source :　　s o t

target :　　s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t				
2	o				
1	s	**0**			
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=1 j=1

$$d[1,1] = \min \left\{ \begin{array}{l} d[0,1]+\text{insert}(t[1]) = 2 \\ d[0,0]+\text{substitute}(s[1],t[1]) = 0 \\ d[1,0]+\text{delete}(s[1]) = 2 \end{array} \right\} = 0$$

最小编辑距离计算示例

source :   s o t

target :   s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t				
2	o	1			
1	s	**0**			
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=1 j=2

$$d[1,2] = \min \left\{ \begin{array}{l} d[0,2]+insert(t[1]) = 3 \\ d[0,1]+substitute(s[2],t[1]) = 3 \\ d[1,1]+delete(s[2]) = 1 \end{array} \right\} = 1$$

最小编辑距离计算示例

source : s o t

target : s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2			
2	o	**1**			
1	s	**0**			
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=1 j=3

$$d[1,3] = \min \left\{ \begin{array}{l} d[0,3]+\text{insert}(t[1]) = 4 \\ d[0,2]+\text{substitute}(s[3],t[1]) = 4 \\ d[1,2]+\text{delete}(s[3]) = 2 \end{array} \right\} = 2$$

最小编辑距离计算示例

source :   s o t

target :   s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2			
2	o	**1**			
1	s	**0**	1		
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=2 j=1

$$d[2,1] = \min \left\{ \begin{array}{l} d[1,1]+\text{insert}(t[2]) = 1 \\ d[1,0]+\text{substitute}(s[1],t[2]) =3 \\ d[2,0]+\text{delete}(s[1]) = 3 \end{array} \right\} = 1$$

最小编辑距离计算示例

source :   s o t

target :   s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2			
2	o	**1**	2		
1	s	**0**	**1**		
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=2 j=2

$$d[2,2] = \min \left\{ \begin{array}{l} d[1,2]+\text{insert}(t[2]) = 2 \\ d[1,1]+\text{substitute}(s[2],t[2]) =2 \\ d[2,1]+\text{delete}(s[2]) = 2 \end{array} \right\} = 2$$

最小编辑距离计算示例

source :   s o t

target :   s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2	1		
2	o	**1**	2		
1	s	**0**	**1**		
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=2 j=3

$$d[2,3] = \min \left\{ \begin{array}{l} d[1,3]+\text{insert}(t[2])=3 \\ d[1,2]+\text{substitute}(s[3],t[2])=1 \\ d[2,2]+\text{delete}(s[3])=3 \end{array} \right\} = 1$$

最小编辑距离计算示例

source :   s o t

target :   s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2	1		
2	o	**1**	2		
1	s	**0**	**1**	2	
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=3 j=1

$$d[3,1] = \min \left\{ \begin{array}{l} d[2,1]+insert(t[3])=2 \\ d[2,0]+substitute(s[1],t[3])=4 \\ d[3,0]+delete(s[1])=4 \end{array} \right\} = 2$$

最小编辑距离计算示例

source : s o t

target : s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2	1		
2	o	**1**	2	1	
1	s	**0**	**1**	2	
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=3 j=2

$$d[3,2] = \min \left\{ \begin{array}{l} d[2,2]+\text{insert}(t[3])=3 \\ d[2,1]+\text{substitute}(s[2],t[3])=1 \\ d[3,1]+\text{delete}(s[2])=3 \end{array} \right\} = 1$$

最小编辑距离计算示例

source : s o t

target : s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2	**1**	2	
2	o	**1**	2	**1**	
1	s	**0**	**1**	2	
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=3 j=3

$$d[3,3] = \min \left\{ \begin{array}{l} d[2,3]+\text{insert}(t[3])=2 \\ d[2,2]+\text{substitute}(s[3],t[3])=4 \\ d[3,2]+\text{delete}(s[3])=2 \end{array} \right\} = 2$$

最小编辑距离计算示例

source : s o t

target : s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2	**1**	2	
2	o	**1**	2	**1**	
1	s	**0**	**1**	**2**	3
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=4 j=1

$$d[4,1] = \min \left\{ \begin{array}{l} d[3,1]+\text{insert}(t[4])=3 \\ d[3,0]+\text{substitute}(s[1],t[4])=5 \\ d[4,0]+\text{delete}(s[1])=5 \end{array} \right\} = 3$$

最小编辑距离计算示例

source : s o t

target : s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

3	t	2	**1**	2	
2	o	**1**	2	**1**	2
1	s	**0**	**1**	2	3
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=4 j=2

$$d[4,2] = \min \left\{ \begin{array}{l} d[3,2]+\text{insert}(t[4])=2 \\ d[3,1]+\text{substitute}(s[2],t[4])=4 \\ d[4,1]+\text{delete}(s[2])=4 \end{array} \right\} = 2$$

最小编辑距离计算示例

source : s o t

target : s t o p

n = length (target) = 4

m = length (source) = 3

Create matrix d [n, m];

source j

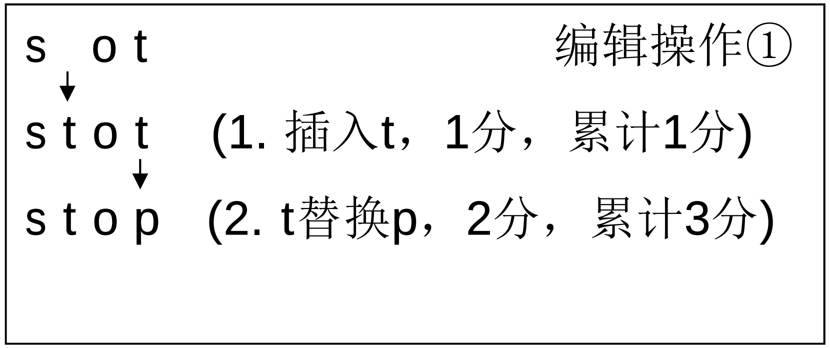
3	t	2	**1**	**2**	**3**
2	o	**1**	2	**1**	**2**
1	s	**0**	**1**	2	3
0	#	s	t	o	p
#	0	1	2	3	4

target i

i=4 j=3

$$d[4,3] = \min \left\{ \begin{array}{l} d[3,3]+\text{insert}(t[4])=3 \\ d[3,2]+\text{substitute}(s[3],t[4])=3 \\ d[4,2]+\text{delete}(s[3])=3 \end{array} \right\} = 3$$

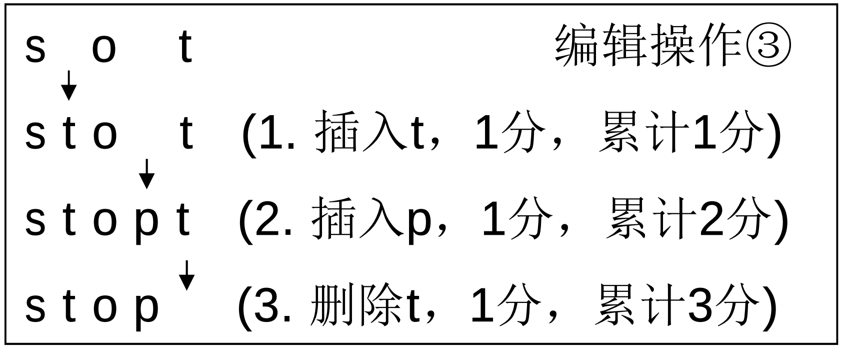
最小编辑距离计算示例



sot 编辑操作①

stot (1. 插入t，1分，累计1分)

stop (2. t替换p，2分，累计3分)

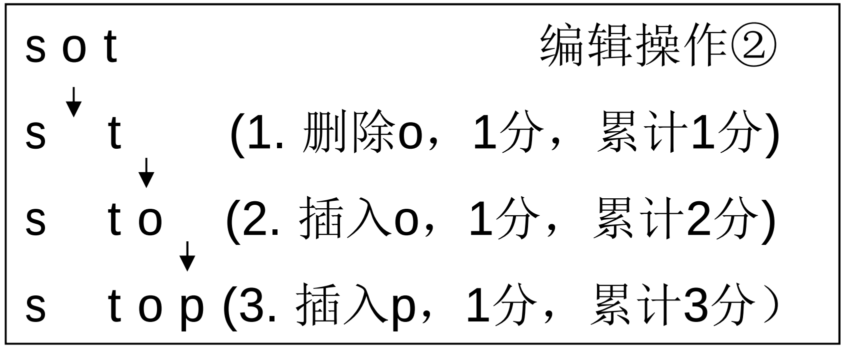


sot 编辑操作③

sto t (1. 插入t，1分，累计1分)

stopt (2. 插入p，1分，累计2分)

stop (3. 删除t，1分，累计3分)

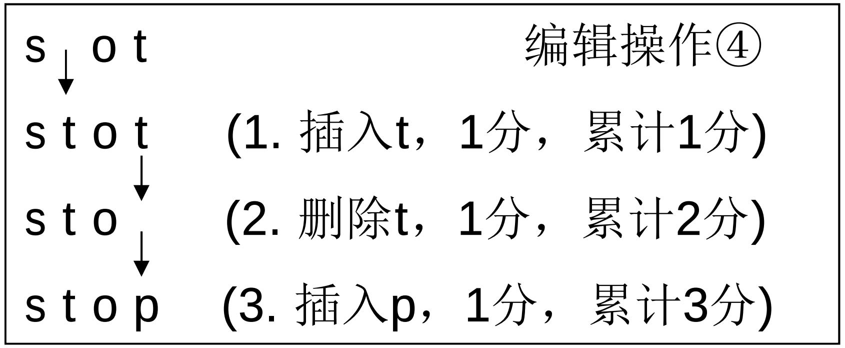


sot 编辑操作②

s t (1. 删除o，1分，累计1分)

s to (2. 插入o，1分，累计2分)

s top (3. 插入p，1分，累计3分）



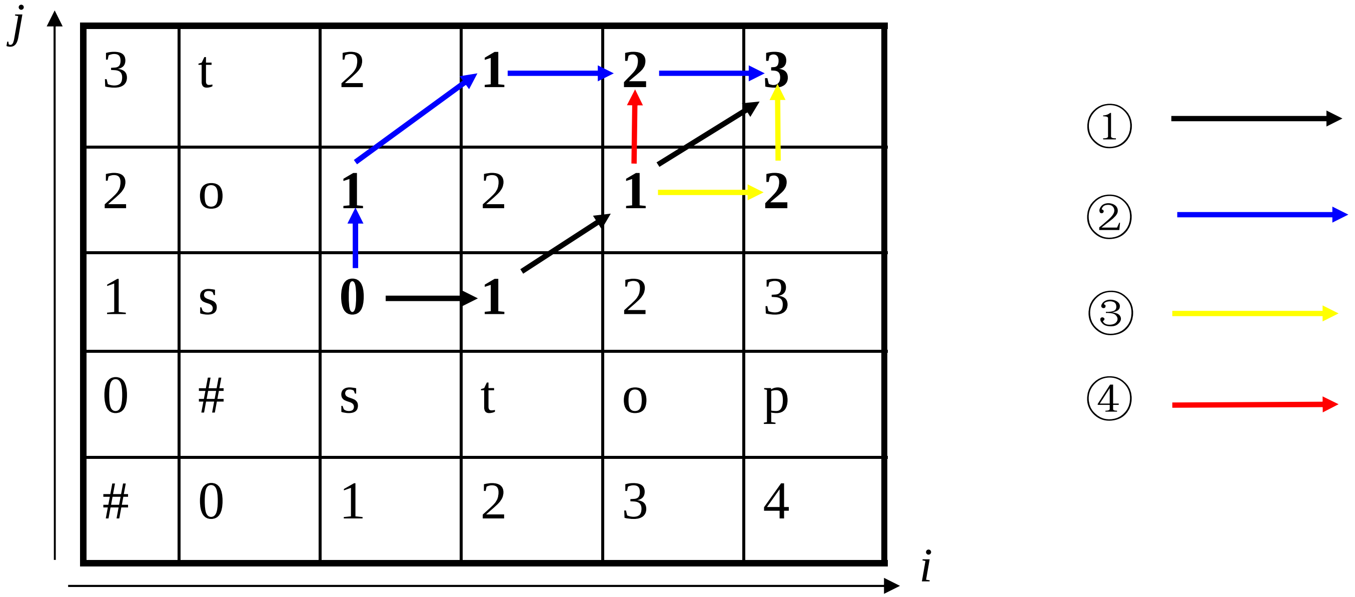
sot 编辑操作④

stot (1. 插入t，1分，累计1分)

sto (2. 删除t，1分，累计2分)

stop (3. 插入p，1分，累计3分)

最小编辑距离计算示例



3	t	2	**1**	**2**	**3**
2	o	**1**	2	**1**	**2**
1	s	**0**	**1**	2	3
0	#	s	t	o	p
#	0	1	2	3	4

① →

② →

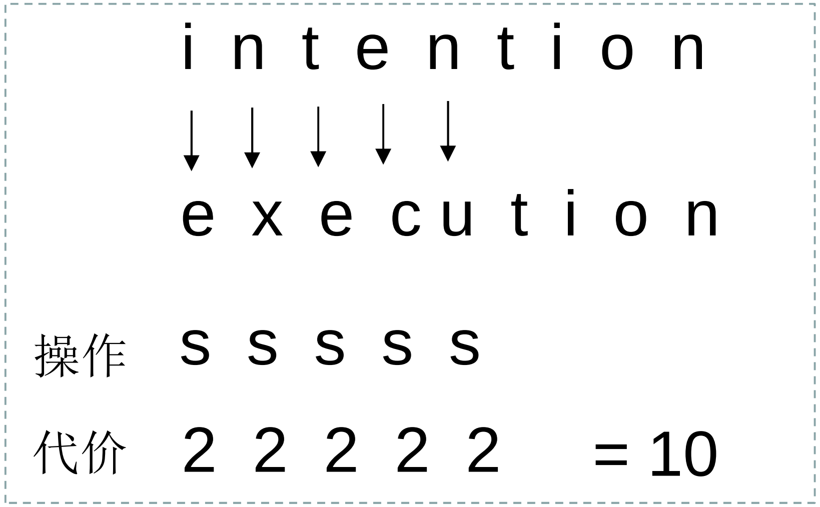
③ →

④ →

最小编辑距离计算练习

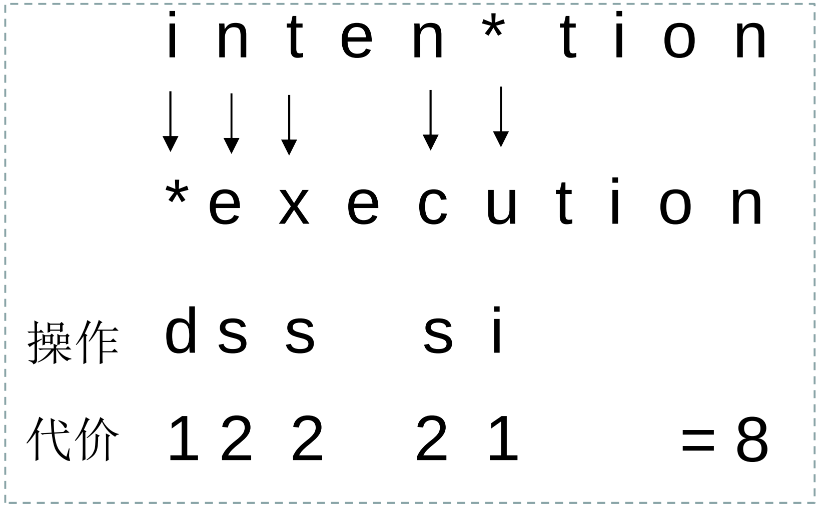
- intention → execution

编辑操作①



	i	n	t	e	n	t	i	o	n	
	↓	↓	↓	↓	↓					
	e	x	e	c	u	t	i	o	n	
操作	s	s	s	s	s					
代价	2	2	2	2	2					= 10

编辑操作②



	i	n	t	e	n	*	t	i	o	n	
	↓	↓	↓		↓	↓					
	*	e	x	e	c	u	t	i	o	n	
操作	d	s	s		s	i					
代价	1	2	2		2	1					= 8

s：替换操作　d：删除操作　i：插入操作

最小编辑距离计算练习

source

n	9									
o	8									
i	7									
t	6									
n	5									
e	4									
t	3									
n	2									
i	1									
#	0	1	2	3	4	5	6	7	8	9
	#	e	x	e	c	u	t	i	o	n

target

最小编辑距离计算练习

source

n	9	8	9	10	11	12	11	10	9	8
o	8	7	8	9	10	11	10	9	8	9
i	7	6	7	8	9	10	9	8	9	10
t	6	5	6	7	8	9	8	9	10	11
n	5	4	5	6	7	8	9	10	11	10
e	4	3	4	5	6	7	8	9	10	9
t	3	4	5	6	7	8	7	8	9	8
n	2	3	4	5	6	7	8	**7**	8	**7**
i	1	2	3	4	5	6	7	**6**	**7**	**8**
#	0	1	2	3	4	5	6	7	8	9
	#	e	x	e	c	u	t	i	o	n

target

参考文献

- Daniel Jurafsky & James H. Martin, 2000, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Chapter 5, section 5.6, pp153-156, Prentice-Hall Inc..