CMPT 413
Computational Linguistics

Anoop Sarkar

http://www.cs.sfu.ca/~anoop

Minimum Cost Edit Distance

- Edit a source string into a target string
- Each edit has a cost
- Find the minimum cost edit(s)

actress
↑ insert(s)
actres
↑ delete(t)
actrest
↑ insert(t)
acrest
↑ insert(a)
crest

```
a c t r e s _ s
  |   | | |
_ c _ r e s t _
```

```
a c t r e s s _
  |   | | |
_ c _ r e s _ t
```

```
a c t r e s s _
  |   | |   |
_ c _ r e _ s t
```

```
a c t r e s s
  |   | | |
_ c _ r e s t
```

minimum cost edit distance can be accomplished in multiple ways

Only 4 ways to edit source to target **for this pair**

Minimum Cost Edit Distance

```
a c t r e s _ s
  |     | | |
_ c _ r e s t _
```

target

source

actress
↑
actres
↑
actrest
↑
acrest
↑
crest

minimum cost edit distance can be accomplished in multiple ways

```
a c t r e s s _
  |     | | |
_ c _ r e s _ t
```

```
a c t r e s s _
  |     | |   |
_ c _ r e _ s t
```

```
a c t r e s s
  |     | | |
_ c _ r e s t
```

Only 4 ways to edit source to target **for this pair**

Levenshtein Distance

- Cost is fixed across characters
  - Insertion cost is 1
  - Deletion cost is 1
- Two different costs for substitutions
  - Substitution cost is 1 (transformation)
  - Substitution cost is 2 (one deletion + one insertion)

Левенштейн Владимир

Vladimir Levenshtein

What's the edit distance?

Minimum Cost Edit Distance

- An alignment between target and source

Find *D(n,m)* recursively

$t_1, t_2, \ldots, \ldots, t_n$

$s_1, s_2, \ldots, s_m$

$$D(i,j) = min \begin{cases} D(i-1,j) & +\text{cost}(t_i, \emptyset) \; \textit{insertion into target} \\ D(i-1,j-1) & +\text{cost}(t_i, s_j) \textit{substitution/identity} \\ D(i,j-1) & +\text{cost}(\emptyset, s_j) \textit{deletion from source} \end{cases}$$

$$D(0,0) = 0$$

$$D(i,0) = D(i-1,0) + \text{cost}(t_i, \emptyset)$$

$$D(0,j) = D(0,j-1) + \text{cost}(\emptyset, s_j)$$

```
Function MinEditDistance (target, source)

 n = length(target)
 m = length(source)
 Create matrix D of size (n+1,m+1)
 D[0,0] = 0

 for i = 1 to n
   D[i,0] = D[i-1,0] + insert-cost

 for j = 1 to m
   D[0,j] = D[0,j-1] + delete-cost

 for i = 1 to n
   for j = 1 to m
     D[i,j] = MIN(D[i-1,j] + insert-cost,
                  D[i-1,j-1] + subst/eq-cost,
                  D[i,j-1] + delete-cost)

 return D[n,m]
```

Consider two strings: target = $g_1a_2m_3b_4l_5e_6$

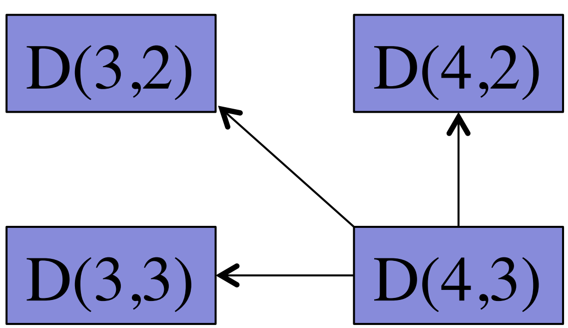
source= $g_1u_2m_3b_4o_5$

- We want to find D(6,5)
- We find this recursively using values of D(i,j) where $i \leq 6$ $j \leq 5$
- For example, consider how to compute D(4,3)

target = $g_1a_2m_3b_4$

source= $g_1u_2m_3$

- Case 1: SUBSTITUTE b_4 for m_3
- Use previously stored value for D(3,2)
- Cost($g_1a_2m_3$b and g_1u_2m) = D(3,2) + cost(b≈m)
- For substitution: D(i,j) = D(i-1,j-1) + cost(subst)



- Case 2: INSERT b_4
- Use previously stored value for D(3,3)
- Cost($g_1a_2m_3$b and $g_1u_2m_3$) = D(3,3) + cost(ins b)
- For substitution: D(i,j) = D(i-1,j) + cost(ins)

- Case 3: DELETE m_3
- Use previously stored value for D(4,2)
- Cost($g_1a_2m_3b_4$ and g_1u_2m) = D(4,2) + cost(del m)
- For substitution: D(i,j) = D(i-1,j-1) + cost(subst)

target

source

		g	a	m	b	l	e
	0	1	2	3	4	5	6
g	1	0 *e*	1	2	3	4	5
u	2	1	2 *s*	3	4	5	6
m	3	2	3	2 *e*	3	4	5
b	4	3	4	3	2 *e*	3 *i*	4
o	5	4	5	4	3	4	5 *s*

```
g a m b l e
|   | |
g u m b _ o
```

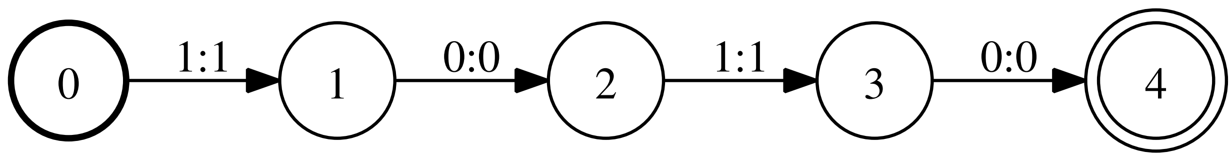
Edit Distance and FSTs

- Algorithm using a Finite-state transducer:
  - construct a finite-state transducer with all possible ways to transduce source into target
  - We do this transduction one char at a time
  - A transition x:x gets zero cost and a transition on ε:x (insertion) or x:ε (deletion) for any char x gets cost 1
  - Finding minimum cost edit distance == Finding the shortest path from start state to final state

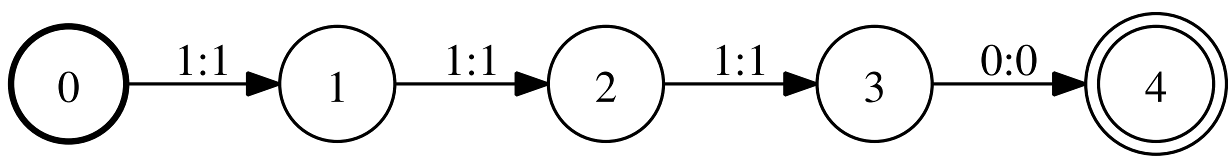
Edit Distance and FSTs

- Lets assume we want to edit source string 1010 into the target string 1110
- The alphabet is just 1 and 0

SOURCE



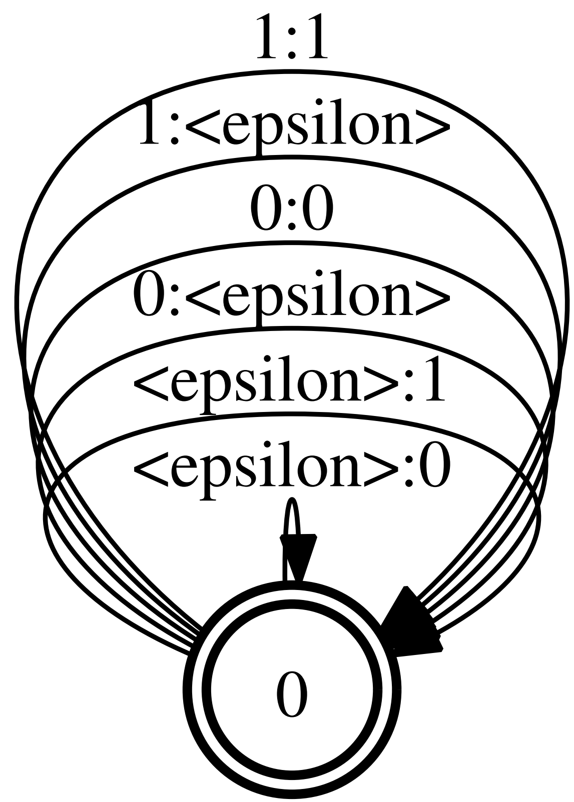
TARGET



Edit Distance and FSTs

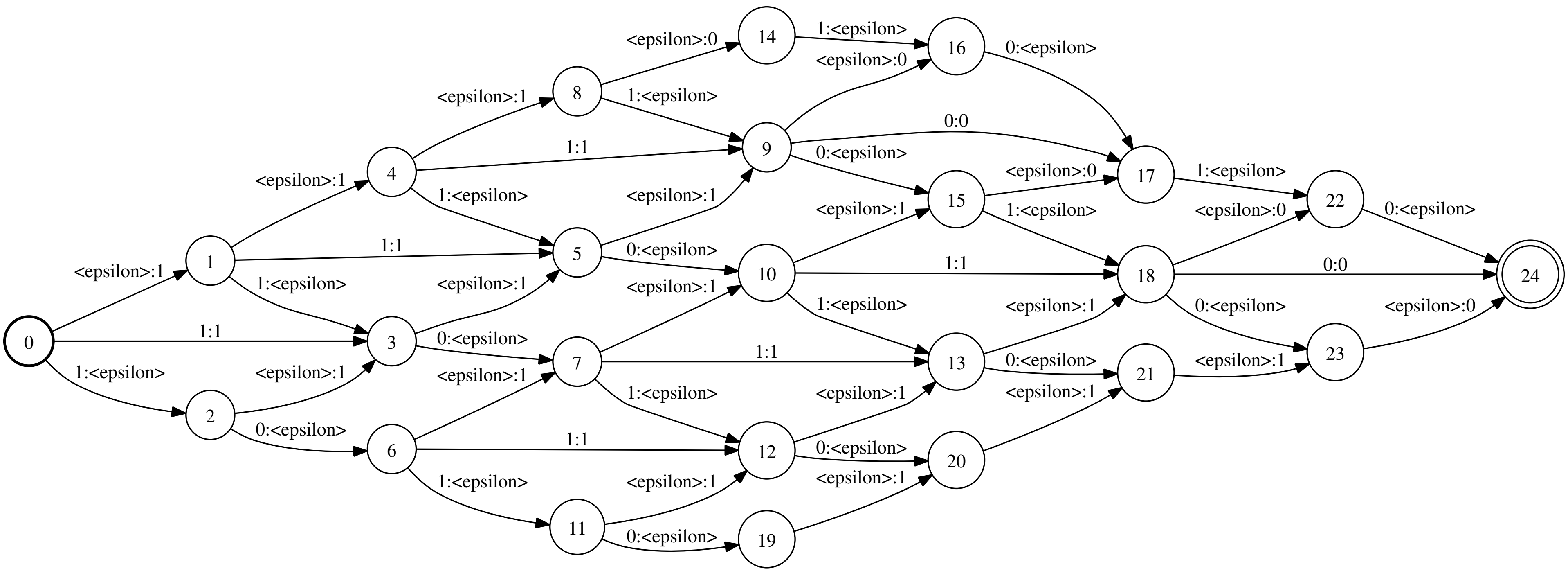
- Construct a FST that allows strings to be edited

EDITS



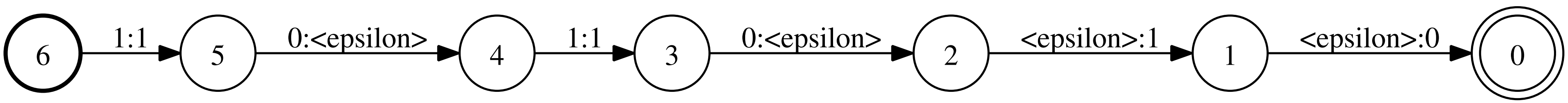
Edit Distance and FSTs

- Compose SOURCE and EDITS and TARGET



Edit Distance and FSTs

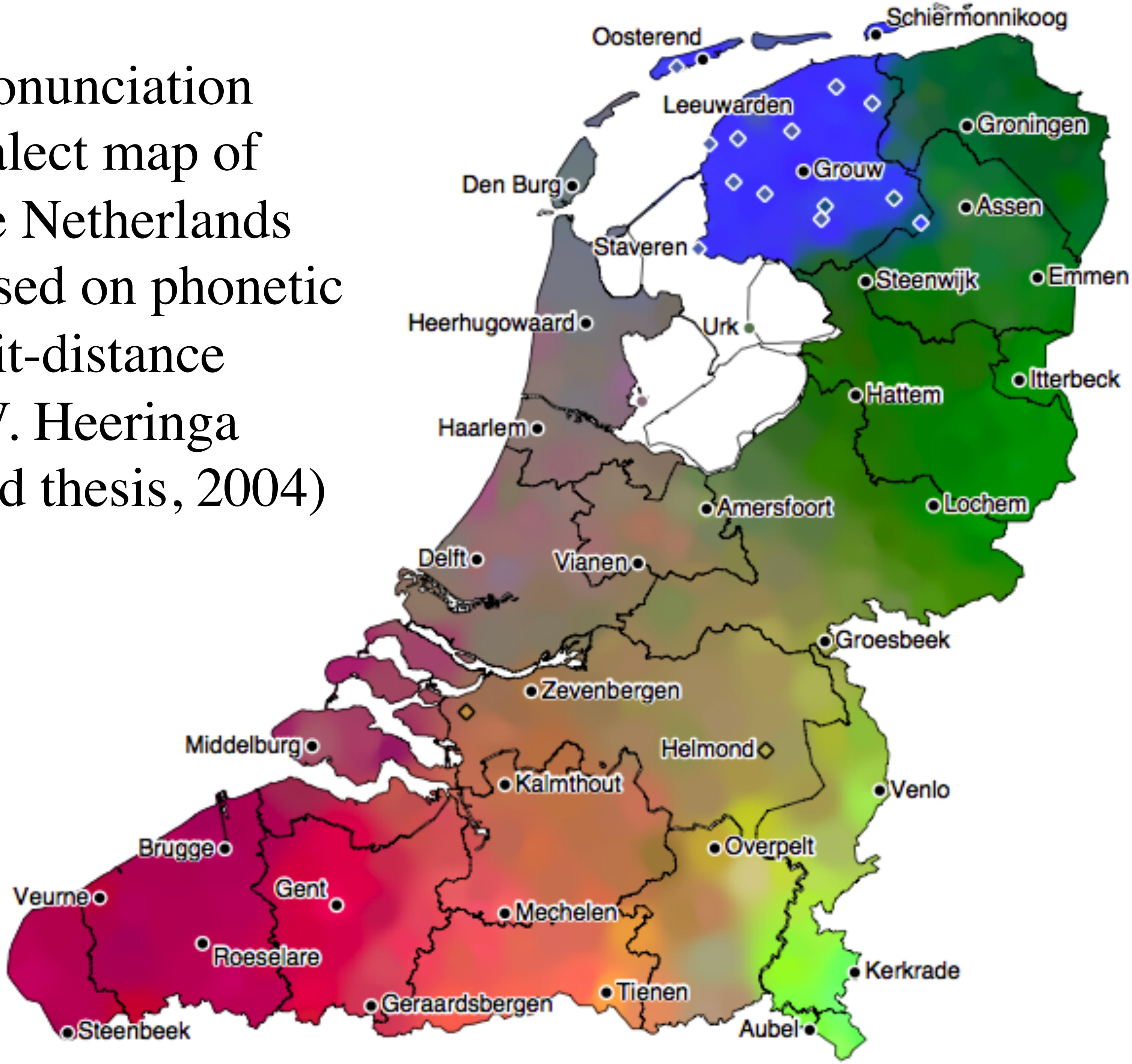
- The shortest path is the minimum edit FST from SOURCE (1010) to TARGET (1110)



Edit distance

- Useful in many NLP applications
- In some cases, we need edits with multiple characters, e.g. 2 chars deleted for one cost
- Comparing system output with human output, e.g. *input:* ibm *output:* IBM vs. Ibm (TrueCasing of speech recognition output)
- Error correction
- Defined over character edits or word edits, e.g. MT evaluation:
  - Foreign investment in Jiangsu ‘s agriculture on the increase
  - Foreign investment in Jiangsu agricultural investment increased

Pronunciation dialect map of the Netherlands based on phonetic edit-distance (W. Heeringa Phd thesis, 2004)



Variable Cost Edit Distance

- So far, we have seen edit distance with uniform insert/delete cost
- In different applications, we might want different insert/delete costs for different items
- For example, consider the simple application of spelling correction
- Users typing on a qwerty keyboard will make certain errors more frequently than others
- So we can consider insert/delete costs in terms of a probability that a certain alignment occurs between the *correct* word and the *typo* word

Spelling Correction

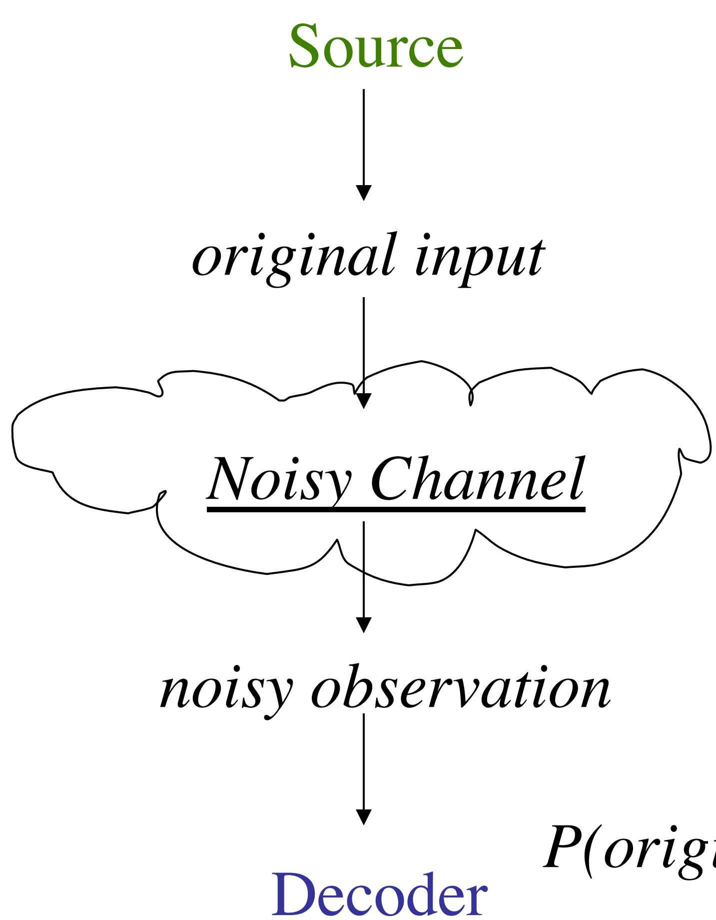
- Types of spelling correction
  - non-word error detection

    e.g. *hte* for *the*
  - isolated word error detection

    e.g. *acres* vs. *access* (cannot decide if it is the right word for the context)
  - context-dependent error detection (real world errors)

    e.g. *she is a talented acres* vs. *she is a talented actress*
- For simplicity, we will consider the case with exactly 1 error

Noisy Channel Model



$P(\textit{original input} \mid \textit{noisy obs})$

Bayes Rule: *computing P(orig | noisy)*

- let $x = original\ input, y = noisy\ observation$

$$p(x \mid y) = \frac{p(x,y)}{p(y)} \qquad p(y \mid x) = \frac{p(y,x)}{p(x)}$$

$$p(x,y) = p(y,x)$$

$$p(x \mid y) \times p(y) = p(y \mid x) \times p(x)$$

$$p(x \mid y) = \frac{p(y \mid x) \times p(x)}{\cancel{p(y)}}$$

Bayes Rule

Chain Rule

$$
\begin{aligned}
p(a,b,c \mid d) = \quad & p(a \mid b,c,d) \times \\
& p(b \mid c,d) \times \\
& p(c \mid d)
\end{aligned}
$$

Approximations: Bias vs. Variance

$$
\begin{aligned}
p(a \mid b,c,d) \approx \quad & p(a \mid b,c) \quad \textit{less } \boldsymbol{\textit{bias}} \\
& p(a \mid b) \\
& p(a) \quad \textit{less } \boldsymbol{\textit{variance}}
\end{aligned}
$$

Single Error Spelling Correction

- Insertion (addition)
  - acress vs. cress
- Deletion
  - acress vs. actress
- Substitution
  - acress vs. access
- Transposition (reversal)
  - acress vs. caress

Noisy Channel Model for Spelling Correction (Kernighan, Church and Gale, 1990)

- t is the word with a single typo and c is the correct word

$$P(c \mid t) = p(t \mid c) \times p(c)$$

Bayes Rule

- Find the best candidate for the correct word

$$\hat{c} = \underset{c \in C}{\arg\max}\, P(t \mid c) \times P(c)$$

$$P(t \mid c) = ?? \qquad\qquad P(c) = \frac{f(c)}{N}$$

C is all the words in the vocabulary; |C| = N

Noisy Channel Model for Spelling Correction (Kernighan, Church and Gale, 1990) single error, condition on previous letter

P(*poton* | *potion)*

$$P(t \mid c) = \begin{cases} \dfrac{del[c_{p-1}, c_p]}{chars[c_{p-1}, c_p]} & (xy)_c \text{ typed as } (x)_t \\ \dfrac{ins[c_{p-1}, t_p]}{chars[c_{p-1}]} & (x)_c \text{ typed as } (xy)_t \\ \dfrac{sub[t_p, c_p]}{chars[c_p]} & (y)_c \text{ typed as } (x)_t \\ \dfrac{rev[c_p, c_{p+1}]}{chars[c_p, c_{p+1}]} & (xy)_c \text{ typed as } (yx)_t \end{cases}$$

P(*poton* | *piton)*

t = *poton*
c = *potion*
del[t,i]=427
chars[t,i]=575
P = .7426

t = *poton*
c = *piton*
sub[o,i]=568
chars[i]=1406
P = .4039

Noisy Channel model for Spelling Correction

- The *del*, *ins*, *sub*, *rev* matrix values need data in which contain known errors

  (**training data**)

  e.g. Birbeck spelling error corpus (from 1984!)

- Accuracy on single errors on unseen data

  (**test data**)

Noisy Channel model for Spelling Correction

- Easily extended to multiple spelling errors in a word using edit distance algorithm (however, using learned costs for ins, del, replace)
- Experiments: 87% accuracy for machine vs. 98% average human accuracy
- What are the limitations of this model?

…*was called a "stellar and versatile* ***acress*** *whose combination of sass and glamour has defined her* …

KCG model best guess is **acres**