CMPT 413: Computational Linguistics
ED4: Variable Cost Edit Distance

Anoop Sarkar

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

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

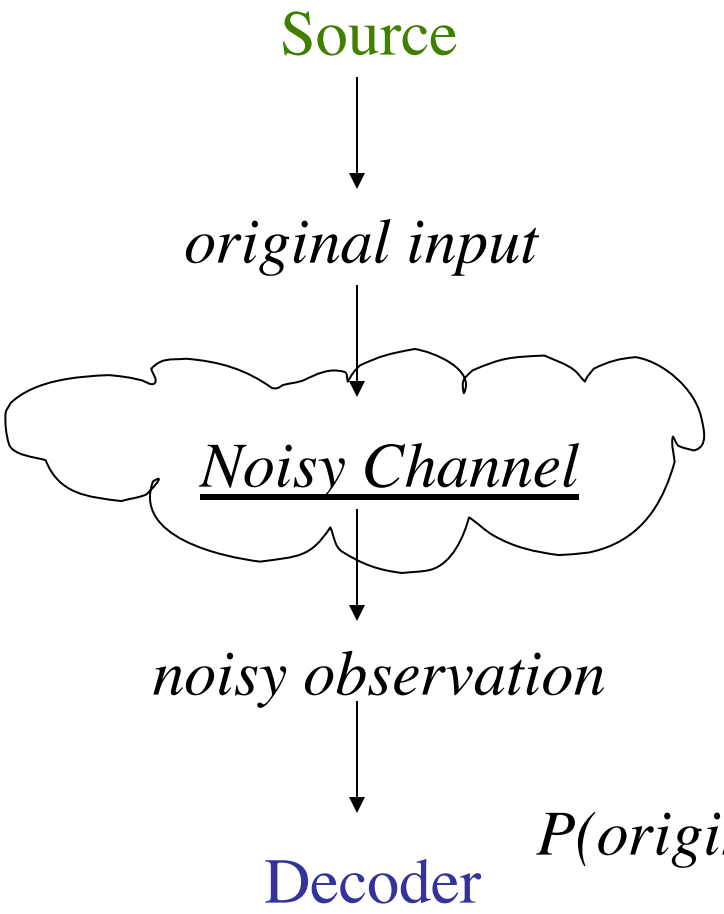
- Types of spelling correction
  - non-word error detection

    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



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*

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

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

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

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(*poton* | *piton*)

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

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

  (http://ota.ahds.ac.uk/texts/0643.html)
- Accuracy on single errors on unseen data

  (**test data**)

from (Kernighan, Church and Gale, 1990)

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X	Y (correct)																								
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Noisy Channel model for Spelling Correction

- Easily extended to multiple spelling errors in a word using edit distance algorithm
- 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**

More on spell checking

- Check out Peter Norvig's introduction to spell checking
  - http://norvig.com/spell-correct.html
- Better version of this appears in a book chapter
  - http://norvig.com/ngrams/