Natural Language Processing

Lecture 8: Information Theory; Spelling, Edit Distance, and Noisy Channels

A Taste of Information Theory

- Shannon Entropy, $H(p)$
- Cross-entropy, $H(p; q)$
- Perplexity

Codebook

Horse	Code
Clinton	000
Edwards	001
Kucinich	010
Obama	011
Huckabee	100
McCain	101
Paul	110
Romney	111

Codebook

Horse	Code	Probability
Clinton	000	1/4
Edwards	001	1/16
Kucinich	010	1/64
Obama	011	1/2
Huckabee	100	1/64
McCain	101	1/8
Paul	110	1/64
Romney	111	1/64

Codebook

Horse	Probability	New Code
Clinton	1/4	10
Edwards	1/16	1110
Kucinich	1/64	111100
Obama	1/2	0
Huckabee	1/64	111101
McCain	1/8	110
Paul	1/64	111110
Romney	1/64	111111

Codebook

Horse	Probability	New Code	Estimated Probability	Code
Clinton	1/4	10		
Edwards	1/16	1110		
Kucinich	1/64	111100		
Obama	1/2	0		
Huckabee	1/64	111101		
McCain	1/8	110		
Paul	1/64	111110		
Romney	1/64	111111		

Three Spelling Problems

1. Detecting isolated non-words
2. Fixing isolated non-words
3. Fixing errors in context

Levenshtein Distance

$$D_{0,0} = 0$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + \text{inscost}(t_i) \\ D_{i,j-1} + \text{delcost}(s_j) \\ D_{i-1,j-1} + \text{substcost}(t_i, s_j) \end{cases}$$

~~Levenshtein~~ Hamming Distance

$$D_{0,0} = 0$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + \infty \\ D_{i,j-1} + \infty \\ D_{i-1,j-1} + \text{substcost}(t_i, s_j) \end{cases}$$

Levenshtein Distance with Transposition

$$D_{0,0} = 0$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + \text{inscost}(t_i) \\ D_{i,j-1} + \text{delcost}(s_j) \\ D_{i-1,j-1} + \text{substcost}(t_i, s_j) \\ D_{i-2,j-2} + \text{transcost}(s_{j-1}, s_j) \text{if } s_{j-1} = t_i \text{ and } s_j = t_{i-1} \end{cases}$$

Three Spelling Problems

- ✓ Detecting isolated non-words
- ✓ Fixing isolated non-words
3. Fixing errors in context

Kernighan's Model: A Noisy Channel

acress

c	freq(*c*)	*p*(*t* | *c*)	%
actress	1343	p(delete t)	37
cress	0	p(delete a)	0
caress	4	p(transpose a & c)	0
access	2280	p(substitute r for c)	0
across	8436	p(substitute e for o)	18
acres	2879	p(delete s)	21
acres	2879	p(delete s)	23

Noisy Channel Model (General)