Collocations

Lecture #5

Introduction to Natural Language Processing
CMPSCI 585, Fall 2004
University of Massachusetts Amherst

Andrew McCallum

Words and their meaning

Some upcoming lectures:

- Word disambiguation
  - one word, multiple meanings
- Word clustering
  - multiple words, “same” meaning
- Collocations - *this lecture*
  - multiple words together, different meaning than than the sum of its parts
  - Simple measures on text, yielding interesting, insights into language, meaning, culture.

Today’s Main Points

- What is collocation?
- Why do people care?
- Three ways of finding them automatically.

Collocations

- An expression consisting of two or more words that correspond to some conventional way of saying things.
- Characterized by limited *compositionality*.
  - *compositional*: meaning of expression can be predicted by meaning of its parts.
  - “strong tea”, “rich in calcium”
  - “weapons of mass destruction”
  - “kick the bucket”, “hear it through the grapevine”

Collocations important for…

- Terminology extraction
  - Finding special phrases in technical domains
- Natural language generation
  - To make natural output
- Computational lexicography
  - To automatically identify phrases to be listed in a dictionary
- Parsing
  - To give preference to parses with natural collocations
- Study of social phenomena
  - Like the reinforcement of cultural stereotypes through language (Stubbs 1996)

Contextual Theory of Meaning

- In contrast with “structural linguistics”, which emphasizes abstractions, properties of sentences
- Contextual Theory of Meaning emphasizes the importance of context
  - context of the social setting (not idealized speaker)
  - context of discourse (not sentence in isolation)
  - context of surrounding words
    Firth: “a word is characterized by the company it keeps”
- Example [Halliday]
  - “strong tea”, coffee, cigarettes
  - “powerful drugs”, heroin, cocaine
  - Important for idiomatically correct English, but also social implications of language use

Method #1
Frequency

80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said

Method #1
Frequency with POS Filter
AN, NN, AAN, ANN, NAN, NNN, NPN

11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	A N
3301	last	year	N N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Method #2
Mean and Variance

- Some collocations are not of adjacent words, but words in more flexible distance relationship
  - she knocked on his door
  - they knocked at the door
  - 100 women knocked on Donaldson’s door
  - a man knocked on the metal front door
- Not a constant distance relationship
- But enough evidence that “knock” is better than “hit”, “punch”, etc.

Method #2
Mean and Variance

Sentence:
Stocks crash as rescue plan teeters.

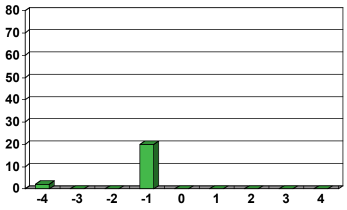
Time-shifted bigrams:

1	2	3
stocks crash	stocks as	stocks rescue
crash as	crash rescue	crash plan
as rescue	as plan	as teeters
...		

- To ask about relationship between “stocks” and “crash”, gather many such pairs, and calculate the mean and variance of their offset.

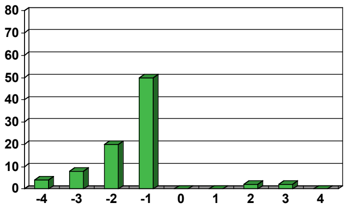
$$\text{mean} = \bar{o} = \frac{1}{n}\sum_{i=1}^{n} o_i \qquad \text{variance} = s = \frac{\sum_{i=1}^{n}(o_i - \bar{o})^2}{n-1}$$

Method #2
Mean and Variance



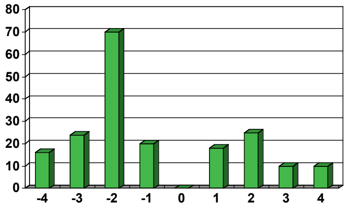
Position of “strong” versus “opposition” (mean=-1.15, deviation=0.67)

Method #2
Mean and Variance



Position of “strong” versus “support” (mean=-1.45, deviation=1.07)

Method #2
Mean and Variance



Position of “strong” versus “for” (mean=-1.12, deviation=2.15)

Method #2
Mean and Variance

dev	mean	count	Word1	Word2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by

Method #3
Likelihood Ratios

- Determine which of two probabilistic models is more appropriate for the data.
  - H1 = hypothesis of model 1
  - H2 = hypothesis of model 2

$$\text{likelihood ratio} = \log\left(\frac{L(H_1)}{L(H_2)}\right)$$

- Hypothesis 1: p(w2|w1) = p = p(w2|~w1)
- Hypothesis 2: p(w2|w2) = p1 ≠ p2 = p(w2|~w1)
- Data
  - N = total count of all words
  - c1 = count of word 1
  - c2 = count of word 2
  - c12 = count of bigram word1word2

Method #3
Likelihood Ratios

- Determine which of two probabilistic models is more appropriate for the data.

	H1	H2
P(w2|w1)	p=c2/N	p1=c12/c1
P(w2|~w1)	p=c2/N	p2=(c2-c12)/(N-c1)
c12 out of c1 bigrams are w1w2	b(c12; c1,p)	b(c12;c1,p1)
c2-c12 out of N-c1 bigrams are ~w1w2	b(c2-c12; N-c1, p)	b(c2-c12; N-c1,p2)

$$\text{likelihood ratio} = \log\left(\frac{L(H_1)}{L(H_2)}\right) = \log\left(\frac{b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c_1, p_2)}\right)$$

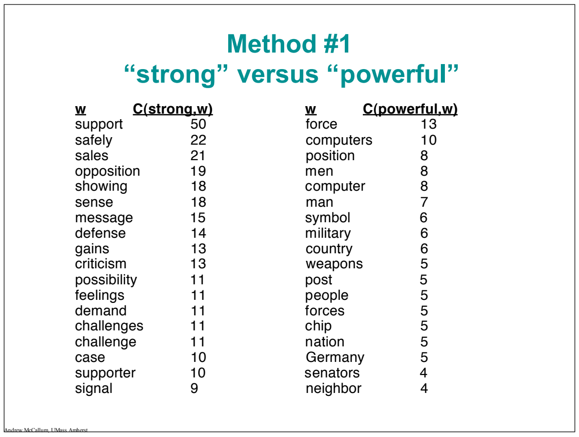
Method #3
Likelihood Ratio example data

-2log l	c1	c2	c12	w1	w2
1291	12593	932	150	most	powerful
99	379	932	10	politically	powerful
82	932	934	10	powerful	computers
80	932	3424	13	powerful	force
57	932	291	6	powerful	symbol
51	932	40	4	powerful	lobbies
51	171	932	5	economically	powerful
51	932	43	4	powerful	magnet
50	4458	932	10	less	powerful
50	6252	932	11	very	powerful
49	932	2064	8	powerful	position
48	932	591	6	powerful	machines
47	932	2339	8	powerful	computer
43	932	396	5	powerful	magnets

Collocation studies helping lexicography

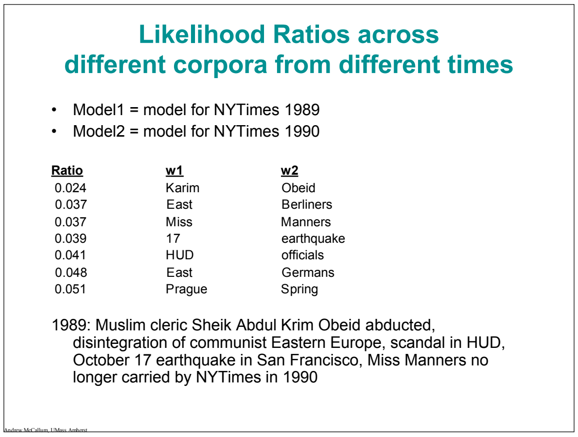
- Want to help dictionary-writers bring out differences between “strong” and “powerful”
  - Understand meaning of a word by the company it keeps.
- Church and Hanks (1989) through statistical analysis concluded that it is a matter of *intrinsic* vs *extrinsic* quality
- “strong” support from a demographic group, means committed, but may not have capability.
- “powerful” supporter is one who actually has capability to change things.

- But also additional subtleties, helps us analyze cultural attitudes
  - “strong tea” versus “powerful drugs”



Method #1
“strong” versus “powerful”

w	C(strong,w)	w	C(powerful,w)
support	50	force	13
safely	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	country	6
criticism	13	weapons	5
possibility	11	post	5
feelings	11	people	5
demand	11	forces	5
challenges	11	chip	5
challenge	11	nation	5
case	10	Germany	5
supporter	10	senators	4
signal	9	neighbor	4



Likelihood Ratios across different corpora from different times

- Model1 = model for NYTimes 1989
- Model2 = model for NYTimes 1990

Ratio	w1	w2
0.024	Karim	Obeid
0.037	East	Berliners
0.037	Miss	Manners
0.039	17	earthquake
0.041	HUD	officials
0.048	East	Germans
0.051	Prague	Spring

1989: Muslim cleric Sheik Abdul Krim Obeid abducted, disintegration of communist Eastern Europe, scandal in HUD, October 17 earthquake in San Francisco, Miss Manners no longer carried by NYTimes in 1990