CS 565: Intelligent Systems and Interfaces

Lecture: Finding Collocations – Alternative Tests

22nd Jan 2016

Semester: Jan - May 2016

Ashish Anand

IIT Guwahati

Finding collocations

- Already Discussed
  - Frequency + PoS-Tag Filter [Count + Domain Knowledge]
  - Mean & Variance [Basic Statistics]
  - t-test [Statistical Test]

- Will Discuss
  - Chi-square Test [Statistical Test]
  - Likelihood Ratio Test [Statistical Test]
  - Mutual Information [Information Theory]

Pearson’s Chi-square Test

- Does not require normal distribution assumption as in t-test
- Test for dependence or association
- Make a frequency or contingency table
- Compare observed and expected frequencies

Chi-square test: contd.

	w1 = new	w1 ≠ new
w2 = companies	8	4667
w2 ≠ companies	15820	14287173

$$X^2 = \sum_{ij} \frac{(O_{ij} - Eij)^2}{E_{ij}}$$

O_{ij}: Observed frequency; E_{ij}: Expected frequency

X^2 is asymptotically χ^2 distributed.

Chi-Square: Other Applications

- Identification of translation pairs in aligned corpora (Church and Gale 1991b).

	w_1	$\neg w_1$
w_2		
$\neg w_2$		

- Metric for corpus similarity (Kilgarriff and Rose, 1998)

	Corpus 1	**Corpus 2**
Word 1	w_{11}	w_{12}
Word 2	w_{21}	w_{22}
Word 3	w_{31}	w_{32}

Likelihood Ratio Test

- Two alternate hypotheses
  - H1: $p(w_2 \mid w_1) = p = p(w_2 \mid \neg w_1)$ -> Independence
  - H2: $p(w_2 \mid w_1) = p1 \neq p2 = p(w_2 \mid \neg w_1)$ -> Association

- Define Likelihood Ratio, $\lambda = L(H_1) / L(H_2)$
  - A number telling how much more likely is one hypothesis over the other.

Calculating Probabilities and Likelihood

- What we do
  - $p = c_2/N$; $p_1 = c_{12} / c_1$; $p_2 = (c_2 - c_{12}) / (N - c_1)$

  c_2: # of occurrence of w_i; c_{12}: # of occurrence of w_{ij}

- Under the hood
  - Maximum Likelihood Estimate

Likelihood Ratio Test

	$\mathbf{H_1}$	$\mathbf{H_2}$
$P(w_2 \mid w_1)$	$p = c_2 / N$	$p_1 = c_{12} / c_1$
$P(w_2 \mid \neg w_1)$	$p = c_2 / N$	$p_2 = (c_2 - c_{12}) / (N - c_1)$
c_{12} out of c_1 bigrams are $w_1 w_2$	$b(c_{12}; c_1, p)$	$b(c_{12}; c_1, p_1)$
c_2-c_{12} out of N-c_1 bigrams are $\neg w_1 w_2$	$b(c_2 - c_{12}; N - c_1, p)$	$b(c_2 - c_{12}, N - c_1, p_2)$

$L(H_1) = b(c_{12}; c_1, p)\ b(c_2 - c_{12}; N - c_1, p)$
$L(H_2) = b(c_{12}; c_1, p_1)\ b(c_2 - c_{12}, N - c_1, p_2)$

$\text{Log}\ \lambda = \log\ (L(H_1) / L(H_2))$
$-2 \log L \sim \chi^2$

$-2 \log \lambda$	$C(w^1)$	$C(w^2)$	$C(w^1 w^2)$	w^1	w^2
1291.42	12593	932	150	most	powerful
99.31	379	932	10	politically	powerful
82.96	932	934	10	powerful	computers
80.39	932	3424	13	powerful	force
57.27	932	291	6	powerful	symbol
51.66	932	40	4	powerful	lobbies
51.52	171	932	5	economically	powerful
51.05	932	43	4	powerful	magnet
50.83	4458	932	10	less	powerful
50.75	6252	932	11	very	powerful
49.36	932	2064	8	powerful	position
48.78	932	591	6	powerful	machines
47.42	932	2339	8	powerful	computer
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip
40.45	932	3694	8	powerful	men
36.36	932	47	3	powerful	486
36.15	932	268	4	powerful	neighbor
35.24	932	5245	8	powerful	political
34.15	932	3	2	powerful	cudgels

Table 5.12 Bigrams of *powerful* with the highest scores according to Dunning's likelihood ratio test.

Source: Table 5.12 [FSNLP]

Mutual Information

- Mutual information: estimation of how much one word tells us about the other

$$I(x', y') = \log_2 \frac{p(x', y')}{p(x')p(y')}$$

$$= \log_2 \frac{p(x'|y')}{p(x')}$$

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

Table 5.14 Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information.

Source: Table 5.14, FSNLP, p 178.

Mutual Information

- Not considered as a good measure
  - Reduction of uncertainty
  - Issues with low frequency words

M.I. - Issue with data sparseness

I_{1000}	w^1	w^2	w^1w^2	Bigram	I_{23000}	w^1	w^2	w^1w^2	Bigram
16.95	5	1	1	Schwartz eschews	14.46	106	6	1	Schwartz eschews
15.02	1	19	1	fewest visits	13.06	76	22	1	FIND GARDEN
13.78	5	9	1	FIND GARDEN	11.25	22	267	1	fewest visits
12.00	5	31	1	Indonesian pieces	8.97	43	663	1	Indonesian pieces
9.82	26	27	1	Reds survived	8.04	170	1917	6	marijuana growing
9.21	13	82	1	marijuana growing	5.73	15828	51	3	new converts
7.37	24	159	1	doubt whether	5.26	680	3846	7	doubt whether
6.68	687	9	1	new converts	4.76	739	713	1	Reds survived
6.00	661	15	1	like offensive	1.95	3549	6276	6	must think
3.81	159	283	1	must think	0.41	14093	762	1	like offensive

Source: Table 5.16, FSNLP, p 181

Reference

- FSNLP: 5.3 – 5.6

- Additional Readings
  - FSNLP – Ch 2 [Background in Probability]

- Suggested background reading
  - Maximum Likelihood Estimation: 3 – 3.2, Pattern classification, Duda, Hart and Stork.