Medidas de Associação

Luís Sarmento
Simpósio Doutoral Linguateca
3, 4 Out 2006

1. Pequena Introdução

Dados quantitativos e co-ocorrências

- A base da linguística empírica consiste na utilização de informação quantitativa (frequências) retirada de texto reais para propor e testar hipóteses
- Uma das grandezas observáveis particularmente interessantes é a informação de co-ocorrência entre palavras/estruturas:
  - o texto não é um simples amontoado aleatório de estruturas e por isso a co-ocorrência em texto de duas estruturas traz consigo informação importante

Quando duas palavras co-ocorrem

- Essa co-ocorrência pode indicar directamente que as palavras estão relacionadas:
  - por composicionalidade: "rede local"
  - afinidade ou relação semântica: "barco…. pesca"
- Indirectamente pode indicar *Semelhança*:
  - **Hipótese Distribucional do Significado**: se duas palavras co-ocorrem com as mesmas palavras então há uma certa semelhança entre elas…

Mas…

- Potencialmente, todas as palavras podem co-ocorrer com todas as outra palavras sem que isso signifique logo que exista uma relação "interessante":
  - "banana" < > "comentário": 14
  - "banana" < > "comida": 7
- Devemos apenas considerar as co-ocorrências "significativas", aquelas em que realmente existe uma "associação" ou uma "dependência" entre os elementos que co-ocorrem…
  - Depende dos corpora utilizados!

Esta apresentação

- Como é que poderemos usar a informação quantitativa relativa a certas estruturas para aferir grau de Associação entre as mesmas e a partir dai tirar conclusões
- As entidades centrais nesta questão são as *funções estatísticas que estimam o grau de associação* entre estruturas usando informação acerca da frequência de co-ocorrências observada:

Medidas de Associação

2. Co-ocorrências

Algumas definições

Definições elementares

- **co-ocorrência**: o evento da ocorrência simultânea de dois ou mais elementos, ditos **elementos co-ocorrêntes** nas condições definidas por um **contexto de co-ocorrência**.
- **contexto de co-ocorrência**: conjunto de restrições lexicais, gramaticais ou de outro tipo, que determinam o âmbito no qual a co-ocorrência é considerada.
  - co-ocorrências fora do contexto de co-ocorrência definido não são contabilizadas.

Elementos Co-ocorrentes

- Podem ser:
  - directa ou indirectamente observados em texto
  - as formas encontradas no texto e contabilizados tal como estão
  - resultado de um processamento linguístico simples (ex: lematização, etc…)
  - elementos mais complexos resultantes de análise linguística (tripletos SVO, etc…)
  - …

Co-ocorrências simples

- Situação mais simples / habitual.
- Consiste na co-ocorrência de duas *palavras simples*, w_x e w_y , podendo o contexto de co-ocorrência *C* variar bastante mediante o objectivo.
  - Por vezes há um processo de lematização para consolidar a contagem das frequências

Contextos possíveis

- Janela:
  - 1 elementos (adjacência)
  - N elementos
- Unidade Estrutural (janela "natural"):
  - Frase
  - Parágrafo
  - Documento

Janela 1 elementos (adjacência)

- Por exemplo, a seguinte frase (note-se a presença do ponto final):
  - w1 w2 w3 w2 w3 .
- permitiria gerar as seguintes **observações** para janela de tamanho 1:
  - o(w1 , w2) = 1
  - o(w2 , w3) = 2
  - o(w3 , w2) = 1
- Eventualmente (dependendo da definição):
  - o(#I #, w1) = 1 o(w3 , .) = 2 o(., #F #) = 1

Co-ocorrência Relacional

- Quando existem ferramentas que permitem identificar *relacionamentos* entre palavras torna-se possível considerar um outros tipo de co-ocorrências que não depende directamente *sequência* das palavras
  - árvores de dependências
  - Tripletos de relações

Situações Possiveis

- tripletos do tipo SVO (Subject - Verb - Object) em que a co-ocorrência involve não **dois** mas sim **três** elementos
  - mas pode ser projectadas numa co-ocorrência de **dois** elementos
- contextos lexicais pouco ambíguos
  - coordenações, caso possessivo em inglês, Padrões de Hearst, etc…

Exemplo simples:

- Coordenações (3 elementos):
  - o(”preto”, ”e”, ”branco”) = c1
  - o(”amarelo”,”ou”,”vermelho”) = c2
  - o(”vidro”,”ou”,”madeira”) = c3
- Podem ser “projectadas” em co-ocorrências de 2 elementos:
  - o_{ou} (”amerelo”,”vermelho”) = c2
  - o_{ou} (”vidro”,”madeira”) = c3

3. Noção básicas de Estatistica

Só aquilo que é mesmo preciso

Alguns conceitos que vamos precisar

- O Saco de Palavras
- A Hipótese Nula da Independência, H_0
- A Tabela de Contingência
- Estimativas e valores esperados

O Saco de Palavras

- Um texto apresenta normalmente uma estrutura sequencial que obedece a certas regras gramaticais e princípios semânticos
- Apesar disso é frequente modelizar estatisticamente o texto considerando um grupo aleatória de palavras, sem noção de sequência.
- Esta simplificação, é conhecida como:
  - **bag-of-words (bow) assumption**
  - **premissa do saco de palavras**

Meter tudo no mesmo saco

- Sob a premissa do saco de palavras qualquer texto - frase, documento, colecção - é visto como um grupo de elementos
- Dependendo da aplicação estes elementos podem ser:
  - palavras simples, n-gramas, estruturas compostas ou outros
- Tais elementos podem ser vistos como um conjunto não ordenado de elementos guardado num saco.
- Neste tutorial iremos usar a letra N para representar o número de elementos do saco

Por exemplo…

- Frase: w1 w2 w3 w2 w3 .
- Saco com N = 6 palavras simples:
  - (w1 : 1) (w2 : 2) (w3 : 2) (. : 1)
- Saco com N=5 bigramas
  - (w1w2 : 1) (w2w3 : 2) (w3w2 : 1) (w3 . : 1)
- Muitas outras possibilidades…

A Hipótese Nula da Independência, H_0

- Na sequência da premissa do Saco de Palavras, é também habitual considerar a hipótese de que os eventos de ocorrência de cada uma das palavras do saco é independente das restantes.
- Se retirarmos do saco aleatoriamente uma palavra wx em nada influenciamos a ocorrência de qualquer outra palavra wy que retirarmos aleatoriamente do saco depois

Mais formalmente…

- Para uma dada função de probabilidade p(x) que descreve a ocorrências das palavras no saco temos que:

$$p_{H0}(X, Y) = p(X)p(Y)$$

- Ou seja, sob a Hipótese H0, as ocorrências de X e Y são totalmente *independentes*

Obviamente…

- que as palavras que ocorrem num determinado texto não são independentes e por isso deverá ser possível reunir evidência empirica que retire sustentação Hipótese H_0.
- O papel das Medidas de Associação é quantificar o grau de "insustentabilidade" de H_0, e fornecer uma medida do grau de associação entre X e Y.
  - É disto que iremos a falar!! :)

A Tabela de Contingência

- Os formalismos empregues na apresentação das Medidas de Associação entre dois elementos X e Y variam consoante os autores que as propõem
- É habitual recorrer a uma uma estrutura de dados padrão onde se explicitam os valores relevantes no cálculo das Medidas de Associação: a Tabela de Contingência.
- Torna-se depois mais simples *formular* e *comparar* as várias Medidas de Associação

Tabela de Contingência (2)

- Para **dois** elementos, X e Y, a Tabela de Contingência (TC) é uma matriz 2 x 2 que contém quatro valores relativos às frequências das quatro combinações possíveis para as ocorrências de X e Y relativamente ao corpus / saco observado

$$TC_{XY} = \begin{bmatrix} o(X,Y) & o(X,!Y) \\ o(!X,Y) & o(!X,!Y) \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Tabela de Contingência (3)

- Verifica-se que :
  o(X, Y) + o(X, !Y) + o(!X, Y) + o(!X, !Y) =
  = a + b + c + d = N
- N é o número total de pares.
  - Corresponde ao total de elementos no saco!
- Vamos ser capazes de formular TODAS as Medidas de Associação com base na TC

Note-se que…

- Por simples inspecção, os valores da TC ficamos com uma ideia acerca do grau de associação de dois elementos
- quanto mais vezes dois elementos co-ocorrerem em simultâneo, o(X,Y) = a, e menos vezes ocorrem separadamente o(X,!Y) = b e o(!X,Y) = c, mais forte deverá ser a sua associação!

Estimativas e Valores Esperados

- Algumas Medidas de Associação são formuladas recorrendo funções de probabilidade relativas à (co-)ocorrências
- Usando *Estimadores de Máxima Verosimilhança* há vários valores de probabilidade podem ser estimados usando os dados existentes na Tabela de Contingência.

Estimação usando valores observados

$$p(X,Y) \simeq \frac{o(X,Y)}{N} = \frac{a}{N}$$
$$p(X,!Y) \simeq \frac{o(X,!Y)}{N} = \frac{b}{N}$$
$$p(!X,Y) \simeq \frac{o(!X,Y)}{N} = \frac{c}{N}$$

$$p(X) \simeq \frac{o(X)}{N} = \frac{o(X,Y)+o(X,!Y)}{N} = \frac{a+b}{N}$$
$$p(Y) \simeq \frac{o(Y)}{N} = \frac{o(X,Y)+o(!X,Y)}{N} = \frac{a+c}{N}$$
$$p(!X) \simeq \frac{o(!X)}{N} = \frac{o(!X,Y)+o(!X,!Y)}{N} = \frac{c+d}{N}$$
$$p(!Y) \simeq \frac{o(!Y)}{N} = \frac{o(X,!Y)+o(!X,!Y)}{N} = \frac{b+d}{N}$$

Estimando sob a Hipótese H_0

$$p_{H_0}(X,Y) = p(X)\cdot p(Y) \simeq \frac{o(X)}{N}\cdot\frac{o(Y)}{N} = \frac{a+b}{N}\cdot\frac{a+c}{N}$$

$$p_{H_0}(X,!Y) = p(X)\cdot p(!Y) \simeq \frac{o(X)}{N}\cdot\frac{o(!Y)}{N} = \frac{a+b}{N}\cdot\frac{b+d}{N}$$

$$p_{H_0}(!X,Y) = p(!X)\cdot p(Y) \simeq \frac{o(!X)}{N}\cdot\frac{o(Y)}{N} = \frac{c+d}{N}\cdot\frac{a+c}{N}$$

$$p_{H_0}(!X,!Y) = p(!X)\cdot p(!Y) \simeq \frac{o(!X)}{N}\cdot\frac{o(!Y)}{N} = \frac{c+d}{N}\cdot\frac{b+d}{N}$$

Valores Esperados sob H_0

- Valores que seria expectável encontrar na C caso H_0 fosse verdadeira

$$e_{H_0}(X,Y) = p_{H_0}(X,Y) \cdot N = \frac{o(X) \cdot o(Y)}{N} = \frac{(a+b) \cdot (a+c)}{N}$$

$$e_{H_0}(X,!Y) = p_{H_0}(X,!Y) \cdot N = \frac{o(X) \cdot o(!Y)}{N} = \frac{(a+b) \cdot (b+d)}{N}$$

$$e_{H_0}(!X,Y) = p_{H_0}(!X,Y) \cdot N = \frac{o(!X) \cdot o(Y)}{N} = \frac{(c+d) \cdot (a+c)}{N}$$

$$e_{H_0}(!X,!Y) = p_{H_0}(!X,!Y) \cdot N = \frac{o(!X) \cdot o(!Y)}{N} = \frac{(c+d) \cdot (b+d)}{N}$$

4. Medidas de Associação

Uma perspectiva geral…

Medidas de Associação (MA)

- É uma formula que calcula o valor da associação existente entre duas variáveis, usando a informação que se encontra na TC correspondente
- Tentam quantificar o grau de dependência entre duas variáveis tendo por base informação acerca das suas co-ocorrências.

Quantificando…

- um elevado valor da MA indica que as variáveis em causa estão ***associadas***, ou seja que a ocorrência de uma delas está associada à ocorrência da outra
- se as variáveis forem independentes então espera-se que a Medida de Associação consiga identificar essa condição e produzir uma valor reduzido, normalmente próximo de 0

MA uni-direcionais / bi-direcionais

- **MA uni-direccional:** se for apenas capaz de identificar a existência de uma associação entre dois elementos X e Y, diferenciando essa situação da ausência da associação / independência
- **MA bi-direcional:** se a medida for também capaz de destinguir entre a situação de associação positiva (X e Y têm tendência a co-ocorrer) e a situação de associação negativa (X e Y têm tendência a excluir-se mutuamente)

A natureza das Medidas

- Segundo (Evert, 2005) há 4 grandes grupos:
  - Testes de Significância estatística
  - Coeficientes de Associação
  - baseadas em conceitos de Teoria de Informação
  - baseadas em Heurísticas diversas

Testes de Significância estatística

- Estas medidas recorrem a todo um conjunto de Testes de Significância *genéricos* disponibilizados pela estatística
- Os testes tentam reunir evidência que permite negar a Hipótese H0, a partir dos dados disponíveis na Tabela de Contingência
- Exemplos:
  - o Teste X^2
  - o Teste de Fisher
  - Razão de Log-Verosimilhança

Coeficientes de Associação

- Estas medidas consistem na aplicação de estimativas de máxima verosimilhança calculadas a partir da Tabela de Contingência como parâmetros de certos Coeficientes de Associação disponíveis na estatística.
- Exemplos:
  - Coeficiente de Dice
  - Coeficiente de Jaccard
  - Risco Relativo.

Medidas baseadas em conceitos de Teoria de Informação

- Estas medidas partem dos conceitos elementares de Entropia e Ganho de Informação para definirem Medidas de Associação.
- Exemplos:
  - Informaçao Mútua
  - Dependência Mútua.

Baseadas em heurísticas

- Todo um conjunto de medidas baseadas em variantes empíricas das anteriores ou medidas que tentam explorar estatisticamente certas intuições linguísticas.
- Exemplos:
  - familia IM^n
  - Coeficiente Combinado IM/t.

Neste tutorial

- Iremos concentrar-nos principalmente:
  - num sub-tipo de medidas baseadas em **Testes de Significância**
  - num sub-tipo de medidas baseada em conceitos de **Teoria de Informação**

5. Medidas baseadas em Testes de Significância

O que são?

- Medidas de Associação que recorrem a um campo da estatística com fortes fundamentos teóricos: os Testes de Significância (TS)
- Os Testes de Significância permitem, com base nos valores observados e presentes na TC, rejeitar ou não uma determinada hipótese acerca da amostra
- A hipótese subjacente ao teste é a Hipótese Nula da Independência H_0

Tipos de MA baseadas TS (1)

- **Testes de Verosimilhança**:
  - estes testes procuram calcular a probabilidade, conhecida como Verosimilhança, da TC conter os valores observados tendo em conta H_0.
  - Valores baixos indicam que H_0 é pouco provável

Tipos de MA baseadas TS (2)

- **Testes de Hipóteses Exactos**:
  - os testes exactos passam por calcular a probalidade de se incorrer num Erro de Tipo I na decisão acerca de H_0, i.e. de rejeitar a injustificadamente.
  - Os Testes de Hipóteses Exactos tentam somar todas as evidências contra H_0 fornecidas por todas as TC possíveis para os dados observados
  - Os testes de Hipóteses Exactos evolvem normalmente bastante computação

Tipos de MA baseadas TS (3)

- Testes de Hipóteses Assimptóticos:
  - assumem o Pressuposto da Normalidade dos dados, o que permite testes mais simples de calcular do que os testes exactos.
  - calculam uma estatística que permite verificar quão bem é que os valores de frequência observados numa determinada amostra correspondem ao valor teoricamente esperado sob H_0
  - em inglês este tipo de teste é frequentemente chamado *goodness-of-fit test*
  - os Testes Assimptóticos são muito populares!
    - Vamos focar nestes testes!

Testes Assimptóticos

- O testes que vamos ver:
  - O teste X^2 de Pearson
  - O teste Z
  - O teste T de Student
  - A razão de Log-Verosimilhança (ou G^2)
- Caso "prático" com que iremos ilustrar:
  - pesquisa de unidades multipalavra bigramas

O teste X^2 de Pearson

- O teste X^2 é um teste unidireccional que consiste em calcular uma estatística baseada nos desvios entre as frequência observadas o(i) e as frequências esperadas sob H0, $e_{H0}(i)$:

$$X^2 = \sum_{i=1}^{k} \frac{(o(i) - e_{H_0}(i))^2}{e_{H_0}(i)}$$

Como funciona?

- a estatística X^2 é uma variável aleatória que pode ser aproximada pela distribuição chi^2 com v = (L_{inhas}-1) · (C_{olunas}-1) graus de liberdade
  - no nosso caso L=2 e C=2, o que implica v=1
- se X^2 for muito baixo, então a amostra segue aproximadamente a distribuição teoricamente esperada sob H_0, não se justificando a sua rejeição
- se X^2 ultrapassar um limite crítico dado pela distribuição chi^2, então rejeitamos H_0

Correcção de Yates

- para os casos em que v = 1 e se as frequências envolvidas forem baixas, X^2 não acompanha bem a distribuição chi^2
- É necessário um factor de correcção:

$$X^2_{Yates} = \sum_{i=1}^{k} \frac{(|o_i - e_i| - 0.5)^2}{e_i}$$

- O factor torna-se insignificante quando as frequências em causa são elevadas (> 10)
- para frequências < 5 o teste X^2 **não** deve ser aplicado

Exemplos…

- Estudar se um dado bigrama w1 w2 possui, ou não, um grau de associação suficientemente elevado para poder ser considerado um unidade multi-palavra
- Dois casos, usando o BACO (Sarmento, 2006):

$$TC_{(rede,local)} = \begin{bmatrix} 2559 & 284164 \\ 352944 & 3105006357 \end{bmatrix}$$

$$TC_{(rede,simples)} = \begin{bmatrix} 17 & 286706 \\ 131089 & 310522821 \end{bmatrix}$$

E os valores esperados sob H_0

$e_{H_0}(rede, local) = \frac{(a+b)\cdot(a+c)}{N} = 32.8$ $e_{H_0}(rede, simples) = 12.1$

$e_{H_0}(rede, !local) = \frac{(a+b)\cdot(b+d)}{N} = 286690.2$ $e_{H_0}(rede, !simples) = 286710.9$

$e_{H_0}(!rede, local) = \frac{(c+d)\cdot(a+c)}{N} = 355470.2$ $e_{H_0}(!rede, simples) = 131093.9$

$e_{H_0}(!rede, !local) = \frac{(c+d)\cdot(b+d)}{N} = 3105003830.8$ $e_{H_0}(!rede, !simples) = 3105228207.1$

$X^2_{Yates}(rede, local) = 194398.1$ $X^2_{Yates}(rede, simples) = 1.60$

Para alfa = 0.05 temos que chi$^2_{alfa}$ = 3.841

Uma amostra

- Tuplos w1w2 escolhidos aleatóriamente e ordenados por valor de X^2

Bigrama "rede w_2"	o("rede", w_2)	X^2_{Yates}
rede viária	3901	31832453.9
rede eléctrica	2053	1281978.6
rede porta-objectos	99	527942.8
rede local	2559	194398.2
rede rcts	109	84347.6
rede pan-europeia	42	42608.8
rede toner	142	27193.3
rede al-qaeda	58	15953.0
rede criminosa	42	7997.0
rede ibérica	71	4562.7
rede internet	421	3043.9
rede u.s.	27	2136.4
rede aceitação	55	1230.5
rede aérea	40	798.1
rede exploração	66	553.4
rede compatível	32	311.3
rede criada	27	173.5
rede operacional	23	57.65
rede podendo	22	38.67
rede apresentação	35	14.71
rede deverão	16	6.209
rede simples	17	1.597
rede toda	19	0.1462
rede conjunto	19	0.1068
rede apenas	38	0.1055
rede acesso	40	0.06707
rede sociedade	37	0.0003121
rede escolas	19	2.001e-06

O teste Z

- A utilização da medida Z para a identificação do colocações remonta à década de 70 (Berry-Rogghe), sendo um teste relativamente simples de calcular.
- Uma das principais referências na utilização da medida Z é o sistema Xtract (Smadja, 1993)

Teste Z

- Decorre do Teorema do Limite Central
  - Quando se realizam n_{amo} amostras de uma população cuja distribuição não é conhecida mas cuja média μ_{pop} e o desvio padrão σ_{pop} são conhecidos, a distribuição do valor da média das n amostras, μ_{amo}, tenderá para uma Distribuição Normal com média μ_{pop} e com desvio padrão $\sigma_{pop} / \sqrt{n_{amo}}$. Assim sendo, a estatística:

$$Z = \frac{\mu_{amo} - \mu_{pop}}{\sigma_{pop} / \sqrt{n_{amo}}}$$

tenderá para uma Distribução Normal Padrão.

No nosso caso temos n = 1 amostra (1 corpus)

Depois de algumas substituições…

- Ficamos com uma formula mais simples que encontramos na literatura:
- Ou:

$$Z = \frac{o(w_1, w_2) - e_{H_0}(w_1, w_2)}{\sqrt{e_{H_0}(w_1, w_2)}}$$

$$Z = \frac{a - \frac{(a+b)(a+c)}{N}}{\sqrt{\frac{(a+b)(a+c)}{N}}} = \frac{a \cdot N - (a+b)(a+c)}{\sqrt{N \cdot (a+b)(a+c)}}$$

- Que pode ser vista directamente como uma Medida de Associação
  - Já iremos ver…

Teste T de Student

- É utilizado na estatística para comparar duas ou mais amostras retiradas da mesma população
  - Serve, por exemplo, para tentar determinar se as médias das duas amostras são ou não significativamente diferentes (ex: comparar as frequências de uma palavra em dois corpora).
- O teste T, ao contrário, do teste Z, não parte do pressuposto que o desvio padrão da população de onde são retiradas as amostras é conhecido à partida.
- Faz menos pressupostos do que o teste Z

Teste T de Student (2)

$$T = \frac{\mu_{amo} - \mu_{pop}}{S_{pop} / \sqrt{n_{amo}}} \quad (5.18)$$

em que:

- μ_{amo} é o valor da média observada nas amostras
- μ_{pop} é o valor da média da população, que assumimos ser $e(w_1, w_2)$
- n_{amo} é o número de amostras observado
- S_{pop} a estimativa do desvio padrão da população σ_{pop} calculada a partir das n_{amo} amostras

Mas… há aqui um problema

- Como referido em (Evert, 2005, pag. 82), em teoria o teste T **não** é aplicável a dados relativos a frequências de co-ocorrência contabilizados a partir de **um** corpus (amostra)
- Este teste foi desenhado para comparar $n_{amo} > 1$ amostras independentes sobre a mesma população
- Recorrendo a uma única amostra (o corpus) não é sequer possível estimar o desvio padrão da referida população

Mas podemos fazer uma coisa…

- O teste T pode ser visto como uma variante heurística do teste Z em que se assume que o valor do parâmetro p(w1,w2) da distribuição binomial que rege o(w1,w2) é estimado directamente a partir da própria observação, o que nos leva a:

$$T = \frac{o(w_1, w_2) - e_{H_0}(w_1, w_2)}{\sqrt{o(w_1, w_2) \cdot (1 - o(w_1, w_2)/N)}} \simeq \frac{o(w_1, w_2) - e_{H_0}(w_1, w_2)}{\sqrt{o(w_1, w_2)}}$$

- Ou:

$$T = \frac{a - \frac{(a+b)(a+c)}{N}}{\sqrt{a}} = \sqrt{a} - \frac{(a+b)(a+c)}{N\sqrt{a}}$$

Z vs. T

$$Z = \frac{o(w_1, w_2) - e_{H_0}(w_1, w_2)}{\sqrt{e_{H_0}(w_1, w_2)}}$$

$$T = \frac{o(w_1, w_2) - e_{H_0}(w_1, w_2)}{\sqrt{o(w_1, w_2) \cdot (1 - o(w_1, w_2)/N)}} \simeq \frac{o(w_1, w_2) - e_{H_0}(w_1, w_2)}{\sqrt{o(w_1, w_2)}}$$

- Uma das características apontadas apontadas à medida T é a de que terá menos tendência que a medida Z para sobre-estimar do valor de associação (Evert, 2005).

Z vs. T

Bigrama "rede w_2"	o("rede", w_2)	Z	T
rede lusoswap	3874	6457.5	62.2
rede pública	4935	1122.1	70.5
rede electrossoldada	43	652.8	6.55
rede consular	242	412.7	15.5
rede globo	409	299.2	20.3
rede alargada	259	273.3	16.1
rede viária...	5	232.7	2.23
rede deverm	5	196.7	2.23
rede sismológica	20	177.9	4.47
rede bracara	56	157.1	7.50
rede pedofila	5	139.1	2.23
rede naric	13	123.6	3.60
rede sedimentológica	8	104.1	2.83
rede rodoviÃ¡ria	14	90.3	3.74
rede crossover	27	80.3	5.21
rede phasac	22	71.7	4.71
rede anti-pobreza	6	64.5	2.45
rede temática	81	57.0	9.23
rede craft	13	53.2	3.62
rede mpls	8	47.3	2.83
rede representada	43	42.8	6.71
rede departamentais	14	38.8	3.77
rede coaxial	14	35.3	3.78
rede 230v	7	32.7	2.66
rede gospel	10	29.8	3.19
rede imobiliÁria	9	27.3	3.03
rede terciária	6	24.7	2.47

Razão de Log-Verosimilhança (G^2)

- Os testes X^2, Z e T baseiam-se no pressuposto de que as variáveis aleatórias em causa possuem uma distribuição aproximável pela Distribuição Normal.
- Mas a aproximação da Distribuição Binomial (discreta) pela Distribuição Normal (contínua), só é válida para casos em a Variância da Distribuição Binomial ultrapassar um determinado limite ($\sigma > 5$)
- Isto quer dizer que para acontecimentos "raros" a Condição de Normalidade não se verifica e os resultados dos testes X^2, Z e T podem ser inválidos

Razão de Log-Verosimilhança (G^2)

- (Dunning, 1993) propôs um teste que não depende tanto da Condição de Normalidade
- O teste é conhecido como Log-Likelihood Ratio, ou numa tradução livre para Razão de Log-Verosimilhança.
- para simplficar, este teste é referido G^2
- Com o G^2 procura-se medir quanto é que um determinado evento é "surpreendente" ou não, mesmo que ocorra apenas uma vez
- Para casos em que o presuposto da normalidade se verifica, o teste G^2 é equivalente ao X^2

Detalhes matemáticos…

- São complicados. Envolvem a noção de Função de Verosimilhança, e uma dedução longa :(
- Por isso, vamos directos à fórmula que é compreensivel (mais ou menos)

$$G^2 = 2 \cdot [o(w_1, w_2) \cdot log\frac{o(w_1, w_2)}{e(w_1, w_2)} + o(w_1, !w_2) \cdot log\frac{o(w_1, !w_2)}{e(w_1, !w_2)} + o(!w_1, w_2) \cdot log\frac{o(!w_1, w_2)}{e(!w_1, w_2)} + o(!w_1, !w_2) \cdot log(\frac{o(!w_1, !w_2)}{e(!w_1, !w_2)}]$$

- É uma combinação linear dos logaritmos das razões entre as frequências *observadas* e *esperadas*

Formulação alternativa

- A formulação de G^2 mais habitual recorre aos dados da TC:

$$G^2 = 2 \cdot [a \cdot log(a) + b \cdot log(b) + c \cdot log(c) + d \cdot log(d) - (a+b) \cdot log(a+b) - (a+c) \cdot log(a+c) - (b+d) \cdot log(b+d) - (c+d) \cdot log(c+d) + (a+b+c+d) \cdot log(a+b+c+d)]$$

- É esta que é habitual encontrar na literatura

Exemplos de utilização do G^2

- Para ocorrências relativamente frequentes pode não haver "grandes" diferenças entre G^2 e X^2

$G^2(rede, local) = 17283.5$ $\quad X^2_{Yates}(rede, local) = 194398.1$

$G^2(rede, simples) = 1.76$ $\quad X^2_{Yates}(rede, simples) = 1.60$

- Mas para casos raros, a diferença pode ser muito significativa…

Casos Raros

- A diferença é muito grante para certos tuplos, com o teste X^2 a ***sobre-estimar*** frequentemente o valor da associação relativamente ao calculado pelo G^2

Bigrama "rede w_2"	o("rede", w_2)	G^2	X^2_{Yates}
rede lisboa	4	137.2	76.1
rede visa/	4	50.7	4416.2
rede 192.200.135.0	2	37.1	12183.5
rede 14.1.1-	2	33.3	8121.3
rede giganet	2	28.7	3478.8
rede end-to-end	3	24.3	318.9
rede redistribui	2	20.8	550.9
rede anodada	1	18.5	2707.1
rede ctc-ctc	1	18.5	2707.1
rede tática	2	14.7	118.2
rede cáritas	3	13.8	49.7
rede racks	1	12.5	337.5
rede conguito	1	11.3	192.4
rede fax*	1	10.3	116.7
rede serra	1	9.43	5.05
rede nexia	1	8.63	49.1
rede destaques	1	7.74	4.12
rede contrapõem	1	6.98	21.0
rede duplique	1	6.30	14.5
rede prelude	1	5.70	10.4
rede dossier	1	5.11	2.63
rede ambulatória	1	4.49	5.20
rede assuntos	2	4.03	2.30
rede cablagens	1	3.60	2.91
rede ouvindo	2	3.19	2.95
rede www.ies.pt	1	2.75	1.52
rede reactiva	1	2.30	1.01
rede estudante	2	1.93	1.00
rede carla	1	1.31	0.48
rede informar	1	1.01	0.32

Algumas conclusões

- Testes de Significância Assimpóticos
  - Partem de pressupostos assimptóticos:
    - Pressuposto de Normalidade (com mais ou menos importância)
    - Aproximam estatística discretas de Distribuições continuas
  - Fáceis de calcular: 1 fórmula - 1 iteração a partir de TC
  - Podem divergir para casos raros, com tendência para a sobre-estimação. Estes são os casos dificeis…
- Outros testes de Significância (de Verosimilhança e Exactos): fica para a proxima :)

6. Medidas de Associação baseadas em conceitos de Teoria de Informação

Mas apenas um exemplo…

Informação Mútua

- Uma das Medidas de Associação mais usadas em Linguística Computacional
  - Descrita em (Curch & Hanks, 1990)
- Relaciona a probabilidade de co-ocorrência entre dois elementos, X e Y, com a probabilidade individual de X e de Y

Informação Mútua (2)

$$IM(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

- A fórmula possui uma leitura muito intuitiva
- Se dois acontecimentos forem fortemente dependentes:
  - P(X,Y) >> P(X) P(Y) e IM >> 0
- Se H_0 se verificar:
  - P(X,Y) = P(X) P(Y) e IM = log (1) = 0
- Se X e Y tiverem tendência a ser exclusivos:
  - P (X, Y) < P (x) P (y) e IM < 0

Adaptando aos parâmetros da TC

- Usando as Estimativas de Máxima Verosimilhança de P(X,Y), P(X) e P(Y):

$$IM(x,y) = log_2 \frac{\frac{a}{N}}{\frac{a+b}{N}\frac{a+c}{N}} = log_2 \frac{a \cdot N}{(a+b)(a+c)}$$

- A aplicação é também directa sobre os parâmetros da TC

Críticas à Informação Mútua (1)

- tende a sobre-estimar o grau de associação quando os eventos são raros, isto é quando os valores de P(X) e de P(Y) são muito reduzidos
  - promoção de ocorrências raras - palavras muito pouco frequentes, erros ortograficos - em detrimento dos casos realmente mais significativos
  - para corrigir este efeito foi proposto um factor extra de correcção (Lin & Pantel 2002)

Críticas à Informação Mútua (2)

- Apesar de ser eficiente em reconhecer a existência de associações fortes, não é robusta no que diz respeito à ordenação podendo levar a alguns erros na comparação
- não faz o melhor uso da informação existente nas células b e c da TC que, por representarem contagem maiores, permitem estimativas mais precisas do que a estimativa fornecida pelo valor "a"

7. Algumas conclusões

Medidas de Associação

- É um assunto relevante para diversas tarefas de PLN
- Há imensas estratégias base para o desenvolvimento de MA, com diferentes fundamentos teóricos ou heurísticos
- Pela sua omni-presença é importante conhecer bem cada uma das MA, para compreender bem os impactos da sua utilização e não ter "surpresas

Bibliografia

- Church, K. & Hanks, P. Word association norms, mutual information, and lexicography Computational Linguistics, 1990, 16(1), 22–29
- Dunning, T.E. Accurate methods for the statistics of surprise and coincidence Computational Linguistics, 1993, 19(1), 61–74
- Evert, S. The statistics of word cooccurrences : word pairs and collocations Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2005
- Lin, D. & Pantel, P. Concept Discovery from Text COOLING 2002, 2002
- Sarmento, L. BACO - A large database of text and co-occurrences Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), 2006
- Smadja, F. Retrieving collocations from text: Xtract. Computational Linguistics, 1993, 19(1), 143-177