Cercare collocazioni e altre parole associate

Marco Baroni

23 ottobre 2004

1 Collocazioni & Co.

- Firth: You shall know a word by the company it keeps.
- Identificare sequenze di due o più parole fortemente *associate* è utile a moltissimi fini: estrazione di collocazioni (*strong tea* vs. *powerful tea*), terminologia (*dialisi peritoneale* vs. *dialisi eseguita*), espressioni idiomatiche (*buon vento* vs. *forte vento*), nomi propri e named entities (*King Kong* vs. *Enroll Here*)...

2 Misure statistiche di associazione

- Consideriamo il caso più semplice: bigrammi adiacenti non-etichettati.
- Il metodo:
  - Estrai lista di bigrammi da corpus.
  - Computa *misura di associazione* per tutti i bigrammi.
  - Ordina bigrammi secondo misura di associazione.
  - Cerca cose interessanti al top della lista.

3 La frequenza

- La misura di associazione più semplice è la frequenza.
- La tabella 1 riporta i 10 bigrammi più frequenti nel corpus *ep_it_97.tok*.
- Qual è il problema con la frequenza?

la commissione	7470
signor presidente	6058
che la	5452
per la	5428
della commissione	5314
unione europea	5146
che il	4457
per il	4073
di un	4041
e la	3925

Tabella 1: Bigrammi più frequenti in *ep_it_97.tok*

3.1 La frequenza dei bigrammi va calibrata con la frequenza degli unigrammi che li compongono

- La frequenza semplice non coglie un fatto importante, e cioè che il numero di volte che due parole capitano assieme va considerato in relazione al numero di volte in cui ciascuna delle due parole capita separatamente.
- Il fatto che *che* e *la* capitino frequentemente insieme (5452 volte) va interpretato alla luce del fatto che *che* e *la* sono entrambe parole molto frequenti (*che* capita 74954 volte e *la* capita 66019 volte nel corpus).
- La sequenza *unione europea* capita meno volte di *che la* in termini assoluti (5146 volte); ma d'altronde *unione* e *europea* hanno frequenze molto minori di *che* e *la* (*unione* capita 8954 volte e *europea* capita 8286 volte).
- Intuitivamente, il fatto che le volte in cui *unione* e *europea* co-occorrono costituiscano una proporzione consistente delle volte in cui i due unigrammi occorrono in generale è più significativo del fatto che, in termini assoluti, la sequenza *unione europea* sia meno frequente di *che la*.
- Questo modo di ragionare (considerare la frequenza di co-occorrenze di un bigramma in relazione alle occorrenze totali dei due unigrammi che lo formano) è alla base di tutte le *misure di associazione*.

4 Mutual Information

- La misura di associazione classica in linguistica computazionale è la *mutual information* (MI).
- Introdotta, come strumento linguistico, da: K. Church & P. Hanks. Word association norms, mutual information, and lexicography. ACL 1989, 76-83.
- La formula:

$$MI(w_1, w_2) = \log_2 \frac{\mathrm{P}(w_1, w_2)}{\mathrm{P}(w_1)\mathrm{P}(w_2)}$$

- Questa formula ha un'interpretazione precisa in *teoria dell'informazione*, dove quantifica l'informazione supplementare sulla possibile presenza di w_2 come seconda parola del bigramma una volta che sappiamo che la prima parola è w_1.
- Tuttavia, anche sorvolando su questa interpretazione, ci sono due maniere intuitive di pensare alla MI (ignorando per ora la trasformazione logaritmica):
  - La MI è il rapporto tra la probabilità di occorrenza di un bigramma in un corpus stimata empiricamente (contando quante volte il bigramma capita nel corpus) e la probabilità teorica che ci aspetteremmo se le due parole che lo compongono fossero indipendenti (data dal prodotto delle probabilità empiriche delle due parole).
  - La MI può anche essere vista come il rapporto tra la probabilità della seconda parola nel bigramma dato che abbiamo visto la prima parola e la probabilità della seconda parola indipendentemente dal contesto.
- Con entrambe le interpretazioni, dovrebbe essere chiaro che più alto è il valore della MI più è plausibile che le due parole siano associate.
- La prima interpretazione è ampiamente discussa negli appunti sulla teoria della probabilità.
- Quanto alla seconda interpretazione, si ricordi che la probabilità condizionale è data da:

$$P(w_2|w_1) = \frac{P(w_1, w_2)}{P(w_1)}$$

- Dunque, il rapporto tra la probabilità di w_2 dato w_1 e la probabilità di w_2 indipendentemente dal contesto è data da:

$$\frac{P(w_2|w_1)}{P(w_2)} = \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

- A parte il logaritmo, questa è proprio la formula della MI.

4.1 Calcolare la Mutual Information in pratica

- Come abbiamo visto nelle note sulla probabilità, possiamo ricavare la seguente formula per calcolare $P(w_1, w_2)/(P(w_1)P(w_2))$:

$$\frac{P(w_1, w_2)}{P(w_1)P(w_2)} = \frac{\frac{fq(w_1,w_2)}{N}}{\frac{fq(w_1)}{N}\frac{fq(w_2)}{N}} = \frac{fq(w_1, w_2)}{N} \times \frac{N^2}{fq(w_1)fq(w_2)} = \frac{fq(w_1, w_2)N}{fq(w_1)fq(w_2)} \quad (1)$$

- La MI è data dal logaritmo di questo valore.

- Come senza dubbio ricorderete:

$$\log A \times B = \log A + \log B$$

$$\log \frac{A}{B} = \log A - \log B$$

- Dunque, dall'ultima formula in (1) possiamo ricavare:

$$MI(w_1, w_2) = \log_2(fq(w_1, w_2)) + \log_2(N) - \log_2(fq(w_1)) - \log_2(fq(w_2))$$

- Tuttavia, di solito, ci interessa semplicemente *l'ordine* in cui i bigrammi estratti da un corpus vengono disposti dalla MI.

- In quest'ottica, possiamo tranquillamente ignorare la trasformazione logaritmica, visto che $log(a)$ è maggiore di $log(b)$ se e solo se a è maggiore di b, e dunque l'ordine in cui una lista viene disposta dai valori prima e dopo la trasformazione logaritmica sarà il medesimo.

- Ma, anche al di là delle motivazioni teoriche, l'uso di logaritmi tende ad avere due vantaggi pratici:
  - Evitiamo divisioni e moltiplicazioni che potrebbero portare a valori estremamente alti o bassi (un problema per i computer oltre che per gli umani...)
  - "Appiattiamo" i valori più alti, il che di solito rende i dati più leggibili/gestibili, e corrisponde all'intuizione che la differenza, per dire, tra una coppia che co-occorre 1001 volte e una coppia che co-occorre 1100 volte non sia altrettanto marcata che la differenza tra una coppia che capita 1 volta e una coppia che capita 100 volte.

- Provate a calcolare a mano la misura ricavata nell'equazione (1) per le coppie *che la* e *unione europea* sulla base dei dati dal corpus *ep_it_97.tok* riportati nella tabella 2.

bigramma	fq(w1, w2)	fq(w1)	fq(w2)	N
che la	5452	74954	66019	2915802
unione europea	5146	8954	8286	2915802

Tabella 2: Dati per calcolare la MI a mano

- (Non preoccupatevi del logaritmo.)

bigramma	fq(w1, w2)	fq(w1)	fq(w2)	N	MI
majo rilascia	1	1	1	2915801	6.46
gracias muito	1	1	1	2915801	6.46
aquesta solitud	1	1	1	2915801	6.46
ritt bjerregaar	1	1	1	2915801	6.46
scoop giornalistico	1	1	1	2915801	6.46

Tabella 3: Alcuni dei bigrammi con MI più alta in *ep_it_97.tok*

5 Mutual information e parole a bassa frequenza

- In teoria, la MI sembra una buona idea, ma in pratica ci troviamo di fronte al grosso problema illustrato dalla tabella 3, che contiene 5 delle (molte) coppie in cima alla lista ordinata per MI dei bigrammi in *ep_it_97.tok*.
- Come la tabella 3 suggerisce, la MI tende ad essere molto alta per parole che sono molto rare.
- Analizzando la formula usata per calcolare la MI, dovrebbe essere ovvio perché incontriamo questo fenomeno:

$$\log_2 \frac{fq(w_1, w_2)N}{fq(w_1)fq(w_2)}$$

- Se cerchiamo di determinare l'ordine relativo della MI per coppie prese dallo stesso corpus, possiamo lasciar perdere il logaritmo (vedi discussione in 4.1) e N, che sarà costante per tutte le coppie.
- Dunque, ci possiamo limitare ad analizzare:

$$\frac{fq(w_1, w_2)}{fq(w_1)fq(w_2)} \tag{2}$$

- Ovviamente, questo valore sarà più alto tanto più alto è il numeratore ($fq(w_1, w_2)$) e tanto più basso è il denominatore ($fq(w_1)fq(w_2)$).
- Le due quantità non sono indipendenti: $fq(w_1, w2)$ non può mai essere maggiore di $fq(w_1)$ o di $fq(w_2)$ (visto che la frequenza del bigramma w_1w_2 sarà contata come parte della frequenza di w_1 e di w_2).
- Dunque, nel "migliore" dei casi (le due parole di una coppia capitano *sempre* insieme), avremo che $fq(w_1, w_2) = fq(w_1) = fq(w_2)$.
- In questi casi, se usiamo f per la frequenza comune ai tre termini della formula originaria, (2) diventa:

$$\frac{f}{f^2} \tag{3}$$

- Ora, la crescita di f non è proporzionale a quella di f^2, che cresce più rapidamente, e quindi man mano che il valore di f aumenta la formula della MI produce valori più bassi ($1/1^2 > 2/2^2 > 3/3^2$...)
- La tabella 4 riporta il valore della formula (3) per vari f.

f	f/f^2
1	1
2	0.5
3	0.33
10	0.1
100	0.01
1000	0.001

Tabella 4: f/f^2 in funzione di f

- In concreto, questo implica che se due parole capitano mille volte nel nostro corpus, e tutte le volte co-occorrono, esse avranno una MI notevolmente minore di due parole che capitano una sola volta, e quella volta capitano insieme.
- Il fenomeno della "sopravvalutazione" delle parole a bassa frequenza persiste anche nei casi in cui $f(w_1)$ e/o $f(w_2)$ sono maggiori di $fq(w_1, w_2)$: per esempio, due parole con frequenza 6 che co-occorrono 4 volte avranno una MI più bassa di quella di due parole con frequenza 3 che capitano 2 volte (notate come in entrambi i casi la frequenza di co-occorrenza sia 2/3 della frequenza delle parole).
- Questo va contro la nostra intuizione: la misura d'associazione in tutti questi casi dovrebbe essere uguale, o addirittura più alta per le coppie più frequenti.

5.1 Un altro problema: mutual information calcolata su corpora diversi

- Tornando all'equazone (1), notate che il termine N rappresenta la dimensione del corpus usato per raccogliere le frequenze.
- A parità di occorrenze di un bigramma e degli unigrammi che lo compongono, l'equazione (1) darà un risultato più alto se N è più alto.
- Per esempio, se paragoniamo MI estratte da un corpus di 10000 parole a MI estratte da un corpus di 100000 parole per coppie che abbiano la stessa frequenza (sia come bigramma che in termini di unigrammi) nei due corpora dobbiamo per forza avere che:

$$\frac{fq(w_1, w_2)10000}{fq(w_1)fq(w_2)} < \frac{fq(w_1, w_2)100000}{fq(w_1)fq(w_2)} \quad (4)$$

- Questo risultato è piuttosto controintuitivo: a parità di frequenze, ci aspetteremmo casomai che fossero più significativi i valori raccolti da un corpus di più piccole dimensioni.
- Questo secondo problema è tuttavia in pratica meno importante di quello della sopravvalutazione delle parole a bassa frequenza, visto che è raro paragonare direttamente collocazioni estratte da corpora diversi.

6 Come risolvere il problema delle parole a bassa frequenza

6.1 Soluzione 1: filtrare le parole a bassa frequenza

- La soluzione più ovvia è quella di eliminare i bigrammi con frequenza inferiore a k dalla nostra lista.
- Scegliere un k abbastanza alto da filtrare coppie implausibili ma abbastanza basso da non perdere collocazioni buone è più un'arte che una scienza.
- La tabella 5 riporta i 25 bigrammi con la MI più alta in *ep_it_97.tok* se consideriamo solo i bigrammi con frequenza uguale o maggiore a 100.
- Come si vede, la situazione migliora nettamente (ma 100 è una soglia arbitraria: potremmo aver perso collocazioni importanti che co-occorrono meno di frequente).

6.2 Soluzione 2: Usare la Log-Likelihood Ratio o altre misure d'associazione

- La MI non è l'unica misura di associazione; ci sono altre misure, alcune delle quali sono più robuste di fronte al problema delle basse frequenze.
- Non entrerò nei dettagli di come vengono computate le altre misure (per saperne di più sulle misure di associazione, visitate il magnifico sito `http://www.collocations.de/`).
- Mi limito ad osservare che, in un modo o nell'altro, le misure più robuste di fronte al problema delle basse frequenze prendono in considerazione anche il valore assoluto o relativo della frequenza di cooccorenza di due parole.
- Per esempio, possiamo moltiplicare la MI di una coppia per la frequenza di cooccorrenza della coppia in questione:

$$fq(w_1, w_2) \times \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} = fq(w_1, w_2) \times \log_2 \frac{fq(w_1, w_2)N}{fq(w_1)fq(w_2)} \quad (5)$$

bigramma	fq(w1, w2)	fq(w1)	fq(w2)	N	MI
hong kong	164	167	164	2915801	4.24
socialdemocratici danesi	145	173	183	2915801	4.12
america latina	138	277	141	2915801	4.01
medio oriente	106	262	130	2915801	3.96
geneticamente modificati	151	286	197	2915801	3.89
effetto serra	114	411	114	2915801	3.85
regno unito	379	396	398	2915801	3.85
fonti energetiche	126	315	173	2915801	3.83
nazioni unite	272	427	276	2915801	3.83
com c	197	206	420	2915801	3.82
criminalità organizzata	169	345	235	2915801	3.78
van den	104	522	104	2915801	3.75
van velzen	121	522	121	2915801	3.75
carne bovina	225	496	249	2915801	3.72
gran bretagna	184	590	185	2915801	3.69
libro bianco	188	634	222	2915801	3.59
libera circolazione	332	503	509	2915801	3.58
h oggetto	197	232	680	2915801	3.56
maggioranza qualificata	173	779	191	2915801	3.53
partito popolare	130	465	242	2915801	3.53
prospettive finanziarie	101	268	346	2915801	3.50
libro verde	395	634	587	2915801	3.49
esseri umani	131	146	867	2915801	3.48
svolgerà domani	246	392	644	2915801	3.45
moneta unica	310	457	710	2915801	3.44

Tabella 5: Bigrammi estratti da *ep_it_97.tok* con $fq > 100$ e ordinati per MI

- Con questa formula, siccome teniamo conto anche della frequenza assoluta di cooccorrenza di due parole, le coppie con una bassa frequenza di cooccorrenza vengono sfavorite.
- Per esempio, nella tabella 6 ho calcolato il valore ottenuto con l'equazione (5) per gli stessi valori di f (dove $f = fq(w_1, w_2) = fq(w_1) = fq(w_2)$) usati nella tabella 4, assumendo $N = 100000$.

f	$f \times log_2(f \times 100000/f^2)$
1	16.6
2	31.2
3	45.0
10	132.9
100	996.6
1000	6643.9

Tabella 6: $f \times log_2(f \times 100000/f^2)$ in funzione di f

- Notate come con la formula in (5) i valori della misura d'associazione sono più alti per le parole con cooccorrenza più alta, in accordo con la nostra intuizione.

- Tra le misure alternative robuste rispetto al problema delle parole a bassa frequenza, una delle più note è la *Log-Likelihood Ratio* (LL) (Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19(1): 61-74 (1994)).
- La formula per calcolare la LL è meno intuitiva di quella usata per la MI, ma uno dei termini fondamentali è quello presentato in (5).
- Si può pensare alla LL come al (logaritmo del) rapporto tra la verosimiglianza delle frequenze estratte dal corpus se ipotizziamo che c'è una dipendenza tra w_1 e w_2 e la verosimiglianza delle medesime frequenze se ipotizziamo che non ci sia tale dipendenza.
- Le dieci coppie con la LL più alta in *ep_it_97.tok* sono riportate nella tabella 7.

bigramma	fq(w1, w2)	fq(w1)	fq(w2)	N	LL
signor presidente	6058	7452	9986	2915801	66373.3
unione europea	5146	8954	8286	2915801	52330.6
stati membri	3885	7771	4734	2915801	43983.9
x bc	2208	2247	2208	2915801	35753.4
la commissione	7470	66019	16519	2915801	35071.5
della commissione	5314	30333	16519	2915801	28946.2
dell' unione	3808	20308	8954	2915801	26432.6
parlamento europeo	2661	8477	5211	2915801	24809.2
signora presidente	2040	2721	9986	2915801	20549.8
si tratta	2067	22038	2211	2915801	19330.8

Tabella 7: Bigrammi ordinati per LL in *ep_it_97.tok*

- Qui abbiamo il problema inverso: c'è una tendenza a favorire coppie con parole molto frequenti, come *della* e *la*.
- Dunque, possiamo adottare la soluzione inversa a quella proposta per la MI, eliminando coppie in cui la frequenza di uno dei due termini è molto alta.
- La tabella 8 riporta le coppie con la LL più alta una volta che si eliminino le coppie in cui la frequenza di una delle due parole sia di 10000 occorrenze o più.
- Si potrebbero anche filtrare i bigrammi sulla base di un valore minimo per la LL ma ordinarli per MI, o filtrarli sulla base di un valore minimo di MI ma ordinarli per LL.
- Last but not least, si potrebbero calcolare sia MI che LL (e magari altre misure), e cercare collocazioni e altre sequenze interessanti in cima alla lista ordinata con entrambe le misure (visto che tendono ad identificare coppie diverse).

bigramma	fq(w1, w2)	fq(w1)	fq(w2)	N	LL
signor presidente	6058	7452	9986	2915801	66373.4
unione europea	5146	8954	8286	2915801	52330.6
stati membri	3885	7771	4734	2915801	43983.9
x bc	2208	2247	2208	2915801	35753.4
parlamento europeo	2661	8477	5211	2915801	24809.2
signora presidente	2040	2721	9986	2915801	20549.8
nell' ambito	1260	4602	1787	2915801	14472.3
quanto riguarda	1345	6634	2265	2915801	13606.7
nei confronti	990	3598	1068	2915801	13004.0
l x	888	954	2247	2915801	12660.1
onorevoli colleghi	990	1946	2303	2915801	11952.3
stati uniti	918	7771	938	2915801	10802.2
degli stati	1281	6471	7771	2915801	8993.2
diritti umani	684	2708	867	2915801	8846.9
all' interno	947	8618	1688	2915801	8827.6
fondi strutturali	612	1493	878	2915801	8491.9
lo sviluppo	944	5317	2912	2915801	8422.7
signor commissario	1005	7452	3173	2915801	8187.5
ad esempio	930	8382	2082	2915801	8134.6
d' accordo	903	4680	3446	2915801	7849.7
dell x	541	569	2247	2915801	7675.0
primo luogo	670	1932	1964	2915801	7553.1
regno unito	379	396	398	2915801	7247.6
conferenza intergovernativa	479	1305	544	2915801	7192.1
presidente onorevoli	858	9986	1946	2915801	7153.3

Tabella 8: Bigrammi estratti da *ep_it_97.tok* con $fq < 10000$ per ciascun componente e ordinati per LL

7 La selezione dei bigrammi

- Per ora, abbiamo assunto che i bigrammi da ordinare usando una misura d'associazione siano semplicemente tutti i bigrammi estratti da un corpus.
- Questa non è quasi mai una buona idea.
- Per esempio, è quasi sempre meglio escludere bigrammi che contengono function words (articoli, preposizioni, ecc.), ed eliminare immediatamente i bigrammi che capitano meno di un certo numero di volte.
- Potremmo anche cercare bigrammi che si conformano ad un certo modello: per es., se siamo interessati all'estrazione di termini tecnici, potrebbe essere interessante analizzare tutte le coppie di parole che capitano nella struttura: *X di Y*.
- Se il corpus è stato annotato morfosintatticamente (POS-tagged), potremmo estrarre solo le coppie che hanno una certa struttura morfosintattica: per esempio, coppie di forma NOME AGGETTIVO.
- Spesso esistono tools appositi per l'estrazione di coppie che si conformano ad un certo pattern, ma in generale non c'è praticamente nulla che non si

possa fare con gawk, sed, egrep, gli altri command-line tools ed un po' di fantasia.

7.1 Bigrammi non-direzionali a lunga distanza

- Si possono anche raccogliere bigrammi formati da parole non adiacenti, e senza considerare la direzionalità.
- Per esempio, potremmo raccogliere tutte le co-occorrenze di due parole in una finestra di 20 parole (i.e., ci possono essere al massimo 18 parole tra le due parole in questione), con una distanza minima di 3 parole (i.e., ci devono essere almeno 3 parole tra l'una e l'altra), e senza considerare l'ordine in cui esse capitano (*cane blah blah blah guinzaglio* e *guinzaglio blah blah blah cane* vengono considerate come due occorrenze dello stesso bigramma "cane guinzaglio").
- La tabella 9 mostra i 25 bigrammi con MI più alta che sono stati estratti in questa maniera dal corpus *ep_it_97.tok*.

coppia	MI
hong kong	3.96
razzismo xenofobia	3.78
gas serra	3.76
annuncio h	3.46
industriale monetari	3.30
uomo violazioni	3.16
umani violazioni	3.15
h interrogazione	2.91
socialdemocratici votato	2.89
oriente pace	2.80
trasporti turismo	2.79
all bc	2.77
derivanti reti	2.77
euro monete	2.76
fonti rinnovabili	2.74
difesa esteri	2.72
h n	2.70
energetiche energia	2.69
libera merci	2.67
giustizia interni	2.63
votazione voto	2.61
ricerca tecnologico	2.56
economici industriale	2.52
agricoltura rurale	2.51
ecu importo	2.50

Tabella 9: "Collocazioni a lunga distanza" in *ep_it_97.tok*

- Che tipo di coppie si trovano, in questa maniera?

8 Oltre i bigrammi

- Oltre i bigrammi...
- ... è difficile andare.
- Problemi di formulazione delle statistiche e, soprattutto, di data sparsity.
- Una soluzione: riformula il problema in termini di bigrammi.
- Per esempio, invece di cercare sequenze tipiche VERBO AGGETTIVO NOME, prima cerca collocazioni AGGETTIVO NOME, e poi VERBO (AGGETTIVO NOME), dove (AGGETTIVO NOME) viene trattata come un'unità.

9 UCS

- Utilities for Cooccurrence Statistics, `http://www.collocations.de/`
- S. Evert. The Statistics of Word Cooccurrences: Bigrams and Collocations. PhD Thesis, University of Stuttgart (2004).
- Vedi handout a parte.