Misure di Associazione e Parole Caratteristiche di un Corpus

Marco Baroni

29 aprile 2005

1 La teoria

- Torniamo alla forma condizionale della Mutual Information presentata nell'handout sulle collocazioni (come al solito, ignorando il logaritmo):

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

- Qui, la Mutual Information è interpretata come il rapporto tra la probabilità di incontrare la parola w_2 se abbiamo appena visto w_1 e la probabilità di incontrare la parola w_2 se non sappiamo nulla sul contesto (la stessa formula si ricaverebbe considerando la probabilità di w_1).

- In termini più astratti, possiamo calcolare il rapporto tra la probabilità di un evento A dato B e la probabilità di A in generale:

$$\frac{P(A|B)}{P(A)} = \frac{P(A, B)}{P(A)P(B)}$$

- Veniamo alla situazione in cui abbiamo due corpora, uno specialistico e uno di riferimento, e vogliamo cercare le parole più caratteristiche del corpus specialistico.

- Ora, l'evento A potrebbe essere l'evento che una certa parola che abbiamo pescato da uno dei due corpora sia *peptic*, e l'evento B potrebbe essere l'evento che la medesima parola sia una parola estratta dal corpus specialistico.

- Dunque:

$$\frac{P(w = \texttt{peptic}|corpus(w) = \texttt{spec})}{P(w = \texttt{peptic})} = \frac{P(w = \texttt{peptic}, corpus(w) = \texttt{spec})}{P(w = \texttt{peptic})P(corpus(w) = \texttt{spec})}$$

- In questo caso, la MI è data dal rapporto tra la probabilità che la parola sia *peptic* dato che sappiamo che è una parola presa dal corpus specialistico e la probabilità che la parola sia *peptic* indipendentemente dal corpus da cui è presa.

- Ovviamente, è anche possibile (ma, a parer mio, meno intuitivo) interpretare la MI come il rapporto tra la probabilità empirica (verificata sui dati) che la parola sia *peptic* e appartenga al corpus specialistico e la probabilità di co-occorrenza di queste due proprietà che ci aspetteremmo teoricamente assumendone l'indipendenza.

- Con entrambe le interpretazioni, ci aspettiamo che le parole tipiche del corpus specialistico abbiano una MI alta, ossia che la loro probabilità di capitare nel corpus specialistico sia più alta della loro probabilità di occorrenza indipendentemente dal corpus considerato.

- Stime delle probabilità:

$$P(w = \texttt{peptic}) = \frac{fq(\texttt{peptic})}{N_{spec} + N_{gen}}$$

$$P(corpus(w) = \texttt{spec}) = \frac{N_{spec}}{N_{spec} + N_{gen}}$$

$$P(w = \texttt{peptic}, corpus(w) = \texttt{spec}) = \frac{fq(corpus(\texttt{peptic}) = \texttt{spec})}{N_{spec} + N_{gen}}$$

- Anche la Log-Likelihood Ratio, come praticamente qualsiasi altra misura d'associazione, può venire usata per cercare parole fortemente caratteristiche di un corpus.

- Di nuovo, osserveremo con la MI un bias in favore delle parole rare e con la Log-Likelihood Ratio un bias in favore delle parole frequenti (e nel caso della MI dovremo probabilmente filtrare per frequenza minima, nel caso della Log-Likelihood Ratio per frequenza massima).

2 Cercare parole tipiche in pratica

```
# ottenere conteggi da cqp, per esempio:

[];

count by word > "temp.spec.fq.data";

oppure

count by word  %c > "temp.spec.fq.data";

# fuori da cqp, ripulire i dati:

grep -v ^# temp.spec.fq.data | gawk '{print $2 "\t" $1}' > spec.fq.data

# calcolare statistiche di occorrenza in corpus specialistico e generale

compute_corpus_comparison_stats_from_frequency_lists.pl
                              spec.fq.data gen.fq.data > spec.stats

# liste di frequenza di riferimento: possono essere utili quelle in
# shared_data/europarl_fq_lists (in versioni case sensitive e non)

# per esempio:

compute_corpus_comparison_stats_from_frequency_lists.pl
               spec.fq.data ~/shared_data/euparl-it.fq > spec.stats

# campi nel file in output (tab-delimited):

parola fq_spec fq_tot mi log_likelihood

# dove fq_tot e' la somma delle frequenze della parola nel corpus
# specialistico e generale

# usare gawk per filtrare, sort per ordinare -- ad esempio:

# ordina per MI le parole di almeno quattro catatteri che capitano
# almeno 50 volte nel corpus specialistico

gawk 'length($1)>3 && $2>49' spec.stats | sort -nrk4 | more

# ordina per log-likelihood le parole di almeno quattro catatteri che
# non capitano piu' di 400 volte nel corpus specialistico e salva le
# prime 100 in un file
```

```
gawk 'length($1)>3 && $2<401' spec.stats | sort -nrk5 | head -100 > ll.words
```