Misure di Associazione e Parole Caratteristiche di un Corpus

Marco Baroni

12 novembre 2004

1 Mutual Information

- 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 mininma, nel caso della Log-Likelihood Ratio per frequenza massima).

2 Cercare parole tipiche con UCS

- Per cercare collocazioni, forniamo come input a `ucs-make-tables` una lista di bigrammi.
- Siccome quando cerchiamo parole tipiche di un corpus non siamo interessati alla co-occorrenza di parole con altre parole ma alla co-occorrenza tra parole e corpora, il "trucco" è semplice: forniamo in input una lista di coppie parola-corpus.
- Per esempio, mettiamo che i due corpora siano:
  - Specialistico:

```
salve
io
sono
il
corpus
specialstico
ulcera
peptica
ulcera
duodenale
```

– Di riferimento:

```
io
invece
sono
il
corpus
di
riferimento
```

- Le coppie da dare in input a `ucs-make-tables` saranno:

```
salve   SPEC
io      SPEC
sono    SPEC
il      SPEC
corpus  SPEC
specialistico    SPEC
ulcera  SPEC
peptica SPEC
ulcera  SPEC
duodenale       SPEC
io      GEN
invece  GEN
sono    GEN
il      GEN
corpus  GEN
di      GEN
riferimento     GEN
```

- Dato un input di questo genere, l'output di `ucs-make-tables` avrà il seguente aspetto:

```
id           l1  l2     f   f1      f2        N
--  -----------  ----  --  ---  ------  -------
 2       ulcera  SPEC   5    9  125494  3519567
```

```
 3    duodenale  SPEC   2   10  125494  3519567
 5      peptico  SPEC   2   17  125494  3519567
```

- Per ciascuna coppia parola-corpus, `f` è la frequenza di occorrenza della parola nel corpus, `f1` la frequenza di occorrenza della parola in entrambi i corpora, `f2` il numero di parole nel corpus, e `N` il numero totale di parole in entrambi i corpora.
- Da questi dati, *ucs-add* può calcolare misure d'associazione quali Mutual Information e Log Likelihood Ratio, *ucs-select* ci aiuta a filtrare, *ucs-sort* a cercare le parole coi valori più altri, ecc.
- Il programma/wrapper `prepare_corp_comp_table.pl` prepara le tabelle in un sol colpo, buttando via le coppie parola-GEN, a cui di solito non siamo interessati (perché non cerchiamo le parole "tipiche" del corpus di riferimento).