TALN
Corpus-based computational linguistics

Seminario del corso di Trattamento Automatico del Linguaggio Naturale

(a. a. 2001 – 2002)

Daniele Barsocchi

Introduzione

Gli strumenti informatici ed i metodi statistici applicati alla linguistica hanno dato un forte impulso agli studi che mirano ad **analizzare quantitativamente il linguaggio**, da un punto di vista scientifico.

Un **corpus** (pl. corpora) è un'insieme materiale di enunciati su cui si fonda la descrizione grammaticale di una lingua. Spesso con corpus si indica una raccolta sistematica di testi, in genere selezionati per scopi precisi.

Nell'analisi automatica del linguaggio naturale l'utilizzo di dati estratti da corpora linguistici di dimensioni ragionevolmente grandi è spesso determinante per ottenere risultati affidabili e significativi. Sia le ricerche di linguistica tradizionale sia quelle di linguistica computazionale fanno spesso riferimento a queste informazioni.

Liste e classi di frequenza del lessico

Nell'ambito della ricerca lessicografica sono stati compilati numerosi dizionari che si richiamano alla frequenza d'uso delle parole, e che permettono di individuare le effettive **abitudini linguistiche** degli utenti di una data lingua.

Lista di frequenza del lessico: elenco dei lemmi (ed eventualmente delle rispettive forme) del corpus di riferimento, accompagnati dall'indicazione della frequenza d'uso.

Classi di frequenza del lessico: corrispondono a partizioni del lessico, in ordine di frequenza, composte da 500 lemmi.

Rapporto tra forme, classificazioni e lemmi

Il termine "parola" spesso risulta ambiguo.

Una stessa *forma superficiale* può avere svariate *classificazioni* (categorizzazioni) riconducibili allo stesso *lemma* o a lemmi diversi.

Rapporto tra forme, classificazioni e lemmi

Quindi ad una stessa forma possono corrispondere classificazioni diverse, riconducibili allo stesso lemma.

Tali classificazioni possono anche appartenere alla stessa categoria grammaticale (**P**art **O**f **S**peech), come il sostantivo "libertà" (invariante nel numero) o l'aggettivo "elegante" (invariante nel genere, o neutro).

LIBERTA' → sostantivo sing. sostantivo plur. → LEMMA libertà

ELEGANTE → sost. maschile sost. femminile → LEMMA elegante

Rapporto tra forme, classificazioni e lemmi

Ancora più evidente è il caso di molti verbi le cui tre forme singolari del congiuntivo presente sono omografe (che io/tu/egli mangi)

MANGI → congiuntivo pers. 1, congiuntivo pers. 2, congiuntivo pers. 3, indicativo pers. 2, imperativo pers. 3 → LEMMA mangiare

Frequenze di lemmi: osservazioni

E' importante sottolineare che esiste un grande divario tra le frequenze dei singoli lemmi, infatti nella realtà d'uso, le **parole vuote** (aventi soltanto valore grammaticale: articoli, alcune preposizioni e congiunzioni) e i termini con una scarsa specificazione semantica (fare, cosa, essere, ecc...) assumono un ruolo predominante, e rappresentano di fatto, nella comunicazione scritta ancor più che in quella orale, la quasi totalità del lessico utilizzato.

I 500 lemmi più frequenti coprono ben l'80 – 90% del totale occorrenze, e i restanti hanno frequenze talmente basse da rappresentare appena il 10 – 20% dell'intero lessico.

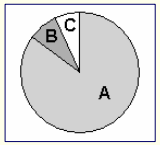
Graficamente, la situazione di un corpus di carattere generale, può essere così riassunta:

Frequenze di lemmi

A: percentuale delle occorrenze totali di termini del corpus riconducibili a lemmi della prima classe del lessico (cioè ai primi 500 lemmi più frequenti)

B: la percentuale delle occorrenze totali riconducibili a lemmi della seconda classe

C: la percentuale di occorrenze totali riconducibili a lemmi delle classi restanti



Frequenze di lemmi e forme

Poiché i primi 500 lemmi sono i più frequenti si potrebbe pensare che essi generino anche le forme con la frequenza più alta.

In effetti la tendenza generale è questa, tuttavia ci possono essere forme molto frequenti che hanno origine da lemmi poco usati e, viceversa, lemmi più frequenti da cui possono derivare anche forme con una bassa frequenza.

Infine considerazioni interessanti si possono fare anche in merito alla frequenza delle categorie grammaticali calcolata in riferimento all'occorrenza delle forme, confrontata con la distribuzione dei lemmi.

Frequenze di categorie grammaticali e termini

La categoria delle parole vuote (articoli, alcune preposizioni e congiunzioni: sono un insieme chiuso), ad esempio, copre una gran parte del totale delle occorrenze (risultano cioè molto utilizzate), eppure i lemmi di partenza non sono molti.

Considerando invece la categoria grammaticale dei verbi o quella dei sostantivi si verificherà la situazione opposta, perché si tratta di categorie più ricche e differenziate (parole piene: insieme aperto).

La frequenza dei termini di un lessico è legata al corpus di riferimento e può subire notevoli oscillazioni. Questa instabilità riguarda le parole tematiche, ed in particolare quelle con un'alta specificazione semantica. I termini generici, e ancor più le parole vuote, infatti, compaiono in misura pressoché costante in qualsiasi tipo di testo.

Ricorrenza e dispersione

Considerato un insieme limitato di testi, la discrepanza tra le frequenze di parole vuote e termini generici rispetto alle parole tematiche può non risultare in linea con l'andamento generale detto.

Questo rischio è tanto maggiore quanto più ristrette sono le dimensioni del campione e quindi la sua rappresentatività tematica.

Per ovviare agli inconvenienti di un corpus di riferimento finito è necessario considerare di ogni parola, oltre alla **ricorrenza**, la **dispersione**.

Una presenza consistente ma circoscritta, infatti, rivela l'influenza del contesto. In un corpus di testi tendente all'infinito un termine con tali caratteristiche ha meno probabilità di comparire rispetto ad un altro con la stessa frequenza assoluta ma con una diffusione maggiore.

Ricorrenza e dispersione: fattore di dispersione ed indice d'uso

A tal proposito sono stati definiti:

Fattore di dispersione: numero compreso tra zero e uno che indica quanto è uniforme la frequenza del lemma tra le varie fonti. (0 = lemma usato in una sola fonte; 1 = lemma presente con la stessa frequenza in tutte le fonti)

Indice d'uso: prodotto tra la frequenza assoluta del lemma e il suo fattore di dispersione (se la parola è ripetuta uniformemente nel corpus coincide con la frequenza).

Le Treebank

Una treebank è una collezione di frasi codificate secondo il syntactic tagging, vale a dire delle quali è stata data una descrizione sintattica.

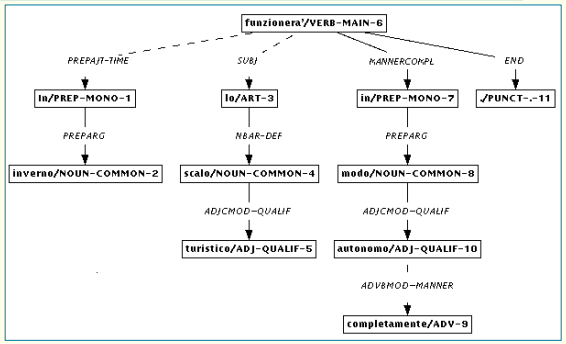
Le tecniche di analisi del linguaggio naturale che fanno uso di teorie statistiche sembrano offrire risultati assai più interessanti se applicate a corpora in cui **i dati sono corredati da una qualche esplicita rappresentazione delle informazioni morfologiche, sintattiche e semantiche**. Per questo motivo, molti sforzi sono stati impiegati negli ultimi anni nella realizzazione delle così dette treebank, banche di alberi sintattici di grandi dimensioni.

Esempio del Turin University Treebank

"In inverno lo scalo turistico funzionerà in modo completamente autonomo."

```
1  In (IN PREP MONO) [6;PREPAJT-TIME]
2  inverno (INVERNO NOUN COMMON M SING) [1;PREPARG]
3  lo (LO ART DEF M SING) [6;SUBJ]
4  scalo (SCALO NOUN COMMON M SING) [3;NBAR-DEF]
5  turistico (TURISTICO ADJ QUALIF M SING) [4;ADJCMOD-QUALIF]
6  funzionera' (FUNZIONARE VERB MAIN IND FUT INTRANS 3 SING)
   [0;TOP-VERB]
7  in (IN PREP MONO) [6;MANNERCOMPL]
8  modo (MODO NOUN COMMON M SING) [7;PREPARG]
9  completamente (COMPLETAMENTE ADV MANNER) [10;ADVBMOD-MANNER]
10 autonomo (AUTONOMO ADJ QUALIF M SING) [8;ADJCMOD-QUALIF]
11 . (#\. PUNCT) [1;END]
```


Esempio del Turin University Treebank



Il progetto SI-TAL

Sistema Integrato per il Trattamento Automatico del Linguaggio naturale

Si tratta di un progetto nazionale (in parte finanziato dal MURST) ormai in fase di conclusione, che aveva l'obiettivo di creare un'infrastruttura nazionale per le risorse linguistiche nel settore del trattamento automatico della lingua naturale parlata e scritta.

Due "realtà" pisane hanno partecipato a SI-TAL:

- il Consorzio Pisa Ricerche (CPR), che ha coordinato il cluster relativo alle treebank
- la Synthema di Pisa.

Tra i vari obiettivi, era prevista la creazione di una treebank sintattico-semantica per la lingua italiana. La verifica e la validità dei dati è stata realizzata dalla Synthema che ha riutilizzando le informazioni estratte per il miglioramento del sistema di traduzione "PeTra".

Conclusioni, il progetto EUROMAP

Si è deciso di terminare questa presentazione fornendo il link ad un progetto europeo nato sia per fornire servizi di informazione e collegamento, sia per lanciare sul mercato i risultati di progetti di ricerca e sviluppo di programmi nazionali ed europei nel settore delle tecnologie del linguaggio naturale. EUROMAP ha l'obiettivo di accelerare il trasferimento tecnologico dal mondo della ricerca al mercato, creando gruppi di interesse tra soggetti nuovi o già esistenti.

http://www.hltcentral.org/page-56.shtml

Dopo una prima fase di attività limitata al territorio nazionale, EUROMAP ha esteso il proprio raggio d'azione a livello transnazionale includendo i paesi dell'UE.

Conclusioni, il progetto EUROMAP

Il progetto EUROMAP ha pubblicato i risultati e le informazioni ottenute sul sito centrale (**H**uman **L**anguage **T**echnologies) all'URL:

http://www.hltcentral.org/

In particolare si segnala la possibilità di scaricare dalla rete, all'indirizzo http://www.cpr.it/euromap/downita.html, diversi documenti tra i quali un glossario bilingue, Inglese-Italiano, dei termini riguardanti le tecnologie del linguaggio naturale.

Riferimenti – Bibliografia

- D. Ambrogi, *Temi scolastici di ragazzi dell'Elba: lessico di frequenza, analisi morfosintattica, varietà sociolinguistica* - Tesi di laurea presso l'Università degli studi di Pisa, facoltà di Lettere e Filosofia, corso di laurea in Lettere Moderne, a. a. 1998-99.
- IBM Italia, *VELI, Vocabolario Elettronico della Lingua Italiana* – IBM Italia, 1989.
- Tullio de Mauro, *Guida all'uso delle parole* – Editori Riuniti Spa, Roma, I edizione ottobre 1997.
- U. Bortolini, C. Tagliavini, A. Zampolli, *Lessico di frequenza della lingua italiana contemporanea*.

Riferimenti – Siti consultati

Treebank:

http://www.cis.upenn.edu/~treebank/home.html	Pennsylvania University (Ingl.)
http://shadow.ms.mff.cuni.cz/pdt/pdt_05.html	Treebank per il ceco
http://www.coli.uni-sb.de/sfb378/negra-corpus/	NEGRA Treebank (tedesco)
http://www.ims.uni-stuttgart.de/projekte/TIGER/	TIGER Project (tedesco)
http://www.di.unito.it/~tutreeb/	Università di Torino (Italiano)

Vari:

http://www.cpr.it/	CPR
http://www.ilc.pi.cnr.it/	ILC, Pisa
http://www.synthema.it/	Synthema srl, Pisa
http://nlp.stanford.edu/links/statnlp.html	An annotated list of resources