

Microdispensa

Laboratorio di informatica specialistica per Scienze dell' Antichità, A.A. 2012/13, 2. semestre, Dr. Paolo Monella, lezione del 05/04/2013. Versione 2.0

Parsing

"Parsing" in inglese significa "analizzare" dal punto di vista linguistico-grammaticale.

Dunque si può fare a livello basilico (individuare la parte del discorso) o a livelli più avanzati (analisi morfologica, ovvero caso, genere numero etc.).

Quando al modo, si può fare in modo automatico (il computer fa il 100% del lavoro, senza aiuto umano) o semi-automatico (l'umano, aiutato dal computer, o viceversa; cioè il computer fa meno del 100% del lavoro).

Uno dei metodi che il computer ha per fare la sua parte di lavoro (che sia il 100% o meno) è l'uso di calcoli statistici (in che posizioni e in che 'compagnia' tende a comparire una certa parola).

I vari 'modi' si possono applicare sia al livello basilico sia ai livelli superiori.

Il crowdsourcing è semplicemente un modo di distribuire la parte umana del lavoro tra moltissime persone, piuttosto che affidarlo ad un solo operatore umano.

- Parsing su testi greco-latini: [Perseus](#)
- Outsourcing → crowdsourcing. Esempi di crowdsourcing:
 - [Wikipedia](#)
 - [Ancient Lives Oxford](#)
 - [Transcribe Betham](#)

WordNet

Synonyms set → synset

WordNet: uno per ogni lingua. [Inglese](#) (il primo), uno in [italiano](#), un progetto europeo per fare varie WordNet per varie lingue ([Multi WordNet](#)).

- sinonimia (scemo/stupido)
- antinomia (bello/brutto)
- iponimia (cane, gatto, Paolo → animale)
- iperonimia (animale → cane, gatto, Paolo)

Semantic technologies

Le tecnologie semantiche sono tecnologie che cercano di codificare/formalizzare non solo segni linguistici ma, in qualche modo, il loro significato (appunto, la 'semantica'). Il Web semantico (Semantic Web) è un progetto ambizioso che si va costruendo con grande lentezza e difficoltà: si tratta di un Web (sostanzialmente Internet) non fatto solo di parole, ma appunto di significati. Cioè in cui vengano codificati/formalizzati non solo i segni linguistici (le parole), ma, come dicevo loro, in qualche modo i loro significati.

Alcuni linguaggi 'semantici' sono RDF, OWL, Linked Data: sono usati per descrivere relazioni tra concetti (le cosiddette 'ontologies', 'ontologie', dal participio presente del greco *eimi*)

Information retrieval vs. (information extraction = data mining)

1. Information retrieval = recuperare formalmente informazioni formalizzate (es.: l'orario dei treni).
2. Information extraction (= data mining): il computer 'ragiona' sui testi (cioè li rielabora, come se li 'capisse') e tira fuori informazione nuova.

Nel caso 1, il computer si comporta come gli studenti che imparano a memoria e restituiscono esattamente le parole del libro. A una domanda 'di ragionamento', rispondono '... ma questo non c'è nel libro!'.

Nel caso 2, il computer si comporta come gli studenti che studiano in modo intelligente ed 'estraggono', appunto, informazioni dal materiale tramite il ragionamento.