Linguistica Computazionale 2004/05
Presentazione del Corso

5 ottobre 2004

1 Informazioni generali

Docente Marco Baroni

Email baroni@sslmit.unibo.it

Telefono 0453/374744

Homepage del corso http://www.e-learning.sslmit.unibo.it/CL0405

Orario lezioni Lunedì e venerdì 12:00/13:30 in aula Rosiello (terzo piano di Palazzo Montanari)

Ricevimento Lunedì 14:30/16:00 in aula Rosiello

2 La linguistica computazionale

- Aka: natural language processing, language engineering.
- Intersezione tra linguistica e informatica.
- Alcune applicazioni:
  - Traduzione automatica;
  - Lessicografia e terminologia;
  - Information extraction e information retrieval;
  - Text mining, web mining;
  - Filtri anti-spam;
  - Natural language interfaces;
  - Text-to-speech, speech recognition;
  - Correzione ortografica, predictive typing;
  - ...

- Metodi
  - Creazione, annotazione e analisi di corpora;
  - Pattern matching, espressioni regolari;
  - Probabilità, statistica;
  - Machine learning: supervisionato, non supervisionato, debolmente supervisionato;
  - Knowledge-intensive, knowledge-free, knowledge-poor;
  - Combinazione di metodi e knowledge-sources;
  - La WWW come corpus;
  - ...

3 Il corso

3.1 Perché studiare linguistica computazionale?

- Per sfruttare in maniera indipendente la grossa mole di materia prima linguistica oggi disponibile (in primis, la rete).
- Per comunicare con gli informatici con cognizione di causa.
- Alfabetizzazione informatica.
- Prospettive lavorative e accademiche.
- Per capire meglio le lingue e la linguistica.
- Per praticare metodo sperimentale.
- It's fun!

3.2 Prerequisiti

- Familiarità minima con il computer: più o meno, essere in grado di accendere e spegnere un computer e di compiere semplici operazioni (per esempio, fare una ricerca in rete con Google).
- Inglese (per fare le letture).
- Se i computer vi repellono e liste e numeri vi inorridiscono, il corso ovviamente non fa per voi.

3.3 Metodo e contenuti

- Enfasi su corpora e aspetti terminologici/lessicografici.
- Approccio pratico e "hands-on".
- Dopo lezioni introduttive dedicate a nozioni di base, gli studenti lavoreranno alla creazione di liste di termini e corpora di dominii specialistici, usando testo estratto automaticamente dalla rete e metodi che richiedono intervento manuale minimo.
- Nozioni teoriche e strumenti pratici verranno introdotti man mano che si rendono utili.
- Sarà possibile lavorare su lingue e domini scelti da studenti.

3.4 Programma

1. Comandi Unix per lavorare con corpora e lessici
2. Statistica lessicale: la legge di Zipf, i valori riassuntivi
3. Estrazione di collocazioni e altre parole relate
4. Testo semplice e altri formati per gestire dati testuali
5. La rete come corpus; spiders, Google API e BootCaT Tools
6. Estrazione di termini semplici e composti da un (web-)corpus
7. Annotazione morfosintattica e lemmatizzazione

3.5 Materiali

- Handouts e articoli, che posterò in formato elettronico sul sito del corso.
- Metterò gli articoli non disponibili in formato elettronico in una teca nella portineria di via Oberdan.

3.6 Valutazione

1. 4 compiti a casa (4 punti ciascuno)
2. Lettura attenta di due articoli, da discutere all'esame orale (2 punti ciascuno)
3. Report su esperimento di estrazione terminologica dal web, da discutere all'esame orale (10 punti)