Corso: Linguistica Computazionale (codice esame: 102943) Scienze della Comunicazione
Docente: Cristiano Chesi (chesi@media.unisi.it, 0577-234762) Specialistica in Linguistica
Ricevimento: Lunedì ore 11-13, stanza 327, CISCL, Palazzo S. Niccolò A.A. 2007-08

LINGUISTICA COMPUTAZIONALE

NOTE SUL CORSO

Programma

Il corso si propone di rendere consapevole lo studente delle necessità/difficoltà inerenti al trattamento automatico di lingue diverse in una prospettiva simbolica e linguistico-formale precisa. Alla fine del corso lo studente deve essere in grado di concepire e sviluppare (almeno da un punto di vista teorico) le componenti principali di un modulo di un sistema completo di traduzione automatica potenzialmente funzionante.
Il tema di quest'anno sarà la Traduzione Automatica (Machine Translation, MT). In generale, il problema della traduzione tra sistemi simbolici significativamente strutturati ha stimolato accesi dibattiti in vari campi collegati alla linguistica quali la semiotica (vedere Jakobson 1959 "On linguistic aspects of translation" o i numeri 85-87 di Versus: Quaderni di studi semiotici, 2000, per una panoramica sul tema), la logica e la filosofia (Quine 1964 "Word and object") e l'Intelligenza Artificiale (Hutchins 1986, "Machine Translation: past, present, future" scaricabile gratuitamente da web). Il corso si inquadra in questo dibattito fornendo un'introduzione teorica e pratica al tema della MT da un punto di vista fondamentalmente linguistico formale: nelle prime lezioni saranno approfondite le principali problematiche che un sistema di MT deve fronteggiare (analisi di un testo linguistico, modellizzazione della variazione cross-linguistica, rappresentazione del significato, generazione di espressioni linguistiche etc.). Verranno quindi descritte alcune soluzioni classiche (rule-to-rule, interlanguage etc.) proposte per risolvere questi problemi cercando di evidenziarne le criticità, le difficoltà implementative e le inadeguatezze rispetto alla teoria linguistica che propone un modello di variazione cross-linguistica basato su precisi parametri (Teoria dei Principi e dei Parametri, Chomsky 1981).
Nella corso verrà descritta un'architettura modulare ideale e verranno analizzate in dettaglio le sue principali componenti:
- corpora linguistici / database lessicali;
- analizzatore morfologico (robusto agli errori di spelling);
- analizzatore sintattico (robusto alle malformatezze sintattiche e alle elisioni);
- formalismo da usare per descrivere la variazione cross-linguistica tra le strutture frasali (principi universali e parametrizzazione);
- rappresentazione del significato / conoscenza;
- componente di generazione di espressioni ben formate.

Per seguire il corso non è necessaria nessuna particolare competenza informatica. È invece gradita una spiccata curiosità per tematiche legate alla Linguistica (Generativa), all'Intelligenza Artificiale, alla Psicologia Cognitiva ed eventualmente alla Semiotica.

Valutazione

Gli **studenti frequentanti** verranno valutati in base a: 1. *partecipazione in classe* (*20% del voto finale*); 2. una *tesina* (opzionale, massimo 10 pagine, *40% del voto finale*) in cui si evidenzia un qualche contributo/riflessione su un aspetto/modulo specifico di un sistema di MT; 3. una discussione orale (della tesina, in caso lo studente abbia scelto di farla) + 1 domanda su un testo a scelta tra quelli di base (*40% del voto finale*). L'esame potrà essere sostenuto anche senza la tesina, in tal caso l'orale verterà sugli appunti del corso e su un testo a scelta tra quelli di base.

Gli **studenti non frequentanti** dovranno sostenere un colloquio orale sul testo di Hutchins & Somers (1992) più un testo a scelta tra quelli di base o tra quelli di approfondimento. Anche gli studenti non frequentanti potranno scegliere di analizzare un modulo/aspetto di un sistema di traduzione automatica in una tesina (previa consultazione del docente) che verrà discussa all'orale.

Bibliografia di base

- appunti del corso, reperibili presso: *http://www.ciscl.unisi.it/* (seguire i link "*didattica*" > "*Linguistica Computazionale*")
- Hutchins & Somers (1992) *An introduction to machine translation* London: Academic Press, 1992 (cap. 1-9, più 3 capitoli a scelta) (*http://www.hutchinsweb.me.uk/IntroMT-TOC.htm*)
- Allegranza & Mazzini (2000) *Linguistica Generativa e Grammatiche a Unificazione*. Paravia scriptorium
- Jurafsky & Martin (2000) *Speech & Language Processing*. Pretience Hall, NJ (cap. 1-3,8-11,13-17, 20,21)
- Lenci, Montemagni & Pirrelli (2005) *Testo e Computer: Elementi di Linguistica Computazionale*. Carocci, Roma

Approfondimenti (altri testi di approfondimento potranno essere selezionati tra quelli suggeriti nelle dispense)

- Fong (1991) *Comptational properties of principle-based grammatical theories*. Ph.D. Thesis
- MacWhinney & Snow (1985) *The child language exchange system*. Journal of Computational Lingustics, 12:271-296
- Miller G. (1993) *Five papers on wordnet*.

Programma dettagliato

Lezione 1 - Lunedì 26 Novembre 2007, ore 17:00 - 20:00 (Aula 456)

Presentazione del corso: introduzione alla linguistica computazionale ed alla traduzione automatica - Obiettivi e organizzazione del corso, breve inquadramento interdisciplinare della linguistica computazionale (con particolare riferimento al Natural Language Processing, NLP). Traduzione Automatica (Machine Translation, MT), storia, prospettive e modelli. Alcuni esempi di MT.

Lezione 2 - Venerdì 30 Novembre 2007, ore 16:00 - 19:00 (Aula 401)

Strumenti linguistico-formali & informatici - Grammatiche formali e la gerarchia di Chomsky; grammatiche a struttura sintagmatica e trasformazionali; grammatiche ad unificazione; principi e parametri. Macchine di Turing (universali), concetto di computazione, dati, programmi, input e output; basi dati (corpora, database e strumenti per interrogarli); algoritmi.

Lezione 3 - Lunedì 3 Dicembre 2007, ore 17:00 - 20:00 (Aula 456)

Lessico, analisi morfologica e robustezza agli errori - Lessici computazionali, analisi morfologica e codifica informazioni linguistiche. Robustezza agli errori.

Lezione 4 - Mercoledì 5 Dicembre 2007, ore 16:00 - 18:00 (Aula informatica A, Lettere e Filosofia, Palazzo S. Galgano, Via Roma 47)

Laboratorio su espressioni regolari e analisi morfologica - Analisi morfologica con FST, recupero informazioni da corpora.

Lezione 5 - Lunedì 10 Dicembre 2007, ore 17:00 - 20:00 (Aula 456)

Parsing sintattico: introduzione ad alcuni algoritmi (parsing avanzato: P&P e minimalismo) - Regole di riscrittura, tagging, Top-down Vs Bottom-up parsing, chart parsing, left corner, la programmazione dinamica e l'algoritmo di Earley. Regole di riscrittura Vs. principi, P&P parsers (Pappi, Fong 1991).

Lezione 6 - Mercoledì 12 Dicembre 2007, ore 16:00 - 19:00 (Aula informatica A, Lettere e Filosofia, Palazzo S. Galgano, Via Roma 47)

Laboratorio di parsing - Scrivere grammatiche, valutare l'efficienza degli algoritmi di parsing, comprendere la struttura di un programma di parsing.

Lezione 7 – Lunedì 17 Dicembre 2007, ore 17:00 - 20:00 (Aula 456)

Rappresentazione della conoscenza, recupero di informazioni e disambiguazione - Ontologie, ambiguità, dal lessico alla rappresentazione della conoscenza (passando per wordnet). Classificazione documenti, riassunto, recupero informazioni con l'approccio Bag-of-Words.

Lezione 8 - Mercoledì 19 Dicembre 2007, ore 16:00 - 19:00 (Aula informatica A, Lettere e Filosofia, Palazzo S. Galgano, Via Roma 47)

Laboratorio su ontologie & disambiguazione - Esplorazione di wordnet & individuazione di idiosincrasie cross-linguistiche

*** Vacanze di Natale: 24 Dicembre - 6 Gennaio ***

Lezione 9 - Lunedì 7 Gennaio 2008, ore 17:00 - 20:00 (Aula 456)

Generazione - Sistemi reversibili, generazione diretta ed indiretta, generazione via transfer e via interlingua.

Lezione 10 - Mercoledì 9 Gennaio 2008, ore 16:00 - 19:00 (Aula informatica A, Lettere e Filosofia, Palazzo S. Galgano, Via Roma 47)

Laboratorio su sistemi di MT

Lezione 11 - Lunedì 14 Gennaio 2008, ore 17:00 - 20:00 (Aula 456)

Approccio simbolico o subsimbolico al processamento linguistico - Quando le reti neurali sono utili e perché, come funziona una rete neurale che "elabora" il linguaggio naturale (Simple Recurrent Networks), e che aspetti coglie. Come si applicano le reti neurali a problemi di traduzione.

Lezione 12 - Mercoledì 16 Gennaio 2008, ore 16:00 - 19:00 (Aula informatica A, Lettere e Filosofia, Palazzo S. Galgano, Via Roma 47)

Laboratorio sulle reti neurali - Costruzione di alcune reti con T-learn & cluster analysis.