Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

Presentazione del corso

Rocco Tripodi

rocco@unive.it

Informazioni utili

Lezione: giovedì 10:30 – 12:00

Ricevimento: giovedì 12:00 – 13:00

Esame: orale + progetto (facoltativo)

Appelli: I 10-01 → 19-01

II 20-01 → 28-01

Link uni: http://www.unive.it/nqcontent.cfm?a_id=68064&af_id=112525

Link lab: http://project.cgm.unive.it/

Materiale didattico

Lenci A., Montemagni S., Pirrelli V.,
Testo e Computer - Elementi di linguistica computazionale, Carocci Editore, 2005.

Corpora e linguistica in rete
a cura di M. Barbera, E. Corino, C. Onesti, Guerra Edizioni, 2007. (Pagine 25-88)

Jackson P., Moulinier I.,
Natural Language Processing for online applications:text retrieval, extraction and categorization, John Benjamins Publishing Company, 2007. (Pagine 1-68)

Testo e computer

Inquadramento storico della disciplina

Corpora

Codifica digitale

Marcatura

Metodi quantificativi

Ricerca nel testo

Annotazione linguistica

Struttura del nuovo medium

Corpora e linguistica in rete

Definizione del termine corpus:

I. Raccolta ordinata e completa di opere di autori (Devoto - Oli)
II. Campione prelevato a fini scientifici dal linguista (Devoto - Oli)
III. Raccolta di testi in formato elettronico uniformemente trattati

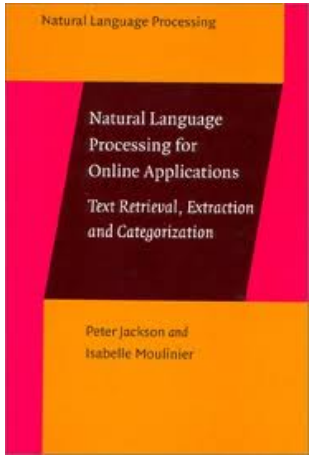
Web as a corpus?

Natura dinamica del linguaggio
Taglio temporale

Motori di ricerca linguistici

Bag of words VS approccio semantico
Esperimenti e prospettive

NLP for Online Applications



NLP: analisi del linguaggio scritto e parlato

Funzionamento dei motori di ricerca

Information Retrieval = indexing + search

Operatori di ricerca

Page Rank

Information Extraction

Text Categorization (Yahoo! Categories)

Text Mining (nuove informazioni)

Cos'è la Linguistica Informatica?

Una disciplina all'interno del Natural Language Processing* (NLP) che si occupa prevalentemente dei corpora e delle metodologie per le quali il computer può essere messo al servizio dell'indagine linguistica e letteraria. Corpus Linguistic.

Gli obiettivi rimangono uguali a quelli della Linguistica tradizionale; cambia però la metodologia di ricerca. Interdisciplinarietà.

*Lo studio del linguaggio naturale mediante la costruzione di modelli computazionali.

Di cosa si occupa la Linguistica Informatica?

Del trattamento automatico del linguaggio tramite la costruzione di modelli informatici atti a fornire una rappresentazione formale* del testo.

Individuazione delle regole astratte che descrivano la competenza della lingua posseduta da un parlante

*Cioè con sintassi e semantica definita in modo preciso

Quali sono le applicazioni classiche della Linguistica Informatica?

Corpora (prossime lezioni)

LIZ: Letteratura Italiana Zanichelli è una banca dati testuale che comprende integralmente 1000 testi letterari italiani.
TLIO: Il Tesoro della lingua italiana delle origini è un database testuale composto da circa 1.780 testi per circa 20 milioni di parole, tratte da scritti in lingua italiana prima del 1375. Basato sul corpus testuale dell'italiano antico dell'OVI (Opera del Vocabolario Italiano).
VELI: Il Vocabolario elettronico della lingua italiana (Tullio De Mauro). È costituito da circa 10.000 lessemi ordinati per frequenza nella lingua italiana.

Perché il testo viene trattato come i dati?

Età dell'abbondanza (overflow informativo)

Produzione massiccia di informazioni, non solo da parte delle aziende di telecomunicazione ma anche e soprattutto dai singoli (e-mail, social networks, pagine web, blogs, ecc).

→ Sistemi di ricerca delle informazioni (prossime lezioni)

Struttura del nuovo medium

Dal problema del chi parla al problema del chi ascolta
Comunicazione telematica basata sul narrow casting

→ Nuove forme di fruizione e presentazione delle informazioni testuali: newsmap

La Galassia Von Neumann

1945: *First Draft of a Report on the Edvac*

1976: M. McLuhan pubblica “La Galassia Gutemberg”

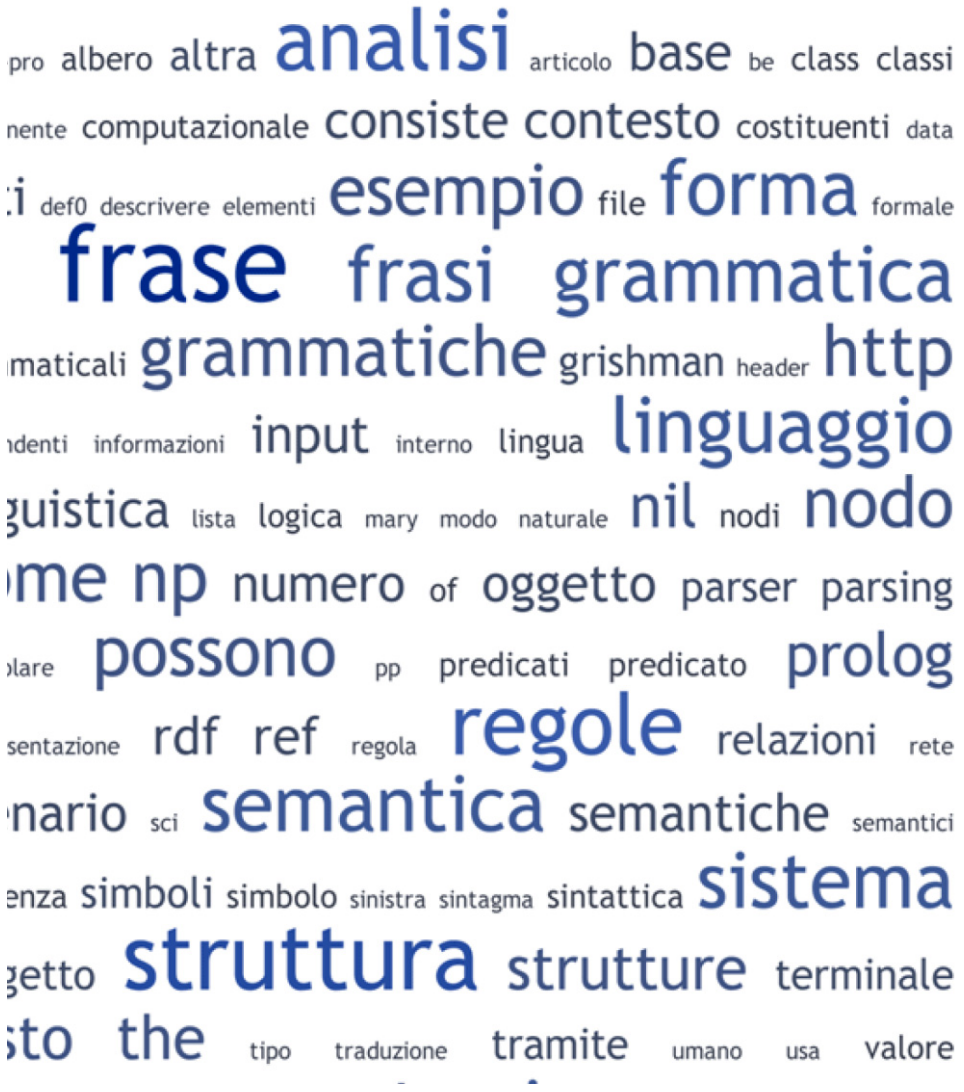
Dalla pergamena al libro
La stampa concentra l’esperienza sulla vista
Ogni tecnologia è una estensione e un potenziamento di un organo umano
Il medium è il messaggio
Media caldi e media freddi
Villaggio globale (tempo reale)

Paul Virilio: Il messaggio non è il medium quanto più la sua velocità

Nuove applicazioni della Linguistica Informatica – Text Visualization 1

Tag Cloud:

- conteggio delle parole
- Vecchie tecniche per nuove applicazioni.
- Il modello *bag-of-words:* il testo viene trattato considerando solamente l'occorrenza (diverso da frequenza) delle parole, tralasciando tutte le altre relazioni.
- Ordine alfabetico*, dimensione del font, colore.



Nuove applicazioni della Linguistica Informatica – Text Visualization 2

Word – Tree o albero dei suffissi

Si fa l'analisi della frase

Le relazioni contano

Ideato allo scopo di migliorare le performance degli algoritmi di ricerca. Il testo inserito in una struttura gerarchica (preparazione dei dati) è più facile da analizzare e consente di ottimizzare i tempi complessivi.

Questo tipo di visualizzazioni consente di ricercare una frase e consultarne tutte le occorrenze nei diversi contesti.

Many eyes

Nuove applicazioni della Linguistica
Informatica – Text Visualization 3

Phrase Net: le relazioni tra le espressioni

Fa ricorso alla tecnica delle espressioni regolari. Per esempio si può vedere quante volte un nome è legato ad un determinato aggettivo.
Ancora una volta le vecchie tecniche rivivono con l'ausilio delle applicazioni in rete. Con delle semplici analisi sintattiche di livello base si riescono ad estrarre informazioni complesse sui testi.

Many eyes

Nuove applicazioni della Linguistica
Informatica – Text Visualization 4

Trascrizione dei dialoghi

Speech recognition. Può essere utilizzata sia per rendere fruibili determinati contenuti da parte dei non udenti che per effettuare analisi testuali e sociolinguistiche.
Identificazione dei parlanti
Attribuzione delle frasi
Nessi causali

Naming names
Democratic Debate

Nuove applicazioni della Linguistica Informatica – Text Visualization 5

TextArc: visualizzazione sintetica dei testi letterari

Un intero libro in una pagina
Le parole più luminose sono quelle più usate, se compaiono al centro vuol dire che sono utilizzate con la stessa frequenza in tutte le parti del testo
Lista delle associazioni.
Ordinamento alfabetico/frequenza (concordanze)
Get thesaurus: di ogni parola offre il campo semantico
Chi conosce poco un'opera riesce ad individuare subito i personaggi e i tratti principali. Chi conosce l'opera può usare i dati per fare analisi sull'uso delle parole e il loro significato all'interno del testo.
TextArc

Nuove applicazioni della Linguistica Informatica – Text Visualization 5

OpenCalais: generazione di metadati “semantici”

- Named entities recognition
- Indicazione del tipo di “fatto”
- Tecniche di NLP per estrarre le classi di appartenenza e risolvere le anafore disseminate nel testo
- Creazione della versione RDF del testo sottoposto
- Funziona bene con i testi giornalistici, meno con i testi narrativi.
- Open Calais

Nuove applicazioni della Linguistica Informatica – Open Amplify 1

OpenAmplify:

Sistema di analisi testuale che si caratterizza per l'analisi dei sentimenti, degli stili e delle azioni
Funziona tramite web service
I tipi di informazione ricavata vengono definiti *signals, poiché si ritiene riescano a descrivere* indicazioni semantiche non direttamente deducibili dal testo; queste in particolare riguardano le attitudini degli autori dei testi stessi.

Topic Analysis: restituisce l'elenco degli argomenti del testo, includendo un grado di *polarità che indica la percezione (positiva o negativa) di un* determinato *topic; una guidance, che indica se sono richiesti o offerti consigli per il topic.*

Nuove applicazioni della Linguistica Informatica – Open Amplify 2

Action Analysis: restituisce l'elenco delle azioni trovate nel testo. Ogni azione è misurata seconda: un grado d'importanza (decisiveness) che indica come deve essere giudicata l'azione, una guidance (come accade per la topic analysis), e una temporality che indica quando l'azione si svolge.

Style analysis: restituisce delle indicazioni sullo stile di scrittura del testo in un tag denominato flamboyance (grado di ornamento del testo) e in un altro: slang, che tiene conto del registro linguistico impiegato.

Demographic Analysis: calcola approssimativamente l'età, il genere, e il grado di scolarità dell'autore e del lettore modello.

Nuove applicazioni della Linguistica Informatica – Open Amplify 3

Topics:
pig, wolf, flute, violin, violinist, house, Jimmy, pig, Timmy, wolf, Tommy, wood, brother, door, flute, violin, straw hut, wish, playing, fortune, story, stick, fun, flute player, fear, danger, build, brick, Jimmy,.
Actions:
open the door, walk, reach a nice wood, reach Jimmy, manage their work, not fear, get out of the woods, build a little house.
Demographics:
Age: Adult
Gender: Neutral
Education: Secondary
Style:
Slang: No Slang
Flamboyance: Somewhat Flamboyant

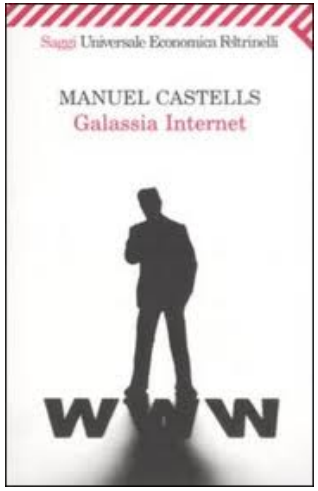
Struttura dei topic

<Topics>
<Domains> </Domains>
<TopTopics> </TopTopics>
<ProperNouns>
</ProperNouns>
<Locations> </Locations>
</Topics>.

Letture consigliate

Marshall McLuhan:
Gli strumenti del comunicare

Manuel Castells:
Galassia Internet



Giuseppe O. Longo:
Il nuovo Golem