Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 – 2011

Le origini

Rocco Tripodi

rocco@unive.it

Quadro storico - culturale

Inquadramento storico:

- Seconda metà del Novecento (Guerra Fredda)
- Utilizzo dell'informatica per gestire operazioni seriali (1950-1980) prevalentemente nei centri di ricerca e delle grandi aziende

Influenze culturali

- 1948: Claude Shannon pubblica *"A Mathematical Theory of Communication"*
- 1957: Noam Chomsky pubblica *"Syntactic Structures"*

Linguistica strutturale

De Soussure: arbitrarietà del linguaggio

non esiste corrispondenza tra nomi e oggetti designati ma tra il concetto (significato) che si ha dell'oggetto e un'immagine acustica (significante) e questo legame è arbitrario. Pensiero e suono sono due masse amorfe che la lingua articola in unità fonico-concettuali.
Es: albus/candidus Bianco opaco/Bianco Brillante

Rapporti sintagmatici: relativi alla combinazione (*articolo* → *nome*)

Rapporti paradigmatici: relative alle associazioni (*vecchio* evoca *nuovo*)

Campo lessicale: il vocabolario di una lingua è un sistema articolato in sottoinsiemi (campi) ognuno dei quali copre un'area concettuale (es: colori)

Hjelmslev: come i fonemi sono studiati in base ai tratti fonologici astratti così il significato può essere descritto come una configurazione di tratti semantici (es. scapolo = +maschio +adulto –sposato)

Categorizzazioni

Processo in base al quale le entità del mondo vengono raggruppate in classi per via delle somiglianze e delle differenze che si individuano in esse.

Labov (1977): la lingua traduce significati in suoni attraverso la categorizzazione della realtà in unità discrete

Una categoria è definita in base ad una serie di proprietà necessarie e sufficienti che colgono l'essenza dei membri (es: animale - bipede) e altre proprietà accidentali (es: biondo, laureato,ecc.)

Sono discrete, hanno confini chiari e definiti (non ammettono la vaghezza)

Sono internamente non strutturate, i membri sono tutti equivalenti

Gli approcci moderni alla categorizzazione ammettono la vaghezza e capovolgono tutti gli assunti elencati

Wittgenstein: le varie attività che chiamiamo giochi non condividono tutte lo stesso insieme di proprietà (somiglianze parziali)

Linguistica generativa 1

Obiettivo: descrivere le competenze linguistiche. Queste hanno quattro caratteristiche:

1. Generativa: la grammatica di una lingua non è un insieme di frasi ma una serie (finita) di regole
2. Fa parte della mente (stato mentale)
3. È in parte innata. La rapidità di apprendimento del linguaggio suggerisce la possibilità che gli essere umani abbiano un corredo genetico per tale funzione
4. La conoscenza innata riguarda solo le regole comuni ad ogni lingua (Grammatica Universale). Aspetti computazionali, che generano le costruzioni sintattiche a cui poi sono assegnate una veste fonetica e una interpretazione semantica

Linguistica generativa 2

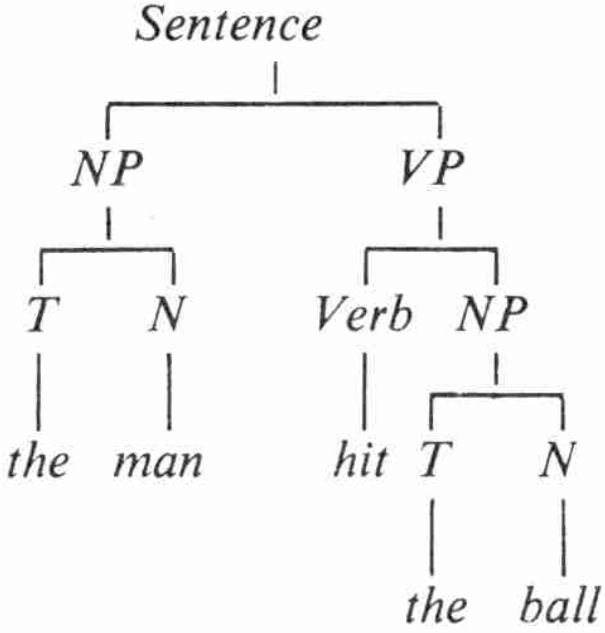
Formalizzazione delle lingue in regole dalle quali è possibile riconoscere le frasi ben formate nella lingua stessa

Grammatica formale: struttura astratta che descrive un linguaggio formale in maniera precisa. Sistema di regole che consente di produrre una serie (anche infinita) di sequenze (finite) di simboli.

Regole di riscrittura

Sentence
NP + VP
T + N + VP
T + N + Verb + NP
the + N + Verb + NP
the + man + Verb + NP
the + man + hit + NP
the + man + hit + T + N
the + man + hit + the + N
the + man + hit + the + ball

Struttura ad albero



Natural Language Processing 1

Far sì che le macchine esibiscano comportamenti linguistici intelligenti, fornendogli un modello computazionale che descriva il funzionamento del linguaggio

Modello *bottom-up* (Katz-Fodor)

Viene presa in input la struttura ad albero di una frase che viene processata (composizionalmente) a partire dai nodi terminali corredati da caratteristiche sintattiche e semantiche
Criticità: non tutte le frasi che non rispettano i vincoli delle features semantiche sono semanticamente inaccettabili
Es: Mi si è addormentata una gamba
Potrebbe essere considerata semanticamente anomala

Natural Language Processing 2

Modello *top-down*

- Viene preso in considerazione il contesto (Minsky)
- Nella mente ci sono gli schemi generali dei concetti (frames)
  - Es: automobile → veicolo, serve a spostarsi, ad una velocità, ecc
- Il frame è un nodo di una rete semantica in cui è condensata la conoscenza su un determinato oggetto.

Script: è un tipo di frame specializzato nella rappresentazione di storie ed eventi (situazioni stereotipate)

- Ruoli concettuali: attore che compie le azioni
- Regole concettuali: condizioni di entrata e uscita dalla script
- Primitive di azione: mangiare – pagare - uscire

Analisi narratologica

1928: Vladimir Propp dimostra che il repertorio delle fiabe di magia russe poteva essere ridotto ad un insieme di 31 *funzioni narrative* (danneggiamento,partenza,lotta,vittoria,ecc.)

Studio degli elementi invariabili del testo, fino ad individuare un'unica matrice strutturale abbastanza astratta da essere condivisa da tutti i testi narrativi

Schema narrativo canonico: è il modello, proposto da Greimas, che si propone di rendere conto della sintassi narrativa di tutti i testi.

Attanti: ruoli sintattici (aiutante, destinante,opponente)

Attori: personaggi caratterizzati che incarnano i ruoli attanziali

Analisi quantitativa

Si pone l'accento sull'uso che viene fatto del linguaggio

Chomsky faceva distinzione tra *competenza* (conoscenza linguistica) e *esecuzione* (uso della competenza nelle situazioni comunicative) tralasciando i dati quantitativi perché inerenti la sfera dell'*esecuzione*

Sinclair: senza l'analisi dei dati reperibili in ampie collezioni di testi (corpora) non vi oggettività nella ricerca in quanto mancano strategie di misurazione
Una gran parte dei significati delle parole sono depositate nella parte superficiale del linguaggio
Studio delle strutture in cui i termini occorrono

I corpora classici

La linguistica è una disciplina relativamente nuova

Nel secolo scorso si distingue come scienza autonoma.

Antichità: fino all'epoca medievale si registrano studi di grammatica (arte dello scrivere) oppure di dialettica e retorica (arte del dire) che proponevano norme da seguire relativamente alla lingua letteraria scritta e parlata. Segue l'era delle cosi dette "grammatiche filosofiche": lo studio del linguaggio e della grammatica era inglobato all'interno della filosofia.

Lo studio del linguaggio basato sulla raccolta sistematica di corpora di dati non è una conseguenza dell'uso dell'informatica, ma è conosciuto da molto prima che il computer nascesse. Questo sistema è stato utilizzato nell'Ottocento per la ricostruzione di lingue scomparse. L'attenzione sul lato quantitativo del linguaggio si registra anche nel opera di Cruden(1736), che pubblica le concordanze dell'Antico e Nuovo Testamento.

I primi passi

I corpora venivano usati prima dell'avvento dell'informatica per studiare la lingua o il lessico di un autore o opera

Padre Roberto Busa: studi pionieristici nel campo dell'analisi linguistica dei testi letterari. Durante gli anni '50 matura l'idea di creare un corpus in formato elettronico delle opere di San Tommaso D'Aquino (*Index Thomisticus*) con oltre 10 milioni di parole – Link

Brown Corpus: è il primo corpus ad essere stato creato per studiare un tipo particolare di lingua (inglese – americano degli anni '60)
1 milione di parole provenienti da più ambiti culturali

Raccolta dei dati

Cosa vogliamo ricercare determina la scelta della fonte.
giornali – dialoghi – trascrizione di lettere – SMS – ecc.

Le fonti possono essere:

Naturali: nel caso in cui i dati testuali vengano prelevati dal loro contesto naturale, conservando tutta la loro naturalezza.
Si intendono scoprire fenomeni inediti

Controllate: ottenute tramite la somministrazione di test creati ad-hoc dal linguista ad un gruppo di utenti.
Si intende verificare un determinato comportamento o uso.

Nel periodo attuale le fonti sono diventate più accessibili ed è anche aumentata la quantità di dati testuali.

Tipi di corpora - 1

Generalità: dipende dalla varietà di testi che sui quali è stato costruito. Si va dai *corpora specialistici* (dominio tematico ristretto, linguaggi settoriali o caratterizzati) ai *corpora generali* (usati spesso per lo studio di una lingua, per sviluppare un dizionario o una grammatica)

Modalità: corpora in lingua scritta, parlata, mista.

I corpora in parlato devono contenere le registrazioni audio del parlato (in database appositi) vengono usati per il riconoscimento e sintesi del parlato. Corpora multimodali (audio-visivo) per il tracciamento degli aspetti gestuali, mimici e emozionali

Tipi di corpora - 2

Lingua: corpus monolingua o multilingue.

- Corpus paralleli: contengono e tracciano le traduzioni tra i testi (vengono usati per la traduzione automatica)
- Corpus comparabili: non contengono la traduzione ma i testi sono stati scelti con gli stessi criteri

Cronologia: sincronici e diacronici

- Se i testi contenuti sono stati prodotti nello stesso arco temporale si fanno analisi sul tipo particolare di lingua in uso se invece si hanno testi raccolti in periodi lunghi si studia il mutamento linguistico

Tipi di corpora - 3

Integrità dei testi: raccolta di testi interi o di campioni

Codifica: in base alle etichette che descrivono il tipo di informazione (metadati) corpora annotati morfologicamente, sintatticamente, semanticamente, ecc)

L'unità di misura è il numero di parole (tokens). I corpora attuali si aggirano attorno ai 100 milioni di parole ma nonostante la loro grandezza rimangono realtà chiuse. Per questo Sinclair ha ideato i *monitor corpus* che sono collezioni aperte di documenti. Alla quantià deve corrispondere la qualità dei dati affinché un corpus possa essere considerato come campione (sottoinsieme di una popolazione) di una lingua

Rappresentatività

Il corpus deve essere una rappresentazione in scala ridotta della popolazione. La selezione deve tener conto di tutti i tratti e le variabili di una lingua

Diversi tipi di testo veicolano scopi comunicativi diversi ed ognuno ha un proprio stile e contenuto.

Bilanciamento: definizione dei confini spaziali e temporali della lingua, differenziazione del tipo di testi (stratificazione della popolazione).Solo così un corpus può essere considerato come un modello fedele del lessico e della grammatica di una lingua

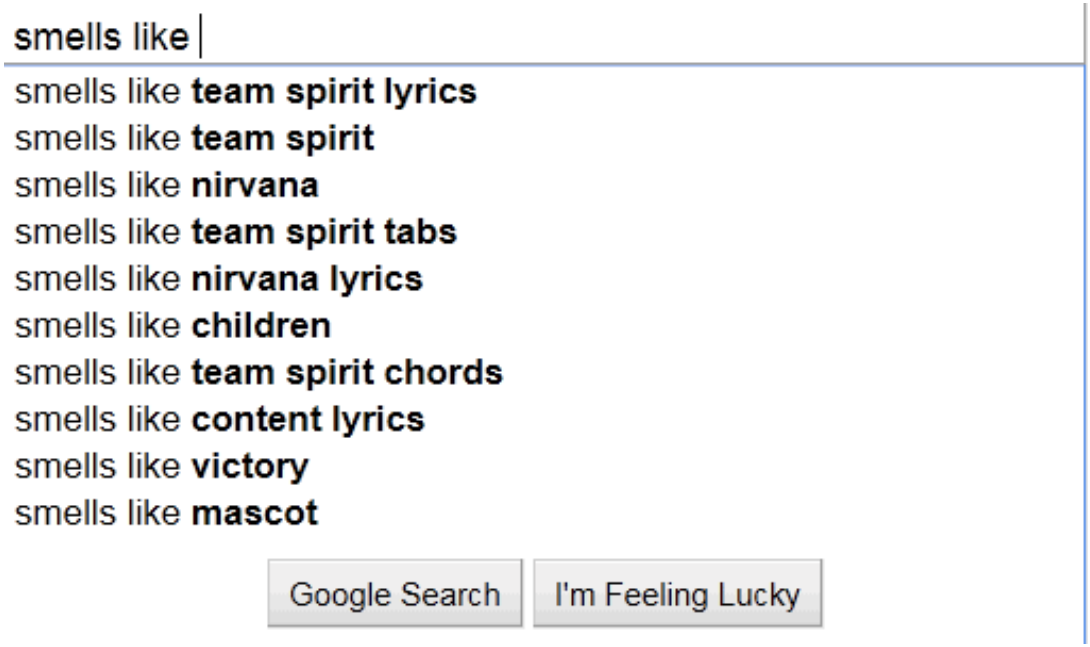
Chomsky: *"Nessun corpus è perfetto!"*
Arbitrarietà e parzialità nei metodi di rappresentazione scelti

Altri usi

Corpus come modello di riferimento

- Per valutare il funzionamento di algoritmi diversi

Corpus di addestramento: per raccogliere dati quantitativi sui fatti osservati dal corpus (parole, significati, categorie lessicali) per trasformare le regolarità riscontrate in informazioni per effettuare previsioni (es: significati ambigui, google translate)



Il web come corpus?

Web miniera d'oro per i linguisti

- 2003: 2000 miliardi di parole
- 75% delle pagine web indicizzate è in inglese contro lo 0,9% di pagine in italiano
- Multilingue, possibilità di corpora paralleli
- Attestazioni di errori ortografici
- Volatilità dei testi, dinamicità (consente di attestare immediatamente i neologismi es: messaggiare)

Bibliografia

Chomsky, N.
2002, *Syntactic Structures,* Walter De Gruyter, Berlin

Gambarara, D. (a cura di)
1999, *Semantica*, Carocci Editore, Roma

Greimas, A.
1968, *La semantica strutturale*, Rizzoli, Milano

Labov, W.
1977, *Il continuo e il discreto nel linguaggio*, il Mulino, Bologna

Propp, V.
1988, *Morfologia della fiaba*, Einaudi, Torino

Shannon, C.
"A mathematical theory of communication" July and October 1948 editions of the *Bell System Technical Journal*

Progetti di ricerca

Il giornalismo nell'era dei dati - Link

Generative Literature – Link

Motori di ricerca semantici e linguistici

Strumenti di traduzione automatica - Link