Linguistica Computazionale – Lezione 11

NLP avanzato: Information retrieval e traduzione automatica

Mercoledì 13 Aprile 2005

Cristiano Chesi, chesi@media.unisi.it

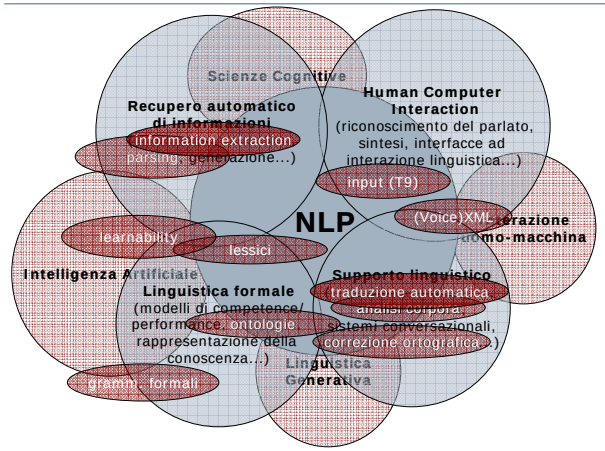
Introduzione al Parsing (sintattico)

- Indice
  - Information Retrieval
    - IR Vs. IE
    - approcci simbolici (rule-to-rule) e statistici (stand alone)
    - disambiguazione
    - rappresentazioni vettoriali di query e documenti
  - Traduzione automatica
    - variazione linguistica
    - metafora del transfer
    - interlingua ed ontologia

Letture, approfondimenti

- **Bibliografia essenziale**
  - Jurafsky & Martin (2000) *Speech & Language Processing*. Prentice Hall, NJ (Cap. 17, 20)
- **Approfondimenti**
  - Hutchins W.J. and Somers H.L. (1992) *An Introduction to Machine Translation*, Academic Press, London.
  - Gruber (1993) *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. In Formal Ontology in Conceptual Analysis and Knowledge Representation, Nicola Guarino and Roberto Poli, Kluwer Academic Publishers.

Mappa del NLP



Cosa abbiamo fatto, cosa dobbiamo ricordarci (in breve)

Lezioni 1 e 2

gramm. formali

- linguaggi regolari
- gerarchia di Chomsky

Lezione 3

correzione ortografica, input (T9), ubiquitus computing, (Voice)XML

- come si rappresenta un problema (definizione precisa problema, raccolta dati, classificazione, ricerca soluzioni/modelli formali)
- n-grammi, distanza minima
- vincoli hardware / linguaggi taggati (XML)

Lezione 4 (lab 6)

lessici, ontologie, analisi corpora

- corpora e databases
- espressioni regolari e SQL
- analisi morfologica (FSA, FST, stemming)
- lessici e ontologie (es. Wordnet)

Lezione 5 e 7 (lab 6, 8)

learnability

- modelli formali (apprendibilità nel limite)
- modelli cognitivi (es. P&P, triggers)
- modelli subsimbolici (reti neurali, TLearn)

Lezione 9 e 10 (lab 12)

parsing

- calcolo complessità
- algoritmi top-down, bottom-up
- programmazione dinamica (Earley
- unificazione, tratti e P&P

Lezione 11 (lab 12)

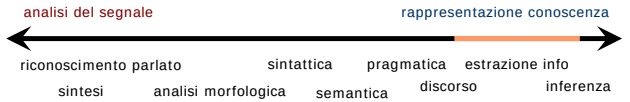
traduzione automatica, information extraction

- IR Vs. IE
- rappresenazione documenti, queries e distanze
- metafora del transfer

Dove siamo

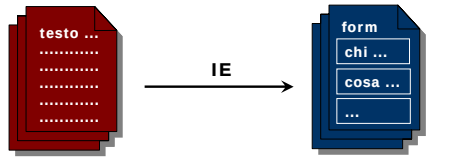
Definizione del problema

- data una query *q* identificare un gruppo di documenti significativi rispetto a *q*
- dove si colloca il problema:

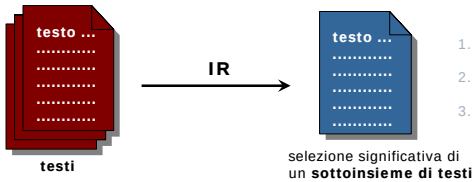


Information Extraction, IE Vs. Information Retrieval, IR

Information retrieval



1. recupero solo informazione rilevante
2. query specifica rispetto a template predefiniti
3. analisi linguistiche



1. recupero tesi
2. query generale
3. approcci statistici, keywords

Specifiche per l'IR

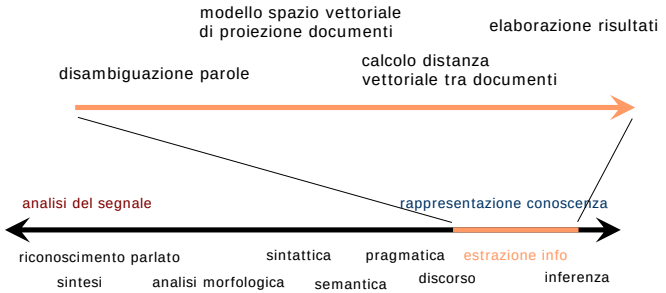
Information retrieval

- **assunzioni**:
  - le necessità dell'utente vengono **espresse da parole** (query testuali: sequenza di keywords)
  - l'unico contenuto semantico è veicolato dalle parole contenute nei documenti a cui si applica la query
- **problematiche classiche**
  - ambiguità lessicale
    (es. cane = animale, cane = parte metallica pistola)
  - rappresentazione dei documenti in uno spazio informativo (multidimensionale) definito

Tipologie di approccio al problema

- **rule-to-rule**
  - analizzatore morfologico > analizzatore sintattico > analizzatore semantico > IR
  - processamento lessicale-semantico
- **stand alone**
  - analisi indipendenti (testi = **bag of words**)
- entrambi gli approcci richiedono la definizione di un modello di spazio vettoriale su cui proiettare i documenti correlati alle queries fatte

Scomposizione del problema



Valutazione di un sistema di IR

- **Rilevanza**
  giudizio di un utente umano rispetto ad un testo recuperato in riferimento ad una query
- **copertura** (**recall**) =
  numero di documenti rilevanti recuperati :
  numero totale di documenti rilevanti nella collezione
- **precisione** (**precision**) =
  numero documenti rilevanti recuperati :
  numero di documenti recuperati
- Banco di prova per i sistemi di: TREC conference

Pre-processing

- **stemming**
  es. algoritmo di **Porter** (vedi lezione 5): serie di FST a cascata per l'inglese del tipo:
  - ATIONAL -> ATE (es. relational -> relate)
  - ING -> ε (talking -> talk)
- **pro e contro**
  - **iper-generalizzazione** (Krovetz 93) es. organization > organ, generalization > generic
  - **ipo-generalizzazione** es. matrices > matrix, European > Europe
  - in termini di **efficienza**:
    aumenta la copertura, ma diminuisce la precisione

Disambiguazione

- **possibili ambiguità**
  - **polisemia**
    Mario lava i ***piatti***
    Mario serve i ***piatti*** che ha preparato

    I Red Hot Chili Pepper hanno ***lanciato*** un nuovo disco
    Mario ha ***lanciato*** il bastone al cane

    Mario ha ***lanciato*** i ***piatti*** dalla finestra
  - **uso metaforico**
    Mario ha ***mangiato la foglia***

Disambiguazione (rule-to-rule)

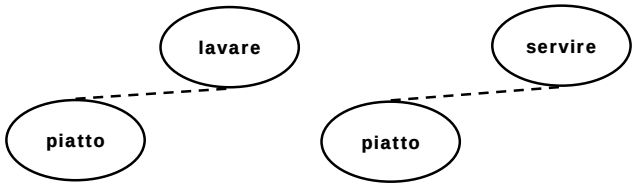
- l'approccio **semantico composizionale** ignora il problema dell'ambiguità
- in un **approccio integrato** (rule-to-rule) le metodologie di disambiguazione fanno uso principalmente di due strumenti:
  - **restrizioni selettive** (Katz e Fodor 1963, Hirst 1987)
    bloccare le violazioni di restrizione del tipo
    $\text{piatto}_{\text{portata}}$ > edibile, cucinabile …
    $\text{piatto}_{\text{scodella}}$ > lavabile …
  - **gerarchie di tipi**
    sfruttare dei vincoli derivanti dalla rete semantica formata dai concetti
    (es. Wordnet)

Disambiguazione (rule-to-rule)

- **restrizioni** sulla **selezione verbale**
  $\text{servire}_{\text{di portata}}$ > $\text{paziente}_{\text{[edibile]}}$
  $\text{lanciare}_{\text{di evento}}$ > $\text{paziente}_{\text{[eventivo]}}$
- **prerequisiti**
  analisi sintattica (almeno gruppi verbali e sottocategorizzazione) + specificazione di tratti semantici
- **problemi**
  - contesti troppo generali per poter applicare questo tipo di selezioni:
    es. che tipo di ***piatto*** ci raccomanda?
  - usi metaforici: violazioni produttive delle restrizioni semantiche (es. mangiare > $\text{paziente}_{\text{[edibile]}}$ in *mangiare la foglia;* oppure *bere un bicchier d'acqua*)

Disambiguazione (rule-to-rule)

- **vincoli derivanti dalla gerarchia semantica**
  associazione probabilistica tra il predicato e la classe che domina tale predicato
- **selectional association (Resnik 1997)**
  sfrutta la relazione di iperonimia in **WordNet** combinata con le informazioni recuperate da un corpus taggato



Disambiguazione (stand alone)

- **approccio statistico**
  l'idea è quella di acquisire informazioni semplicemente dalle parole (**bag of words**) piuttosto che da pre-analisi linguistiche
- **feature vectors**
  - **input:** parole target **(parole da disambiguare) +** contesto **(porzione di testo intorno ai target)**
    es. Marco lava i **piatti** con il detersivo
  - **rappresentazione tratti** in un vettore:
    - **collocazione** (finestra ad es. di 3 parole prima e 3 parole dopo il target)
      [Marco, N, lava, V, i, D, con, P, il, D, detersivo, N]
    - **co-occorrenza** (presenza di parole altamente frequenti in testi campione usati come training per la disambiguazione)
      guanti, lavare, acqua, detersivo > [0, 1, 0, 1]

Disambiguazione (stand alone)

- massima probabilità di un significato s (appartenente ad un insieme S di significati possibili), dato un vettore di tratti V:

$$\hat{S} = \underset{s \in S}{\operatorname{argmax}}\, P(s|V)$$

- secondo il teorema di Bayes:

$$\hat{S} = \underset{s \in S}{\operatorname{argmax}}\, \frac{P(V \mid s) \bullet P(s)}{P(V)}$$

- **Naive Bayes classifier**

$$P(V \mid s) \approx \prod_{j=1}^{n} P(v_j|s) \qquad \hat{s} = \underset{s \in S}{\operatorname{argmax}}\, P(s) \prod_{j=1}^{n} P(v_j|s)$$

Disambiguazione (stand alone)

- **Decision list classifiers**
  parole chiave collegate da operatori booleani costituiscono regole di classificazione:

regole		significato
detersivo, ¬pasta	⇒	$\text{piatto}_{\text{scodella}}$
lavare, lavastoviglie	⇒	$\text{piatto}_{\text{scodella}}$
pasta, carne	⇒	$\text{piatto}_{\text{portata}}$
servire, gustoso	⇒	$\text{piatto}_{\text{portata}}$
	...	

Disambiguazione (stand alone)

- **bootstrapping by seeds**
  l'idea è quella di iniziare l'apprendimento riducendo il training set ad un insieme limitato di esempi prototipici rispetto ad un determinato senso per poi cercare di generalizzare su un corpus più grande

testo prototipico		significato
can che abbaia non morde	⇒	$\text{cane}_{\text{animale}}$
caricò il cane prima di premere il grilletto della pistola	⇒	$\text{cane}_{\text{parte\_pistola}}$
	...	

Disambiguazione (stand alone)

- Apprendimento non supervisionato per **agglomerative clustering**
- **Approcci basati su dizionari di macchina**
  - sfruttare le definizioni delle parole come vettori di significato (Lesk 1986); accuratezza del 50-70%
  - considerare nel vettore anche le parole nella cui definizione si ritrova la parola target (Morris 1985)
  - utilizzare, ove presenti, anche i codici tematici (subject codes, Guthrie e al. 1991); accuratezza del 47-72%

Modello spazio vettoriale

- ogni vettore è rappresentativo dei termini ***t*** presenti nella query

$$\vec{q}_k = (t_{1,k}, t_{2,k} \ldots t_{n,k})$$

- e nei documenti

$$\vec{d}_j = (t_{1,j}, t_{2,j} \ldots t_{n,j})$$

(***n*** è il numero totale di termini nella collezione, ***t*** può essere presente, **1**, o assente, **0** nel vettore; Salton 1971)

- la **rilevanza** di un documento rispetto ad una query è data dal numero di termini ***t*** in comune

Modello spazio vettoriale

- un vettore può avere **valori pesati** anziché binari, che esprimono l'importanza di ciascun termine

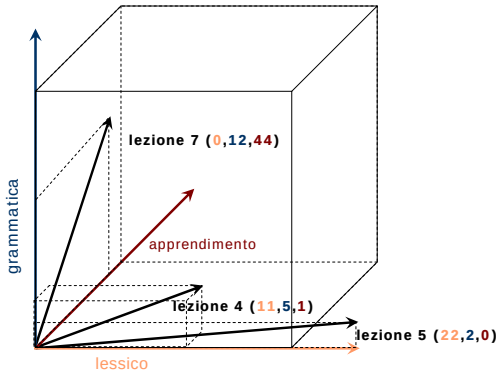
$$\vec{q}_k = (w_{1,k}, w_{2,k} \ldots w_{n,k})$$

$$\vec{d}_j = (w_{1,j}, w_{2,j} \ldots w_{n,j})$$

- matrice di termini per documento (**term-by-document matrix**)

	d_1	d_2	…	d_j
t_1	$w_{1,1}$	$w_{1,2}$		$w_{1,j}$
t_2	$w_{2,1}$	$w_{2,2}$		$w_{2,j}$
…				
t_i	$w_{i,1}$			$w_{i,j}$

Modello spazio vettoriale



Modello spazio vettoriale

- **Normalizzazione**
  conversione di tutti i vettori ad una lunghezza standard: ad esempio dividendo ogni dimensione per la lunghezza totale del vettore

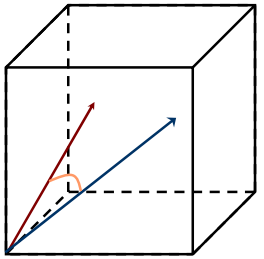
$$A = \begin{pmatrix} .80 & .92 & 0 \\ .40 & .9 & .20 \\ .10 & 0 & .79 \end{pmatrix}$$

lessico
grammatica
apprendimento

lezione 4
lezione 5
lezione 7

Distanza tra vettori

- **Prodotto scalare**
  la misura della similarità tra due vettori è data dal prodotto scalare tra i due vettori normalizzati



$$sim(\vec{q}_k, \vec{d}_j) = \vec{q}_k \bullet \vec{d}_j = \sum_{i=1}^{n} w_{i,k} \times w_{i,j}$$

coseno: da 0 (doc. ortogonali) a 1 (doc. identici)

Selezione del peso specifico dei relativi termini

- **frequenza**
  quante volte un termine appare in un testo (lista delle occorrenze)
- **importanza**
  meno volte un termine appare nell'intera collezione, più significativa è la sua presenza all'interno di un testo
- **stop list**
  lista di parole ad altissima frequenza (parole funzionali tipo articoli, povere di contenuto semantico) che vengono escluse dai vettori ("to be or not to be" > "not", Frakes and Baeza-Yates 1992)

Selezione del peso specifico dei relativi termini

- **usare WordNet per l'IR**

omonimia e polisemia riducono la precisione

iperonimi

word

iponimi

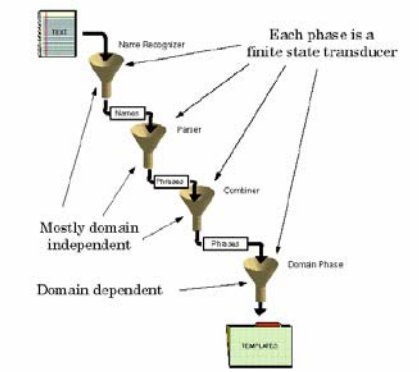
sinonimia e iponimia riducono la copertura

Selezione del peso specifico dei relativi termini

- usare il **feedback** per calcolare la **rilevanza** (Rocchio 1971)
  - **query [utente]**
  - **clustering automatico [sistema]**
  - **presentazione di documenti significativi per cluster**
  - **selezione di un documento**
  - **nuova query con nuovi termini pesati in base ai termini presenti nel cluster a cui appartiene il documento selezionato**

Finite State Automata-based Text Understanding System

- **FASTUS (Appelt & al. 1993)**



Il problema della traduzione

- **Eteronimo** = unità lessicale nella lingua di arrivo che ha forma diversa ma stesso significato dell'unità lessicale della lingua di partenza
- **Divergenze sintattiche**
  - **Divergenze strutturali**
    The man entered the room > L'uomo entrò nella stanza
  - **Divergenze tematiche**
    John likes Mary > A Gianni piace Maria

Il problema della traduzione

- **Divergenze sintattiche** (... continua)
  - **Divergenze categoriali**
    I'm scared > Ho paura
  - **Divergenze di inglobamento**
    To shelve a book > Riporre su uno scaffale un libro
  - **Divergenze lessicali**
    To take a shower > Fare una doccia
- **Ambiguità**
  vedi disambiguazione (slide 11-19)

Il problema della traduzione

Traduzione automatica

- **Divergenze semantiche**
  - **Divergenze di lessicalizzazione**
    towel > asciugamano (parola composta)
    private > soldato semplice
    perifrasi, circonlocuzioni …
    mollica > ?

Il problema della traduzione

Traduzione automatica

- **High quality translation Vs. Rough translation**
- velocizzare i tempi di traduzione umana
  (Computer-Aided Human Translation, CAHT o CAT)
- IR cross-linguistico
- filtraggio informazioni (spam)
- marketing
- traduzione di manuali tecnici (sottolinguaggi)

La variazione linguistica

Traduzione automatica

- **Tipologia** (Croft 1990, Comrie 1989)
  studio delle similarità e delle differenze sistematiche tra le lingue
- **Universali di Greenberg** (1931)
- **Parametri Chomskyani** (Chomsky 1981)
  es. parametro testa-complemento

La variazione linguistica

Traduzione automatica

- **variazione morfologica**

isolanti	<->	**polisintetiche**
Vietnamita, Cantonese		Esquimese
(1 morfema > 1 parola)		(molti morfemi > 1 parola)

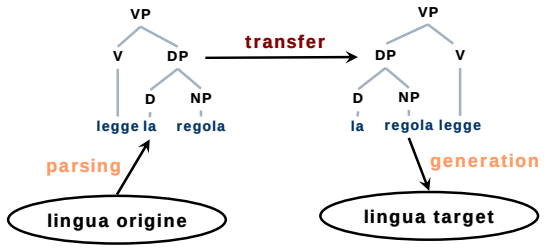
agglutinanti	<->	**a fusione**
Turco		Russo
(1 morfema > 1 tratto)		(1 morfema > molti tratti)

- **variazione sintattica**

SVO (inglese, italiano, francese …)
SOV (Indi, Giapponese …)
VSO (Irlandese, Arabo classico …)

Modello del Transfer

- **Conoscenza contrastiva**
  esplicitare le differenze tra le due lingue è il primo passo verso la traduzione.
  Da questo punto di vista occorre una ristrutturazione linguistica per conformarsi alle regole della lingua target

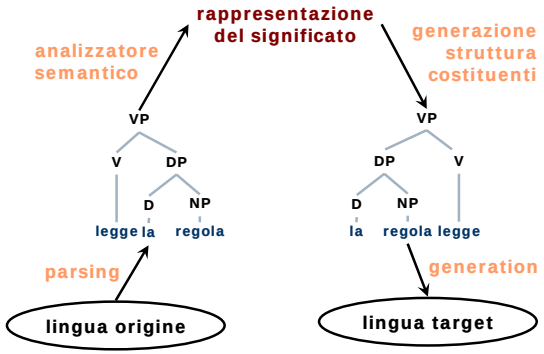


Esempi di Transfer con Context-Free Grammars

- Inglese ⇒ Italiano
  $DP \rightarrow D_1\ Agg_2\ Nome_3$ ⇒ $DP \rightarrow D_1\ Nome_3\ Agg_2$

- Giapponese ⇒ Inglese
  $DP \rightarrow relativa_1\ DP_2$ ⇒ $DP \rightarrow DP_2\ relativa_1$

Modello dell'interlingua



Ontologie

- **Concettualizzazione**
  astrazione (e semplificazione) delle relazioni tra oggetti e concetti in un dominio di conoscenza

- **Ontologia**
  specificazione dettagliata di queste entità e relazioni

  Ogni ontologia definisce un'insieme di **classi**, **relazioni**, **funzioni** e **oggetti** costanti all'interno di un dominio discorsivo, esplicitando un'**assiomatizzazione** in modo da **vincolare interpretazioni** ed **inferenze**

Ontologie

Rappresentazione della conoscenza

- **Knowledge Interchange Format**
  (**KIF**, Genesereth & Fikes, 1992)

```
(defrelation PHYSICAL-QUANTITY
    (<=> (PHYSICAL-QUANTITY ?q)
    (and (defined (quantity.magnitude ?q))
            (double-float (quantity.magnitude ?q))
            (defined (quantity.unit ?q))
                    (member (quantity.unit ?q)
                    (setof meter second kilogram ampere kelvin
                            mole candela)))
```


Prossima lezione

- Laboratorio su Parsing, ontologie e IR
  - scrittura grammatiche/lessico
  - benchmark algoritmi
  - esplorazione di wordnet
    - word sense disambiguation (per il IR)
    - sottocategorizzazione verbale (Levin 93)