

Natural Language Processing

Natural Language Generation
Natural Language Understanding

Informatica e lingue naturali - Isabella Chiari (2004) 1

Natural language processing

- **Definizione:** “lo studio dei sistemi informatici per la comprensione e generazione del linguaggio naturale” (Grisham, 1986: 4)
- Si occupa di tutti i livelli linguistici, ma soprattutto della **sintassi**
 - *Natural Language Processing* (NLP), o trattamento del linguaggio naturale
 - *Natural Language Understanding*, o *Natural Language Analysis*, o analisi del linguaggio naturale

Informatica e lingue naturali - Isabella Chiari (2004) 2

La linguistica computazionale nel paradigma del NLP

```

    graph TD
      LC[Linguistica computazionale] --> NLG[NLG  
Natural Language generation]
      LC --> NLA[NLA  
Natural Language Analysis]
      NLG --> NLG_sub["- generazione di frasi  
- produzione linguistica"]
      NLA --> NLA_sub["- analisi delle frasi  
- riconoscimento di strutture e gerarchie linguistiche a qualunque livello"]
  
```

Informatica e lingue naturali - Isabella Chiari (2004) 3

L'analisi del linguaggio

- **parsing**, la determinazione della struttura morfo-sintattica di una frase data
- Il **parsing** associa a una frase di una lingua naturale una struttura (per esempio una struttura ad albero) che analizza la frase da un qualche punto di vista
 - Parsing sintattico
 - Parsing morfologico
 - Parsing semantico, ecc..

Informatica e lingue naturali - Isabella Chiari (2004) 4

Le tappe principali del parsing

(modif. da Ferrari, 2002: 17)

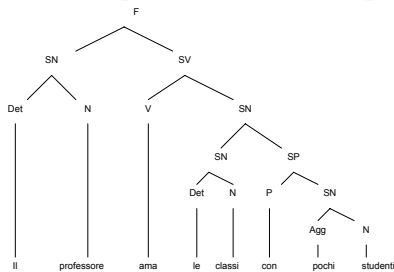
Informatica e lingue naturali - Isabella Chiari (2004) 5

Il parsing sintattico

- l'analisi consiste in una definizione dei sintagmi che compongono la frase nel loro ordine gerarchico
 - scompone la frase nei suoi principali sintagmi
- attribuzione alle parole delle funzioni grammaticali, dei ruoli tematici o logici
- **Output:** un diagramma ad albero che rappresenta le relazioni tra gli elementi della frase
- *Trebanks (Penn Tree-Bank)*

Informatica e lingue naturali - Isabella Chiari (2004) 6

Scomposizione e diagramma ad albero



Informatica e lingue naturali - Isabella Chiri (2004)

7

Il parser deve attribuire diverse etichette

- **etichette di struttura** che sono chiamate **simboli non terminali** (SV, SN, N, Agg)
- le parole che costituiscono la frase (*il, professore, pochi, classi, con*) sono dette **simboli terminali**
- dall'alto verso il basso troveremo sempre al primo posto l'etichetta di frase (F), successivamente troveremo una serie di etichette sintattiche di struttura (SN, SV, SP), nella penultima riga troveremo sempre le etichette delle categorie grammaticali (N, V, Agg, Det), sull'ultima riga sempre i simboli terminali, ossia le singole parole della frase (*le, con, ama*).

Informatica e lingue naturali - Isabella Chiri (2004)

8

Moduli del parser sintattico

- **regole language-dependent** che stabiliscono cosa può essere incluso in ogni tipo di sintagma (regole di struttura per i nodi non terminali)
 - SN → Det + N;
 - SV → V + SN;
 - SP → P + SN
- **vocabolario** per riconoscere a che categoria grammaticale appartiene ogni forma (regole di attribuzione per i simboli terminali)
 - V → *ama*;
 - Det → *il*;
 - N → *professore*

Informatica e lingue naturali - Isabella Chiri (2004)

9

Problemi del parsing sintattico

- omonimi testuali
- ambiguità sintattiche
- La mancanza di analisi semantica rende ambigue le anafore
 - *Gianni le ha parlato del suo cane*
- frasi non grammaticali (come gli accordi *ad sensum*)
 - *il gruppo di studenti andavano verso l'aula*
- Non sono escludibili le frasi sintatticamente ben formate ma non accettabili (semanticamente)
 - *idee verdi prive di colore dormono furiosamente*

Informatica e lingue naturali - Isabella Chiri (2004)

10

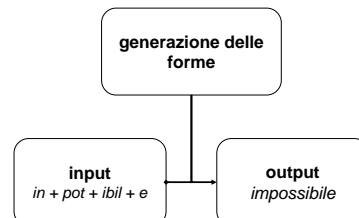
Il parsing morfologico

- **Generazione della struttura morfologica**
 - **Input:** morfemi
 - **Output:** la/e parola/e ben formata
- **Analisi (o comprensione) della struttura morfologica**
 - **Input:** la parola già formata
 - **Output:** la sua analisi morfologica o morfosintattica

Informatica e lingue naturali - Isabella Chiri (2004)

11

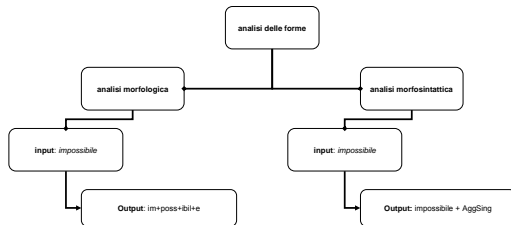
Generazione morfologica



Informatica e lingue naturali - Isabella Chiri (2004)

12

Analisi morfologica



Informatica e lingue naturali - Isabella Chiari (2004)

13

Morfologia a due livelli

- Si basa sulla fonologia generativa
- Prevede una serie **intermedia** di **livelli di rappresentazione** che integrano le regole morfologiche con quelle **morfofonologiche**

Informatica e lingue naturali - Isabella Chiari (2004)

14

applicazioni morfologiche del NLP

- la correzione ortografica (*spell checkers*) di documenti;
- la sillabazione di documenti;
- la lemmatizzazione;
- la preparazione dell'analisi morfologica per il *parsing* sintattico

Informatica e lingue naturali - Isabella Chiari (2004)

15

I correttori ortografici

- **Funzioni:**
 - segnalano i luoghi di possibile errore sia ortografico che di battitura
 - propongono suggerimenti sulle correzioni
- **Moduli del correttore:**
 - **dizionario** con forme flesse delle parole (*dormire*, ma anche *dormito*)
 - **regole di scomposizione morfologica** (il correttore elimina l'affisso, e confronta successivamente solo la radice con un'entrata di dizionario)
 - algoritmi che controllano le sequenze di lettere alfabetiche (**n-grammi**)
 - errori di battitura
 - Problema dei prestiti e degli errori di ortografia (che non violano la fonotassi)

Informatica e lingue naturali - Isabella Chiari (2004)

16

Problemi del correttore

- parole del lessico tecnico-specialistico
 - Possibilità di aggiungere al dizionario nuove forme
 - *Morfo, fono...segnalati* come errori
- Errore omografo con forme esistenti
 - *Nano* per *mano* non è segnalato come errore
 - Possibile integrazione con **parsing sintattico** (che almeno disambigua la categoria grammaticale e genere e numero)
 - **Parsing semantico**

Informatica e lingue naturali - Isabella Chiari (2004)

17

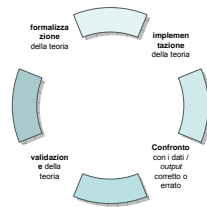
Procedimento per i suggerimenti del correttore

- dall'inizio della parola inserita propone le scelte più vicine (anagrammi, parole in cui solo una lettera cambia, parole in cui si deve inserire o cancellare una lettera)
- confronto degli n-grammi
- le probabilità di certi tipi di errori (derivanti per esempio dalla posizione dei tasti sulla tastiera del computer)
- la vicinanza fonetica (soprattutto per lingue in cui il rapporto tra grafia e fonìa è molto complesso, come l'inglese)
- l'analisi semantica

Informatica e lingue naturali - Isabella Chiari (2004)

18

La fonologia computazionale



Informatica e lingue naturali - Isabella Chiari (2004)

19

Scopi principali della fonologia computazionale

- produrre statistiche di vario genere (per *type*, per *token*, per tipologia di regole, ecc.);
- ordinare secondo diversi criteri il materiale fonologico;
- confrontare i modelli fonologici con le produzioni fonetiche (trascritte), ossia estrarre dati sulla cosiddetta interfaccia fonetica/fonologia;
- produrre esempi di determinate regole fonologiche, o di sequenze fonotattiche (sequenza di fonemi);
- confrontare diverse regole fonologiche fra loro e controllarne le esemplificazioni.

Informatica e lingue naturali - Isabella Chiari (2004)

20

Applicazioni del NLP

- **Correzione grammaticale**
- **Lessicografia**: per l'analisi dei dizionari *corpus-based*
- **Disambiguazione semantica**
- **Indicizzazione automatica**
- **Reperimento dati e informazioni** (*information retrieval*)
- **Traduzione automatica**
- **Riconoscimento vocale**

Informatica e lingue naturali - Isabella Chiari (2004)

21