

Parsing Robusto

Approcci Statistici e Simbolici per il
riconoscimento sintattico su larga scala

R. Basili, F. M. Zanzotto

Dipartimento di Informatica, Sistemi e Produzione
Universita' di Roma Tor Vergata
Rome - Italy

Obbiettivi del corso

- fornire una panoramica di un ambito di indagine molto attivo ma scarsamente sistematizzato
- dimostrare le possibilità di un campo di ricerca già legato a diverse esperienze applicative ed industriali, mediante la descrizione di alcuni sistemi e prototipi

Sommario

- Robust Parsing
 - Obiettivi
 - Principi fondanti
 - Metodi
- Modelli Simbolici
- Modelli Quantitativi

Sommario (2)

- Alcuni Parser Robusti
 - Apple Pie
 - Link parser
 - Chunkers
 - Chaos/Karma
- Acquisizione Lessicale nel Parsing

Parsing: due significati

- ***Parsing*** rispetto a grammatiche formali

- CFG
- LFG
- TAG
- HPSG

- ***Robust parsing***, cioè *tagging*, *chunking*, *partial*, *shallow parsing*

Le due accezioni fanno riferimento a due paradigmi e due comunità di ricerca diversi

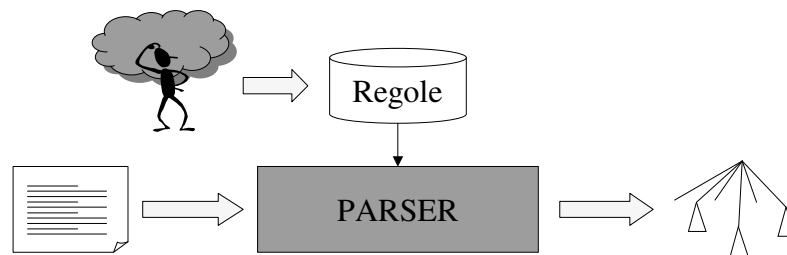
La visione tradizionale

Il ***Parsing*** (da *pars orationis*) ha lo scopo di determinare

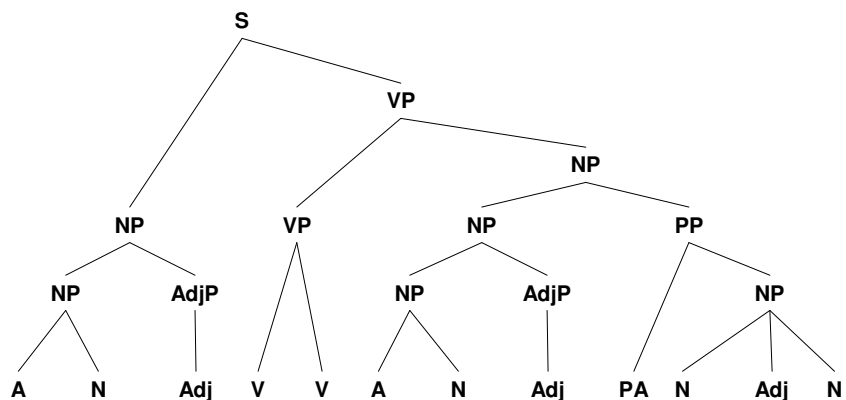
- la *grammaticalità* di una frase in ingresso ($\in L$)
- la(e) *struttura*(e) che giustificano la correttezza della frase nella grammatica
- tutte le *sottostrutture* che giustificano le due precedenti proprietà

Parsing sintattico

- Prospettiva: fornire una interpretazione sintattica delle frasi in linguaggio naturale utilizzando un *modello* del linguaggio.

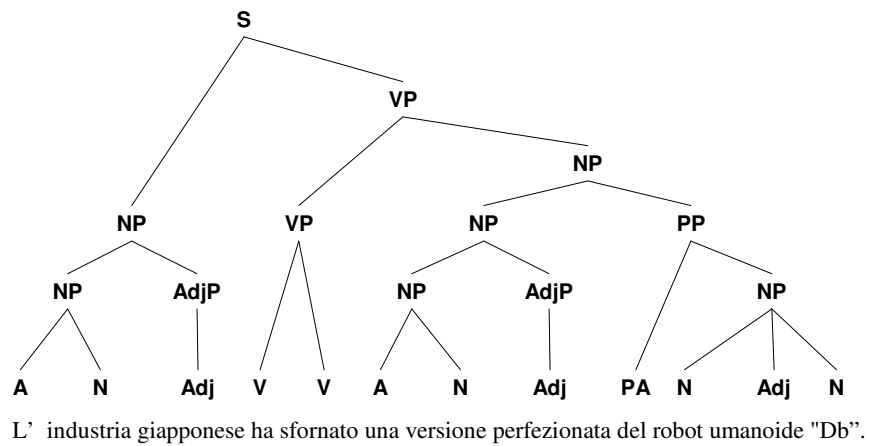


Grammatica, e alberi di derivazione



L' industria giapponese ha sfornato una versione perfezionata del robot umanoide "Db".

Grammatica e alberi sintattici

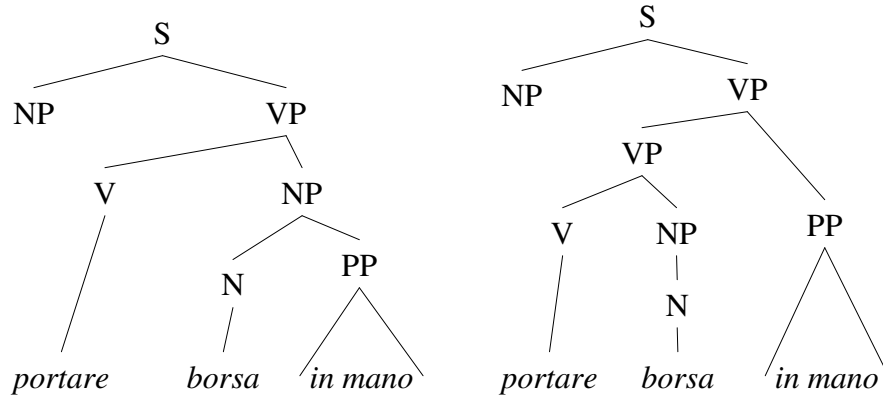


Parsing sintattico

- Può un modello descrivere tutta la fenomenologia del linguaggio naturale?
- Modelli simbolici vs. modelli statistici

Ambiguità

“portare borsa in mano”



Parsing sintattico “monolitico”

- Un formalismo grammaticale
CFG, LFG, HPSG, TAG
- Una grammatica
- Un algoritmo di parsing (nel caso migliore $O(n^3)$)

Parsing Sintattico: limiti

- Limitazioni derivanti dal muoversi all'interno di un livello di interpretazione sintattica (esempi)
 - *Mario guarda l'uomo con la bandiera*
 - *Mario guarda l'uomo con il cannocchiale*
 - *Mario guarda l'uomo con il microscopio*
 - *La vecchia porta la sbarra*
 - *Gianna mi ha prestato il giubbotto di pelle di Leonardo*
 - *Gianna mi ha prestato il giubbotto di pelle di leopardo*

Parsing Sintattico: limiti

- Ambiguità
 - Implicazioni computazionali
 - Qualità dei risultati
- Copertura dei fenomeni
 - fenomeni strettamente lessicali (*strong* vs. *powerful coffee*)
 - strutture extra-linguistiche
 - irregolarità
 - ungrammaticalità

Modelli quantitativi

- The traditional view
 - *"If we rank sentences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list there appears to be no particular relation between order of approximation and grammaticalness"*
(Chomsky, 1957).

Modelli quantitativi

- The traditional view
 - Do any n and ε exists such that:
$$\forall s \text{ grammatical}(s) \Rightarrow P_n(s) > \varepsilon$$
 - But
$$\forall s \text{ grammatical}(s) \Leftrightarrow \lim_{n \rightarrow \infty} P_n(s) > \varepsilon$$

Limiti degli approcci simbolici

- Language Acquisition
- Language Change
- Language Variation
 - Limiti di espressività degli approcci algebrici
 - Grammatiche + preferenze sono necessarie

Limiti degli approcci simbolici

- Legalità sintattica e ambiguità
 - *The a are of I*
 - *the cows are grazing in the meadows*
 - *John saw Mary*
 - Le grammatiche non sembrano poter esprimere le percezioni (graduali) dell'uomo
La probabilità può catturare tali gradi di grammaticalità

Limiti degli approcci simbolici

- Accettabilità ed Elaborazione
 - La accettabilità e' la valutazione (graduale) relativa alle strutture linguistiche contestualizzate
 - La separazione tra la grammatica ed il processore non ha evidenze neurobiologiche
 - I dati linguistici non si offrono ad analisi categoriche

Parsing Sintattico

- Debolezze
 - limitazioni del modello
 - limitazioni di un approccio monolitico
 - espressività
 - complessità
 - inadeguatezza nelle applicazioni

Parsing Robusto

- **Motivazioni**

- Copertura
- Efficienza
- Decisione Categorica
- Fenomeni non linguistici e non grammaticali
- Manutenibilità
- Portabilità

Robust Parser: criteri generali

- *Underspecification*
 - Restrizioni al (co)dominio
 - Rappresentazione
- *Decomposizione*
 - Cascade di riconoscitori
 - Utilizzo di strategie multiple
- *Lessicalizzazione*
 - grammatiche lessicalizzate
 - controllo lessicalizzato

Modelli Simbolici

- Permettono di fornire una definizione intenzionale di una proprietà che poi deve essere verificata (approccio top-down), ovvero tradotta nella sua versione estenzionale
- Vantaggi:
 - Possibile controllo umano del modello
 - Scrittura delle regole generali in maniera semplice
- Svantaggi:
 - Casi particolari non controllabili facilmente

Operazioni Standard

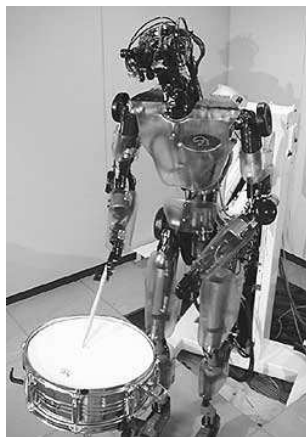
<i>operazione</i>	<i>input unit</i>	<i>output unit</i>
• part-of-speech tagging	word parola	word parola
• chunking	word parola	chunk sequenze
• clause bracketing	chunk sequenze	clause proposizioni

Modelli Quantitativi

- *Tagging* probabilistico
- Grammatiche con peso
- Induzione di preferenze sintattiche
- Induzione di preferenze lessicali

Robust Parsing: Applicazioni

UN ROBOT UMANOIDE CHE GIOCA A TENNIS



L'industria giapponese dei robot, una delle poche a non risentire della grave crisi economica, ha sfornato una versione perfezionata del robot umanoide "Db", fatto nascere due anni fa in un laboratorio vicino a Kyoto. Il robot, che sa anche giocare a tennis, è stato presentato alla stampa dal responsabile del progetto-pilota, Mitsuo Kawato, che studia da tempo l'attività cerebrale dell'essere umano. "Siamo sempre più vicini al sogno di un replicante umanoide in tutto simile a noi nei movimenti", hanno detto parlando del suo ultimo progetto.

Robust Parsing: Applicazioni

- Information Extraction
- Information Retrieval
 - Text Classification
 - Document and Passage Retrieval
 - Shallow Machine Translation
 - CLIR
 - Automatic Hypelinking
- Question Answering

Robustezza

- I modelli sintattici non possono essere esaustivi
 - la complessità del linguaggio naturale
 - limitazione ad un livello dell'interpretazione
- Gli algoritmi di parsing sintattico devono convivere con le eccezioni
 - le carenze del modello
 - le imperfezioni dei locutori

Osservazioni

- Se l'obiettivo sono applicazioni del tipo:
 - Information Extraction (IE)
 - Question Answering (QA)
 - Lexical Acquisition (LA)



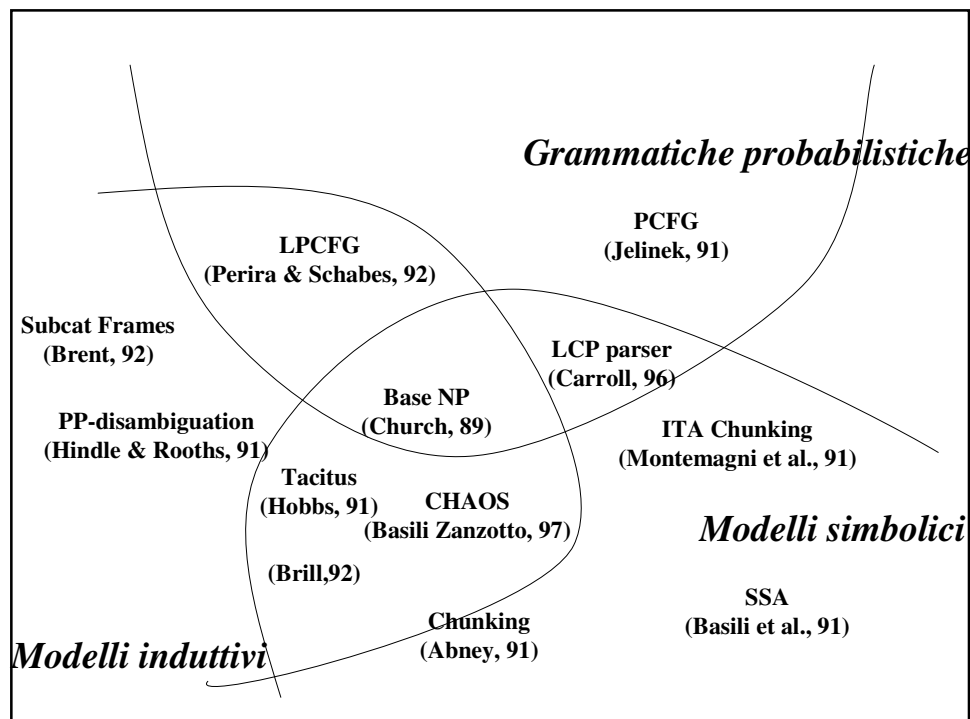
Robustezza = Capacità di fornire comunque qualche informazione

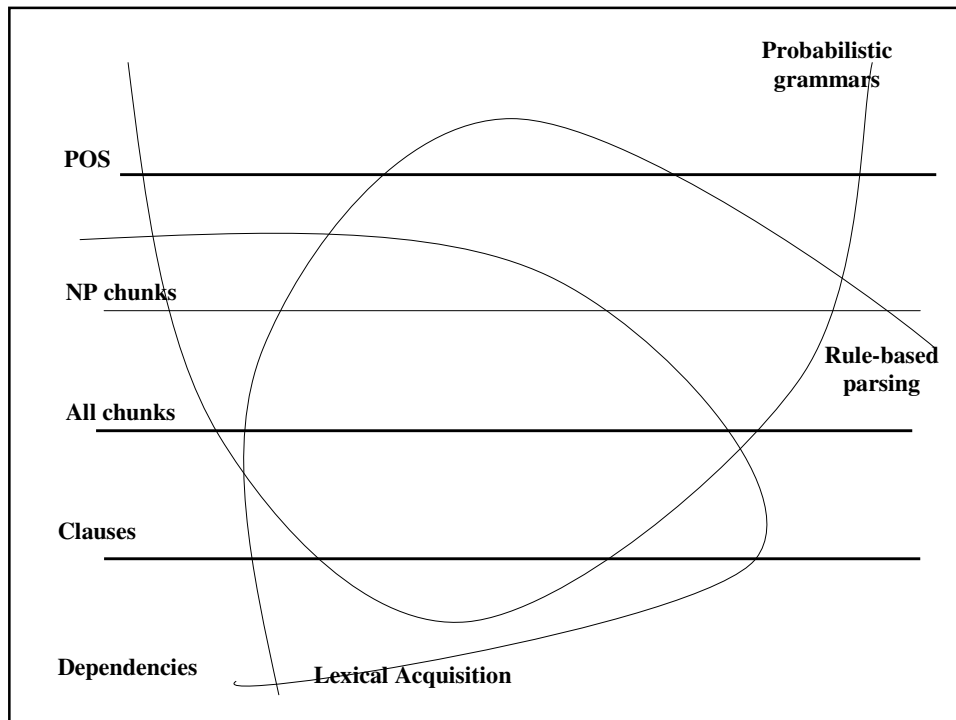
Parsing robusto

- Interessati:
 - “empiricisti”
 - Che lavorano su materiale reale: animo pragmatico
 - Che sono interessati alle prestazioni

Parsing robusto

- Decomposizione del processo in sottoprocessi
- Controllo lessicalizzato
- Ridondanza di sistemi che pur avendo competenze diverse lavorano sullo stesso processo





Parsing Robusto: Modelli Simbolici

- Decomposizione del processo in sottoprocessi
- Ridondanza di sistemi che pur avendo competenze diverse lavorano sullo stesso processo

Decomposizione del processo

- Ricerca di obiettivi (o rappresentazioni) intermedi raggiungibili (e utili)
 - Criteri psicolinguistici
 - Requisiti computazionali
 - Esigenze applicative
- È possibile trovare soluzioni approssimate per problemi più semplici?

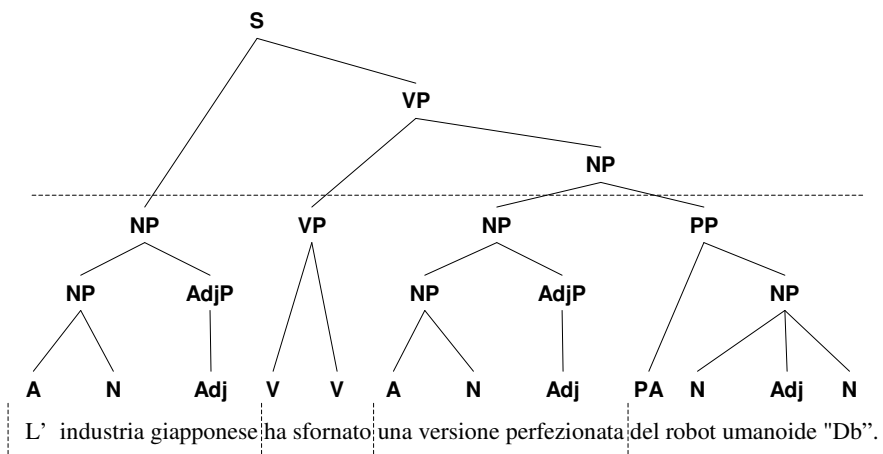
Decomposizione del processo

*|L'industria giapponese| dei robot| una| delle poche|
a non risentire| della grave crisi economica| ha
sfornato| una versione perfezionata| del robot
umanoide "Db",| fatto nascere| due anni fa| in un
laboratorio| vicino a Kyoto|*

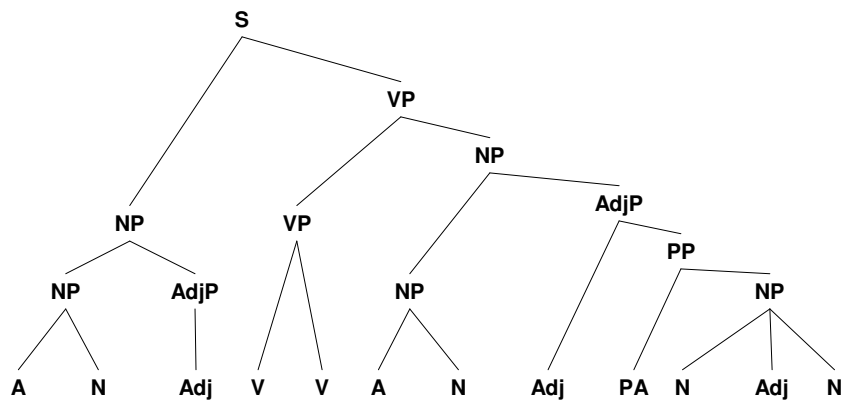
Decomposizione del processo

- Esiste un livello di aggregazione nel testo che si pone tra parole e frasi.
- Gli aggregati non si sovrappongono (i.e. non sono “ricorsivi”)

Chunking: esempio di stratificazione di un albero

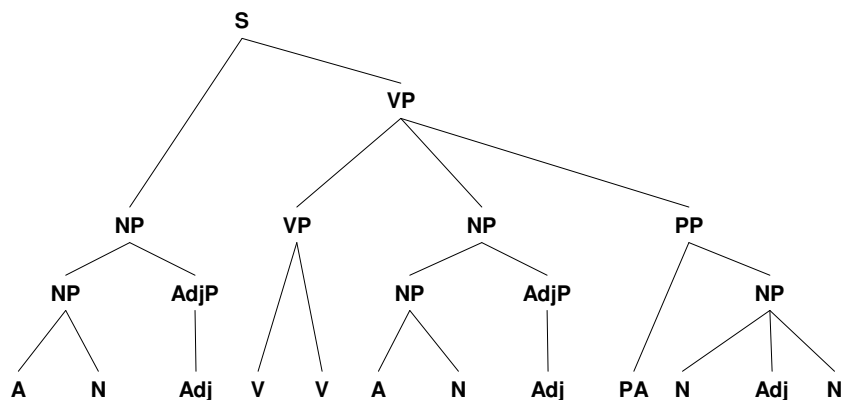


Chunking: esempio di stratificazione di un albero



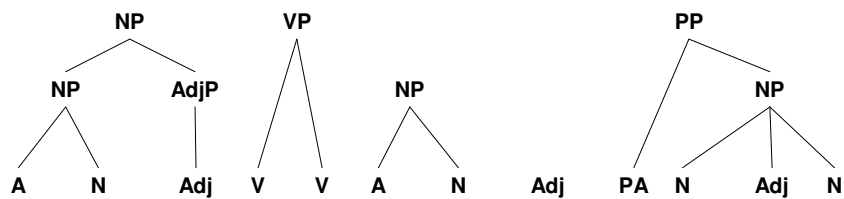
L' industria giapponese ha sfornato una versione perfezionata del robot umanoide "Db".

Chunking: esempio di stratificazione di un albero



L' industria giapponese ha sfornato una versione perfezionata del robot umanoide "Db".

Chunking: esempio di stratificazione di un albero



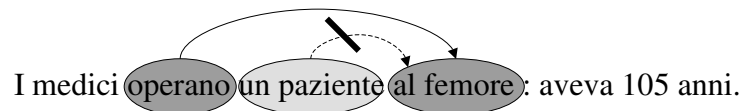
L' industria giapponese ha sfornato una versione perfezionata del robot umanoide "Db".

Decomposizione del processo

- **Chunk:**
 - livello intermedio di rappresentazione
 - giustificato psico-linguisticamente (Abney, 1991)
- **Definizione (intuitiva) di chunk:**
 - Sequenza di parole
 - fortemente connessa
 - con un unico portatore di significato
 - costate alle differenti interpretazioni

Decomposizione del processo

- LESSICALIZZAZIONE: Controllo dell'ambiguità
 - verbi controllano semantica delle proposizioni
 - quindi controllano le relazioni sintattiche



Decomposizione del processo

- Controllo dell'ambiguità
 - verbi controllano semantica delle proposizioni
 - quindi controllano le relazioni sintattiche

SUBJ **operare** *OBJ PP(a)*

| I **medici** | **operano** | un **paziente** | **al** **femore** : | **aveva** | 105 **anni** |

Definizione di chunk

- **Bottom-up:**
Una sequenza di parole che rappresenta il nucleo non ricorsivo di sintagmi nominali, preposizionali, verbali ed aggettivali
- **Top-down:**
Una sequenza di parole le cui relazioni non sono influenzate dal comportamento dei verbi

Chunk: osservazioni

Chunk

nuclei “non ricorsivi”* di sintagmi particolari



Chunk

sono riconoscibili con automi a stati finiti

* “non ricorsivi” = ricorsivi destri che non rimandano a sintagmi superiori

Chunking: prototipi

Prototipo:

- regola per catturare chunk
- esprimibile utilizzando informazione di POS tags tramite
 - *espressioni regolari/trasduttori* (Fastus, Alembic, Chanod&Ait)
 - *marker iniziale e finale* (ACL, 2001)

Chunking: prototipi

Esempi di *espressioni regolari*:

NPK:

Art **N** | Art A **N**

VPK:

V | V **V**

PPK:

P Art **N** | P Art A **N**

Chunking: prototipi

Esempi di *marker iniziale e finale*:

NPK:

MI: *Art* MF: *N*

VPK:

MI: *V* MF: *V*

PPK:

MI: *P* MF: *N*

Chunking: considerazioni

- Identificazione e classificazione:
 - possibile nel livello sintattico
 - risolvibile con macchinari semplici (i.e. FSA)
- La grammatica (ovvero i *prototipi*)
 - indipendente dal dominio di applicazione
- Domanda: *Qualora fosse la sola informazione estratta, sarebbe utile per una qualche applicazione?*

Prerequisiti

- Chunking
 - Part-of-speech tagging
- Riconoscimento dei legami verbali:
 - Individuazione dei limiti delle proposizioni (clause boundary recognition)

Architettura Generica di un RobParser

Part-of-speech tagging

- Definizione del problema

$$w_1 \dots w_n \rightarrow t_1 \dots t_n$$

NNS TO VB IN NNS PRP MD VB

Strategies to use with questions you cannot answer

Part-of-speech tagging

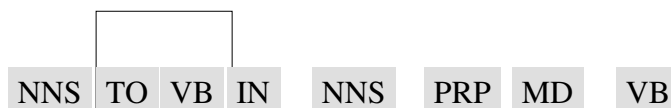
- Origini (1989)
sotto la spinta dell'Information Extraction
alla Message Understanding Conference
- Approcci
 - approcci simbolici (regole trasformazionali,
Brill 94)
 - approcci statistici (a seguire)

POS Tagging basato sulle trasformazioni (Brill, 94)

Dato un primo tagging (dizionario con tag più frequenti),
applicare regole di trasformazione fino a
che l'errore non diminuisca sotto una soglia

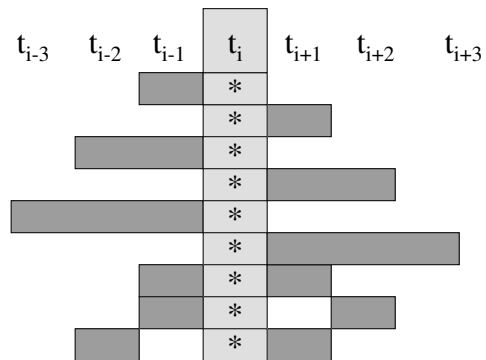
Trasformazioni

- Regole di riscrittura
t1 à t2 se <condizione nello spazio circostante
(triggering environment)>
- Esempio
NN à VB se il tag precedente è TO



Strategies to use with questions you cannot answer

Trasformazioni: Schemi dei triggering environments



Trasformazioni: algoritmo di apprendimento

- Quali trasformazioni?
- Quale ordine di applicazione?

Trasformazioni: algoritmo di apprendimento

```
C0:= Corpus con tag più frequenti
for k:=0 step 1 do
  v:= trasformazione n che minimizza E(n(Ck))
  if (E(Ck) - E(n(Ck))) < ε then break
  Ck+1 :=v(Ck)
  tk+1:= v
end
OUTPUT: sequenza t1, ..., tk
```

POS Tagging basato sulle trasformazioni

- Tagging delle parole sconosciute basato sulla morfologia
 - Tutte le parole sconosciute vengono taggate con NN
 - Il tag viene cambiato seguendo alcune regole trasformazionali **morfologiche**
 - Es: NNà NNS la parola termina con -s

POS Tagging basato sulle trasformazioni

Qualità dell'attività di POS Tagging dipende:

- dall'insieme dei tag obbiettivo
- dalla possibilità di recuperare informazione disambiguante nei contesti di attivazione
Es.: *che* in italiano (pronomi/congiunzione)
- dal materiale di apprendimento

Clause boundary recognition

- Definizione del problema

(L'industria giapponese dei robot, una delle poche (a non **risentire** della grave crisi economica), ha **sforato** una versione perfezionata del robot umanoide "Db", (fatto **nascere** due anni fa in un laboratorio vicino a Kyoto).)

Clause boundary recognition

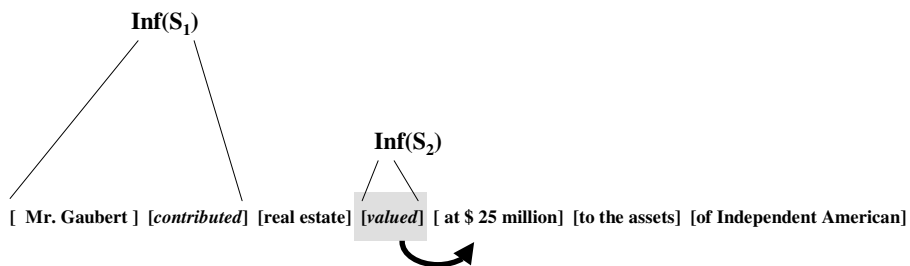
- Proposizioni sono utili per:
 - Conversione Text-to-speech
 - Allineamento di testi
 - Traduzione automatica
- Particolarità
 - Ricorsività non presente nei chunks

Clause boundary recognition

- Tecniche di apprendimento automatico (Shared Task in CoNLL, 2001)
- Tecniche in cooperazione con la ricerca dei legami verbali (Basili, Pazienza, Zanzotto, 1998)

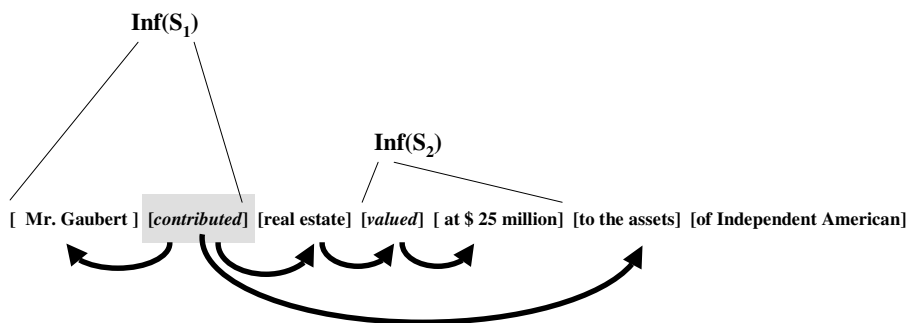
Clause boundary recognition

contribute-NP-PP (to)
value-NP-PP (at)



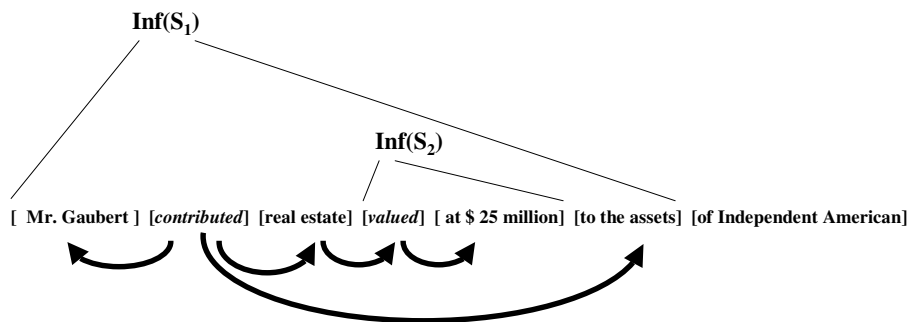
Clause boundary recognition

contribute-NP-PP (to)
value-NP-PP (at)



Clause boundary recognition

contribute-NP-PP (to)
value-NP-PP (at)



Clause boundary recognition

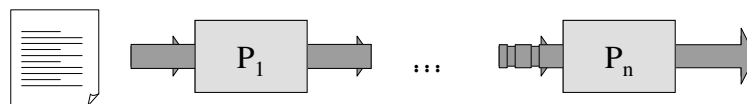
- Algoritmo:
 - Ipotesi iniziale di
 - minima estensione delle proposizioni
 - gerarchia derivata
 - Finché ci sono verbi da analizzare (da destra verso sinistra):
 - Riconoscere il legami verbali
 - Espandere l'estensione minima della proposizione

Controllo del processo

- Passi analisi:
 - POS Tagging
 - Chunking
 - Clause Boundary Recognition
 - Verb Argument Detection

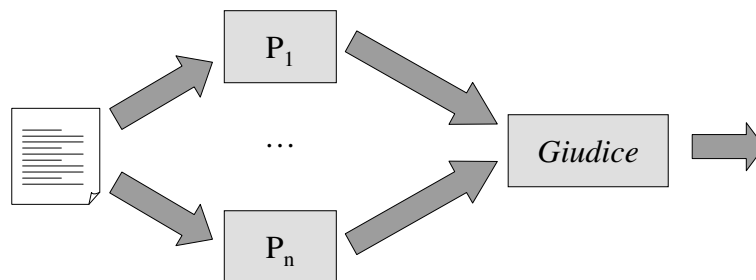
Controllo del processo

- Situazione problematica:
necessità di definire i tipi di dati trattati



Controllo del processo

- Situazione problematica:
necessità di definire i tipi di dati trattati



Formalismo di rappresentazione

Requisiti:

- Rappresentazione di analisi parziali
- Rappresentazione di legami distanti
- *Information hiding*
 - rendere disponibile la *sola* informazione necessaria ...
 - ma capace di esprimere *tutti* i vincoli correnti

Formalismo di rappresentazione

- Rappresentazione a costituenti
 - Context-free Grammar (Tree)
 - Well Formed Substring Table (WFST): *chart*
 - Tree-Adjoint Grammar (TAG)
- Rappresentazione a dipendenze
 - Link Grammar
- Rappresentazione miste
 - *Extended Dependency Graph* (XDG)

Decomposizione del processo

Principi:

- Scegliere i fenomeni trattati in ogni livello
- Scegliere l'algoritmo migliore per ogni task
- Scegliere un opportuno formalismo di rappresentazione

Discussione su Ridondanza (Verbmobil)

- Statistical NLP

RP: Modelli quantitativi

RP: Modelli a dipendenze

- (Collins, 96/7)

Parser *Robusti*

- Apple Pie (Satoshi Sekine, 92,95)
- Obbiettivi Caratteristici
 - Input/Output
- Metodologie
 - e.g. chunking, clause rec
- Prestazioni con citazione

Parser *Robusti*

- Link grammars (Lafferty et al, 94)
- Obbiettivi Caratteristici
 - Input/Output
- Metodologie
 - e.g. chunking, clause rec
- Prestazioni con citazione

Parser *Robusti*

- Chunking (Abney, 91)
- Obbiettivi Caratteristici
 - Input/Output
- Metodologie
 - e.g. chunking, clause rec
- Prestazioni con citazione

Parser *Robusti*

- Chunking (Montemagni, Pirrelli, 97)
- Obbiettivi Caratteristici
 - Input/Output
- Metodologie
 - e.g. chunking, clause rec
- Prestazioni con citazione

Parser *Robusti*

- Fastus (Hobbs, 96/7)
- Obbiettivi Caratteristici
 - Input/Output
- Metodologie
 - e.g. chunking, clause rec
- Prestazioni con citazione

Parser *Robusti*

- Verbmobil (???, 98)
- Obbiettivi Caratteristici
 - Input/Output
- Metodologie
 - e.g. chunking, clause rec
- Prestazioni con citazione
- DA METTERE SOPRA

Parser *Robusti*

- CHAOS / KARMA (Basili, Zanzotto, 98)
- Obbiettivi Caratteristici
 - Input/Output
- Metodologie
 - e.g. chunking, clause rec
- Prestazioni con citazione

RP: Acquisizione Lessicale

- PP-disambiguation
 - Estrazione di *Selectional Pattern* (Grishman, 1994/98)
 - *Transformation-based learning* (Resnik, Brill, 1996)
 - *Semantic preferences* (Basili et al., 1992, 93)
- GALOIS x KARMA, 1997 (Basili et al, 1997)

RP: Acquisizione Lessicale (2)

- IE rules
 - Booting an IE system (Basili, ..., Wilks, 1998)
 - Verb semantic pattern (Basili et al, 2000)

Conclusioni

- Un area di ricerca
 - molto ampia (CL, NLP, Statistica, SWEng)
 - molto attiva
 - relativamente matura ma ...
 - scarsamente sistematizzata

Prospettive di ricerca

- Fondazionali
 - sistematizzazione dei modelli ibridi
 - apprendimento di modelli probabilistici più espressivi
- Rappresentazionali
 - induzione
 - accuratezza delle decisioni grammaticali

Prospettive di Ricerca (2)

- Applicative
 - Principi di implementazione e design
 - Integrazione in scenari applicativi
 - IR
 - Hyperlinking
 - IE
 - Automazione dello sviluppo di risorse