Trattamento del Linguaggio Naturale - L'Approccio Classico

Roberto Basili

Department of Computer Science, System and Production
University of Roma, *Tor Vergata*
Via Della Ricerca Scientifica s.n.c., 00133, Roma, ITALY

e-mail: basili@info.uniroma2.it

Contenuti:

- Introduzione
- Alcune riflessioni sull'approccio vero-condizionale alla semantica
- Il problema della ambiguita' nell'analisi del linguaggio naturale
- Il ruolo del lessico
- Approcci moderni alla semantica lessicale

Trattamento automatico del linguaggio naturale

- Si occupa della definizione, realizzazione e validazione di *modelli computazionali* dei processi linguistici.

- Essi sono caratterizzati da una esplicita descrizione della *conoscenza riguardo al linguaggio*

- Principali Obbiettivi:
  - Analisi degli elementi linguistici (fonetici e morfologici)
  - Analisi della struttura linguistica (sintassi)
  - Analisi dei significati delle espressioni linguistiche
  - Analisi delle forme di comunicazione linguistica dedicate a compiti specifici (applicazioni)

Trattamento automatico del linguaggio naturale (2)

- Il progetto di algoritmi e sistemi software per la elaborazione del linguaggio naturale si e' concentrata su due principali aspetti della conoscenza linguistica:

  - *Competenza*
    Cosa caratterizza la mente ai fini del riconoscimento e della sintesi di dati linguistici?

  - Prestazione (*performance*)
    Quali legami funzionali esistono tra le conoscenze del parlante/ascoltatore e tali processi? (uso del linguaggio)

L'elaborazione del linguaggio naturale (3)

- L'attenzione degli approcci computazionali si e' quindi incentrata su:
  - *Efficenza* dei modelli simulativi ...
  - la loro *adeguatezza empirica* (*accuracy*)
  - e *Plausibilita' psicologica*
  - *Espressivita'* dei formalismi (linguaggi di rappresentazione)
  - Relazione tra i *linguaggi formali* e *naturali*
  - *Architetture* alla base della attivita' complessiva dei processi linguistici

Prospettiva storica

- Anni '40-'56: Fase dei fondamenti
  Teoria degli Automi (Turing),
  Teoria dell'Informazione (Shannon)
  Linguaggi Formali (Chomsky)

- Anni '57-'70
  Approcci Simbolici (AI: Newell, Simon)
  Approcci Stocastici (Harris)

- Anni '70-'83: Fase paradigmatica
  Modelli Markov (Jelinek (IBM-TJW), Baker (CMU))
  Logic Programming (Colmerauer) - DCG (Pereira, Shieber)
  Natural Language Understanding (Allen, Winograd, Wilks, Schank)
  Discourse Modeling (Grosz, Hobbs)

Prospettiva storica (2)

- Anni '83-'96: Nuova Fase Empirica

  Modelli Probabilistici (Jelinek, Baum, Mercer)

  Modelli a Stati Finiti (Fonologia, Morfologia e Sintassi)

  Corpora

  Valutazione delle Prestazioni

- Anni '90-oggi: Fase Ingegneristica

  Modelli Ibridi (Speech vs. Text Processing)

  IR, Information Extraction, Multimodalita'

  Knowledge Management, Ontologie, Semantic Web

Il programma della semantica vero-condizionale

- Stabilire delle linee guida per la nozione di riferimento degli elementi lessicali
  (a partire da una *teoria causale del riferimento*)

- Definire i meccanismi *composizionali* della sintassi
  (semantica di elementi funzionali del linguaggio, es *coordinazioni*)

- Verificare empiricamente il sistema ottenuto rispetto ai fenomeni interni (coreferenza, implicazione) ed esterni (discorso, atti linguistici)

Interpretazione degli elementi lessicali

Espressione	Riferimento	Senso
SN determinati	Individui	Concetti individuali
la stella del mattino	`Venere`	l'ultima stella che scompare al mattino
SN indeterminati	Insiemi	Concetti
le stelle	insiemi	corpi celesti
Gli ingegneri	di individui	professionisti
SV	Insiemi	Concetti
amare_Y	`X`	innamorati di Y
Espressioni	V/F	Pensiero

Benefici

- Enorme stimolo alla riflessione sul significato degli elementi lessicali
- Lo sforzo verso una rappresentazione sistematica (in maggiore o minore analogia con quella mentale)
- La definizione di un sistema linguistico a livelli
  - morfologico
  - sintattico
  - semantico
  - pragmatico

Osservazioni

- L'utilizzo di strumenti logici per lo studio di fenomeni linguistici non corrisponde ad una adesione al programma vero-condizionale

- Processi di analisi possono sfuggire alle regole del sistema vero-condizionale
  (ad es. grammatiche stocastiche)

- Sistemi computazionali enfatizzano il problema della rappresentazione senza per questo determinarne una semantica universale

- Le rappresentazioni intermedie acquisiscono significato all'interno dei processi *consumatori*

Ambiguita' (1)

L'ambiguita' affligge tutti i livelli della informazione linguistica:

- Il ruolo del contesto e' determinante:
  - contesto frasale
  - contesto comunicativo
  - contesto culturale: ontologico e storico

- La *contestualizzazione* coinvolge
  - le parole (cioe' il loro senso)
  - le strategie di interpretazione (e.g. preferenze grammaticali)
  - il dominio conoscitivo

Ambiguita' Morfologica

- E' relativa alla molteplicita' delle forme canoniche da cui una parola in un testo libero e' derivabile
- Overloading (piu' ruoli per la stessa forma di trascrizione)

- Proprieta' morfologiche e ambiguita':
  - categoria sintattica: $imposta_{finestra}$ vs. $imposta_{imporre}$
  - persone verbali: sia, $imponga$
  - genere: $rosa$, boa, ...
  - numero: $radio$, $biro$, ...

Ambiguita' Sintattica

- E' relativa alla molteplicita' delle strutture sintagmatiche che e' possibile assegnare ad una frase
- No Overloading (sempre un ruolo per la stessa forma superficiale, sequenza)

- Proprieta' sintattiche e ambiguita':
  - categoria sintattica: *la vecchia porta la sbarra*
  - dipendenze/modificatori: *devo sempre andare a studiare*
  - ambiguita' sintagmi preposizionali:
    guardo la ragazza nel parco

  $\Rightarrow$ anafora: *tiro l'uovo contro il muro e questo non si rompe*

Ambiguita' Sintattica (2)

- Affligge l'efficenza computazionale del parsing
- La combinazione di forme diverse di ambiguita' produce una esplosione esponenziale di interpretazioni

- Non e' completamente risolubile su base morfologica o sintattica:
  - *la vecchia porta la sbarra*
  - *devo sempre correre per studiare dal giardino al parco*
  - *guardo la ragazza nel parco e nel giardino pero' mi diverto*

Ambiguita' Sintattica (3)

Alcuni approcci:

- Disambiguazione rimandata e rappresentazione separata di interpretazioni (*foreste di alberi*)
- Disambiguazione rimandata e rappresentazione compatta delle diverse interpretazioni (*grafi*)
- Integrazione di piu' livelli: *parsing* lessicalizato o semantico
- Disambiguazione preventiva su base euristica:
- Processamento *parallelo* o *distribuito*

Ambiguita' Semantica

- E' relativa alla molteplicita' dei significati che e' possibile assegnare ad una frase o ad una sua parte
- No Overloading (sempre un ruolo per la stessa forma superficiale, sequenza)
- Proprieta' semantiche e ambiguita':
  - senso: *Roberto e' rosso*
    l'astronomo adora la stella
  - relazioni semantiche: *ho comprato il libro di Rossi*
  - molteplicita' di modelli:
    Ogni uomo ama una donna
    Ogni uomo adora Julia Roberts

Il livello semantico: applicazioni

- Per *comprendere* un testo (o una sua parte)
- Per *tradurre* un testo
- Per *decidere* sulla base dei contenuti di un testo:
  - *ricercare* informazioni (in DB bibliografici, File systems, Web)
  - *classificare* testi (su Web o in un DBMS dedicato)
  - *estrarre* dati o notizie (Monitoring/Controllo di flussi informativi testuali)
  - *apprendere* da collezioni di testi

Ho bisogno di una rappresentazione formale del *significato*.

Ambiguita' e lessico

- *Mario beve la birra in lattina*
  - `bere(m,b)` $\wedge$ `sorgente(b,l)`

- *Mario beve la birra in cucina*
  - `bere(m,b)` $\wedge$ `location(m,l)`

Posso vincolare ulteriormente il lessico per scegliere l'interpretazione giusta?

Ambiguita' e lessico

- *Mario beve la birra in lattina*

  $\rightarrow$ `bere(m,b)` $\wedge$ `sorgente(b,l)`

  $*$ `location(prep(in),m,l)` $\wedge$ `source(prep(in),b,l)`
  (*Un costituente non puo' attivare due interpretazioni*)

  $*$ `bere(m,b)` $\wedge$ `location(m,l)`
  $(location(X, lattina) \Rightarrow \forall Y \quad bere(X, Y) \models \bar{F})$

Ambiguita' e lessico - modelli quantitativi

- *Mario beve la birra in lattina*

  $\rightarrow$ `bere(m,b)` $\wedge$ `source(b,l)`

  * `bere(m,b)` $\wedge$ `location(m,l)` $\wedge$ `source(b,l)`
    $prob(location(m,l) \wedge source(b,l)) = 0$

  * `bere(m,b)` $\wedge$ `location(m,l)`
    $prob(location(m,lattina)) \ll prob(source(b,l))$

R: $\forall X \quad prob(bere(X,b)|location(X,l)) = 0$

Ambiguita' del Senso e lessico

- *L'astronomo sposa la stella*
  $\forall X, Y \quad sposare(X, Y) \Rightarrow umano(Y)$

? *L'astronomo adora la stella*

? *L'auto beve benzina*
  (*) $\forall X, Y \quad bere(X, Y) \Rightarrow potabile(Y)$
  (!) $\forall X, Y \quad bere(X, Y) \wedge combustibile(Y) \Rightarrow consumare(X, Y)$

  `IF` $bere(X, Y) \wedge combustibile(Y)$
  `THEN change` $bere(X, Y)$ `in` $consumare(X, Y)$

Ambiguita' e lessico

- Il lessico e' la sorgente di informazioni determinanti per:
  - la corretta interpretazione semantica
    (e.g. la numerosita' degli argomenti verbali)
  - la disambiguazione delle strutture sintattiche ambigue
    (e.g. regole per il riferimento dei sintagmi preposizionali)
  - la disambiguazione del senso
- la sua codifica nei sistemi computazionali pone il problema di una rappresentazione psicologicamente plausibile (*senso cognitivo* vs. *senso cognitivo*)

Ambiguita' e lessico

La codifica estensionale delle regole lessicali apre numerosi problemi:

- correttezza logica
  (e.g. proprieta' lessicali vs. inferenze contestuali/pragmatiche)

- completezza, (problematicita' di una semantica di mondo chiuso)

- complessita' del design (rappresentazione e modelli computazionali)

- costi di sviluppo

- portabilita' tra applicazioni e domini linguistici

- complessita' dei processi (accesso ed attivazione delle regole)

Approcci correnti al lessico nei sistemi NLP

Modelli linguistic avanzati

- Formalismi algebrici di rappresentazione (FS)
- Ereditarieta' delle parole e delle proprieta' (e.g. classi/categorie lessicali)
- Modelli psicolinguistici
- Modelli generativi

Approcci empirici allo sviluppo del lessico

- Modelli quantitativi delle proprieta' lessicali
- Algoritmi induttivi da campioni (e.g. *corpora*)
- Dizionari elettronici

Semantica delle parole

- Connessioni semantiche
  (*madre*, *genitore*, *femmina*)
  $madre(X) \Rightarrow genitore(X) \wedge femmina(X)$

- Relazioni tra morfologia e semantica
  (*ri-vincere*, *in-solito*)

- Proprieta' di tipo logico
  (*manoscritto giallo* vs. *grande manoscritto* vs. *ipotetico manoscritto*)

Semantica delle parole: Ruoli tematici

I verbi esprimono classicamente le seguenti categorie:

- Stati (*essere ubriachi*, *sostare*)
- Azioni (*correre*, *comprare*, *salire*)
- Eventi telici (*raggiungere*, *scalare*, *tagliare*)

Semantica delle parole: Ruoli tematici (2)

La descrizione dei modi con cui gli individui/entita' partecipano agli stati, azioni ed eventi e' spesso denominata *struttura argomentale* o *struttura tematica* (*Ruoli tematici*)

- Relazioni tra le *posizioni* nei predicati verbali ed i *ruoli* corrispondenti nell'evento descritto
  (*Agente*, *Paziente*, *Tema*)
  (Case relations, Fillmore, 1968)

- Proprieta' caratterizzanti l'evento in termini di primitive *spazio-temporali* (*Relazioni tematiche*, Gruber, Jackendoff)

Semantica delle parole: Ruoli tematici (3)

Due punti di vista sui *Ruoli tematici*:

- Teoria degli argomenti ordinati

- Teoria dei ruoli tematici basati su eventi

Teoria degli argomenti ordinati

- *Vieri spaventa Luisa*
- *Vieri colpisce Luisa*
- *Vieri ama Luisa*

- `spaventa(v, l) - spaventa:` v_S, l_O
- `colpisce(v, l) - colpisce:` v_S, l_O
- `ama(v, l) - ama:` v_S, l_O

Teoria dei ruoli tematici basata sugli eventi

- Introduce un calcolo logico per la descrizione dei predicati (verbali)

- Il calcolo contiene variabili che variano su
  - *eventualita'*, e
  - insiemi di individui, $x, y, z, ...$
  - istanti di tempo e mondi possibili

Teoria dei ruoli tematici basata sugli eventi

- *" Totti colpisce il palo"*
  $\exists e \quad [colpire(e) \wedge AGENT(e) = t \wedge THEME(e) = p]$

- *" Totti ama Maria"*
  $\exists e \quad [ama(e) \wedge ESPERIENTE(e) = t \wedge THEME(e) = m]$

Sommario (1)

- Sono stati introdotti i principali apsetti della ambiguita' linguistica
- E' stato evidenziato il ruolo del lessico nella risoluzione delle principali ambiguita'
- Sono state sintetizzate le tecniche principali alla rappresentazione (computazionale) dell'informazione lessicale
- E' stata introdotta la nozione di *struttura argomentale* e di *ruolo tematico* nella rapresentazione dei verbi di una lingua

Bibliografia

- *Intelligenza Artificiale*, S. J. Russel, P. Norvig, Prentice Hall Int., Chapter 22.3-22.8, 23, 1998.

- *NLP In Prolog*, G. Gazdar, C. Mellish, Chapter 7, 8, 1998.

- *An Introduction to Unification-based Approaches to Grammar*, S. Shieber, Chapter 1, 2, 7, 8, CSLI Lecture Notes, n. 4, 1986.

I sistemi di NLP

Le fasi principali dei processi di interpretazione linguistica in un sistema NLP sono:

- Pre-elaborazione del testo
- Analisi morfologica
- Analisi sintattica
- Analisi semantica
- Fase pragmatica (*task-oriented*)

Pre-elaborazione testuale
Obbiettivi

- Normalizzazione dei testi in ingresso rispetto a formati (e.g. HTML, PS, PDF)

- Selezione dell'informazione testuale (*zoning*)

- Tokenizzazione

- Trattamento di fenomeni speciali:
  - Punteggiatura
  - Forme di descrizione speciali (e.g. numeri)
  - Fenomeni di dominio (e.g. URLs, Strutture dei domenti - commi, ...)

Risultato: Segmenti di testo tokenizzati

Analisi Morfologica
Obbiettivi

- Individuazione di *Nomi Propri* (e.g. A.S. Roma)
- Dictionary Look-up per le *classi aperte* di tokens (e.g. nomi, verbi, aggettivi, avverbi)
- Analisi delle forme terminologiche (o *entita' nominate*, e.g. *Dipartimento di Scienze della Formazione*)

Risultato:

- Sequenze di tokens analizzati
- Relazione uno a molti tra tokens e forme lessicali
- Trattamento dei tokens sconosciuti

Analisi Morfologica
Tecniche

- Ricerca in database morfologici

- Trattamento sintattico della morfologia (FSA)
  `<NPlur> -> <NRoot>`+*i*
  `<NSing> -> <NRoot>`+*o*

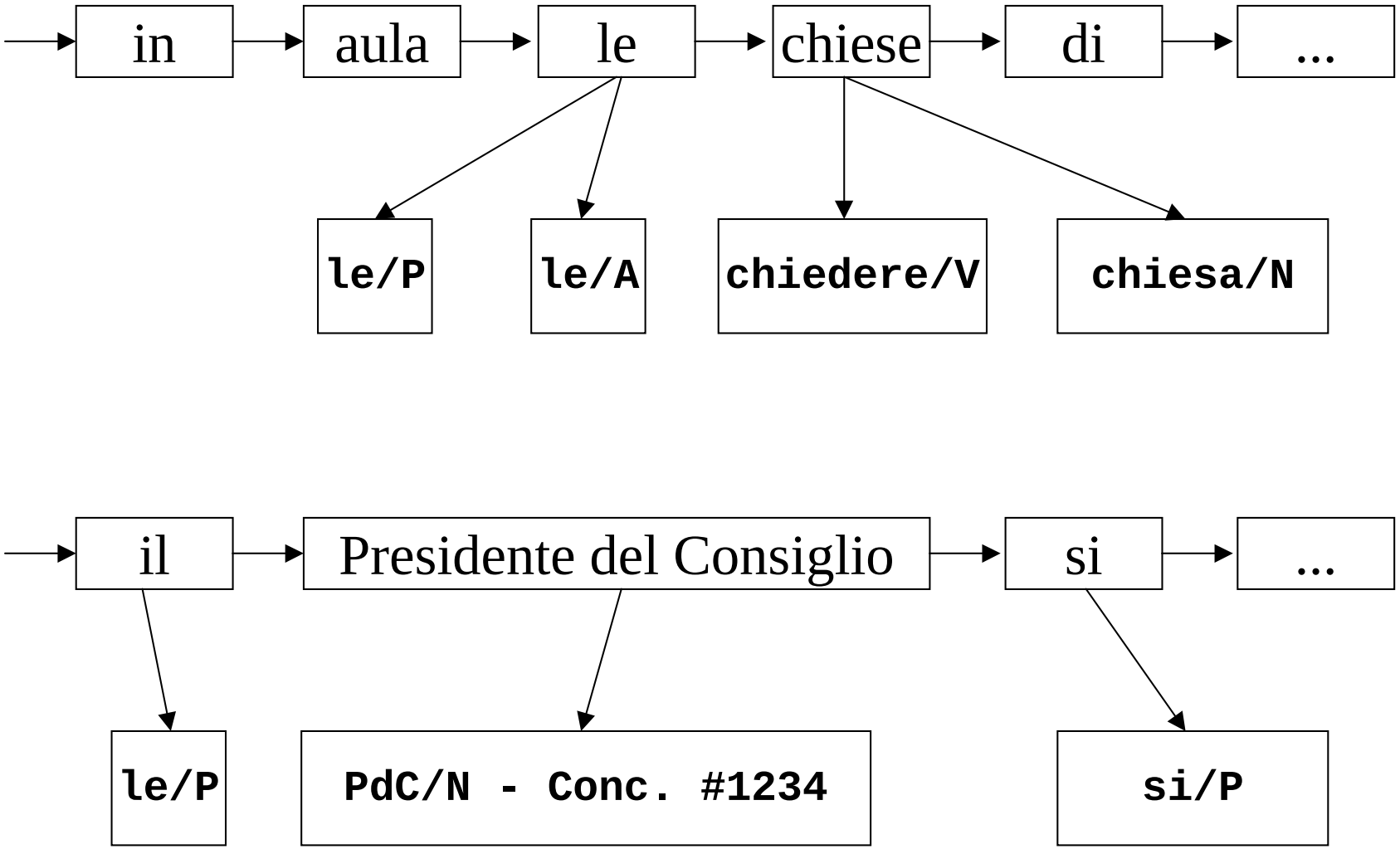
- Adozione dei principi della morfologia generativa (Ereditarieta')
  `<VInfinito> -> <VRoot1Con>`+*are* (e.g. *guard-are*, *am-are*))

Problemi:

- La relazione tokens e interpretazioni e' molti-a-molti
  `X, <le>, Y` $\rightarrow X$, $\{$*articolo, pronome*$\}$, Y
  `X, il, Presidente, del, Consiglio, Y` $\rightarrow X$, $\{$*ncom.sing*$\}$, Y

- Trattamento dei tokens sconosciuti

Analisi Morfologica



Analisi Sintattica
Obbiettivi

- Riconoscimento/Accettazione/Riscrittura delle proprieta' strutturali delle frasi

- Classificazione linguistica dei legami tra costituenti
  `<VPTrans> -> <V> <NP>`

- Determinazione delle unita' semantiche riflesse dalla sintassi
  `<NP> -> <Art> [<Agg>*] <N>`

Risultati:

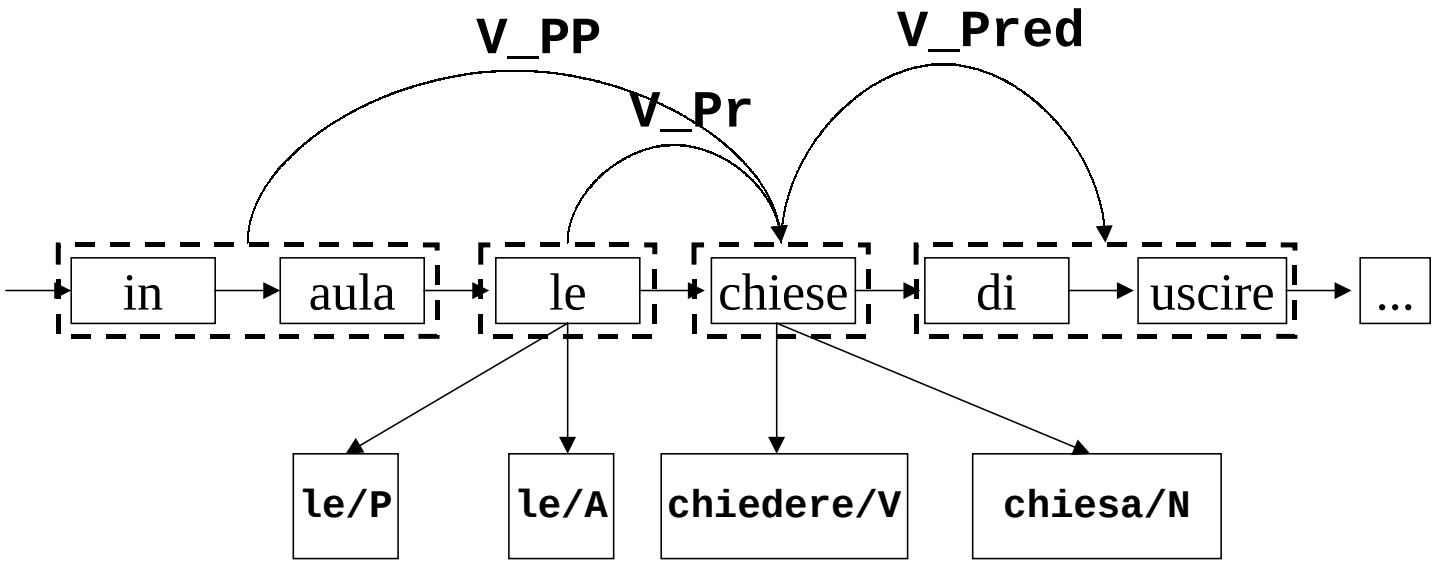
- Relazione tra entita' morfologiche (*grafi*)

- Sottosequenze (e.g. sequenze di *chunks*)

- Disambiguazione morfosintattica

- Individuazione della struttura frasale (i.e. proposizioni)

Analisi Sintattica - Rappresentazione

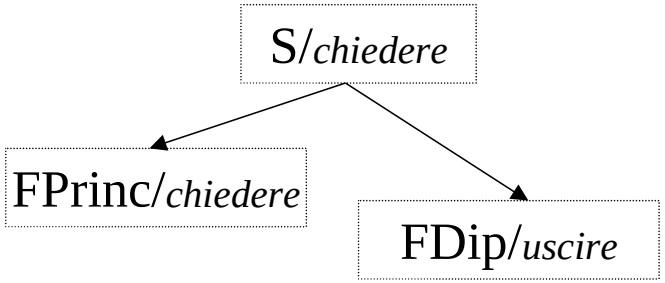


chunks

dipendenze

S/*v*

proposizioni



Analisi Sintattica
Tecniche

- Disambiguazione del Part-Of-Speech tag (*POS tagging*)
  (*SELECT 1 tag per token*)

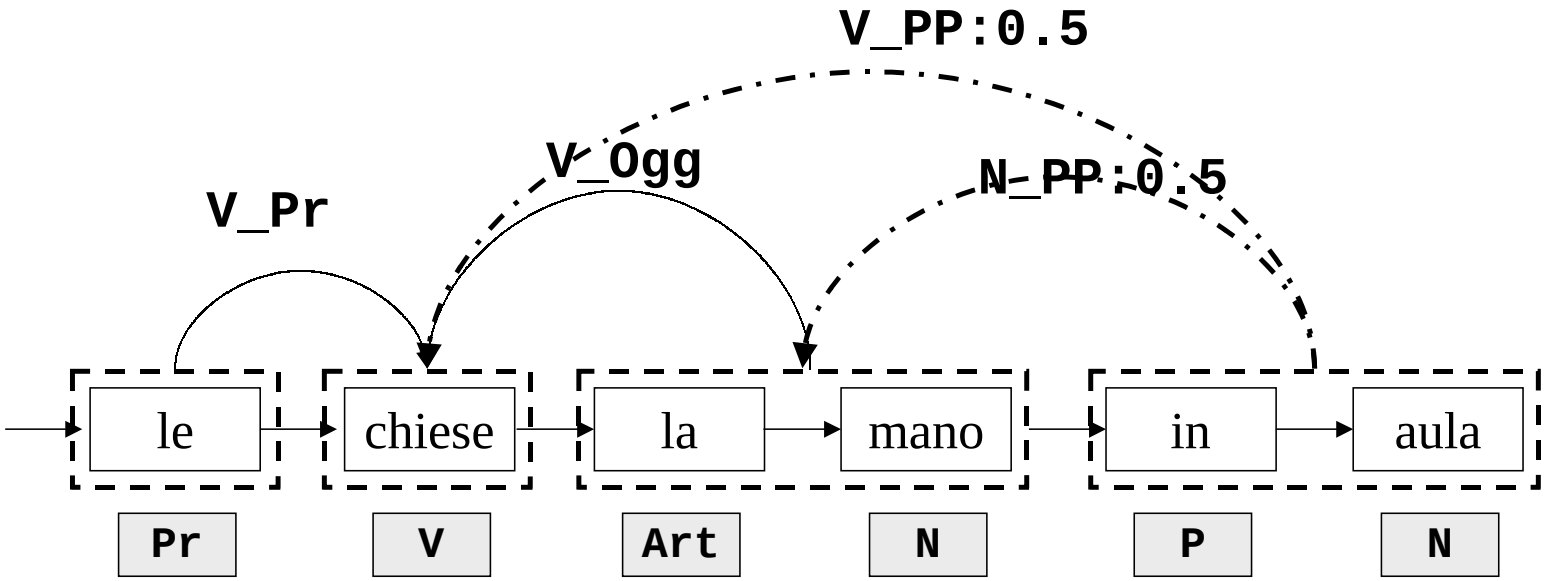
- Decomposizione del processo di analisi grammaticale
  (*Stratificazione della grammatica*)
  (e.g. *NP kernels first*)

- Analisi lessicalizzata
  `IF(` $VP(chiedere) \in S$ `) THEN Look_FOR(` NP_dativo `)`

- Strategia: *Disambigua il piu' tardi possibile*

- Modelli quantitativi della ambiguita' grammaticale

Analisi Sintattica - Rappresentazione



POS **POS tags**

chunks

dipendenze non ambigue

dipendenze ambigue

Analisi Semantica
Obbiettivi

- Determinare la trascrizione (sintattica) del significato delle frasi in ingresso

- Disambiguazione delle unita' semantiche riflesse dalla sintassi (WSD)

```
np(Sem) --> n(Sem)
n(human(X) ∧ female(X)) --> [stella]
n(star(X)) --> [stella])
vp(Espr) -->  vp(X^Espr),  n(X), {sem_check(Espr,X)}.
```

- Interpretazione di tutti i legami sintattici

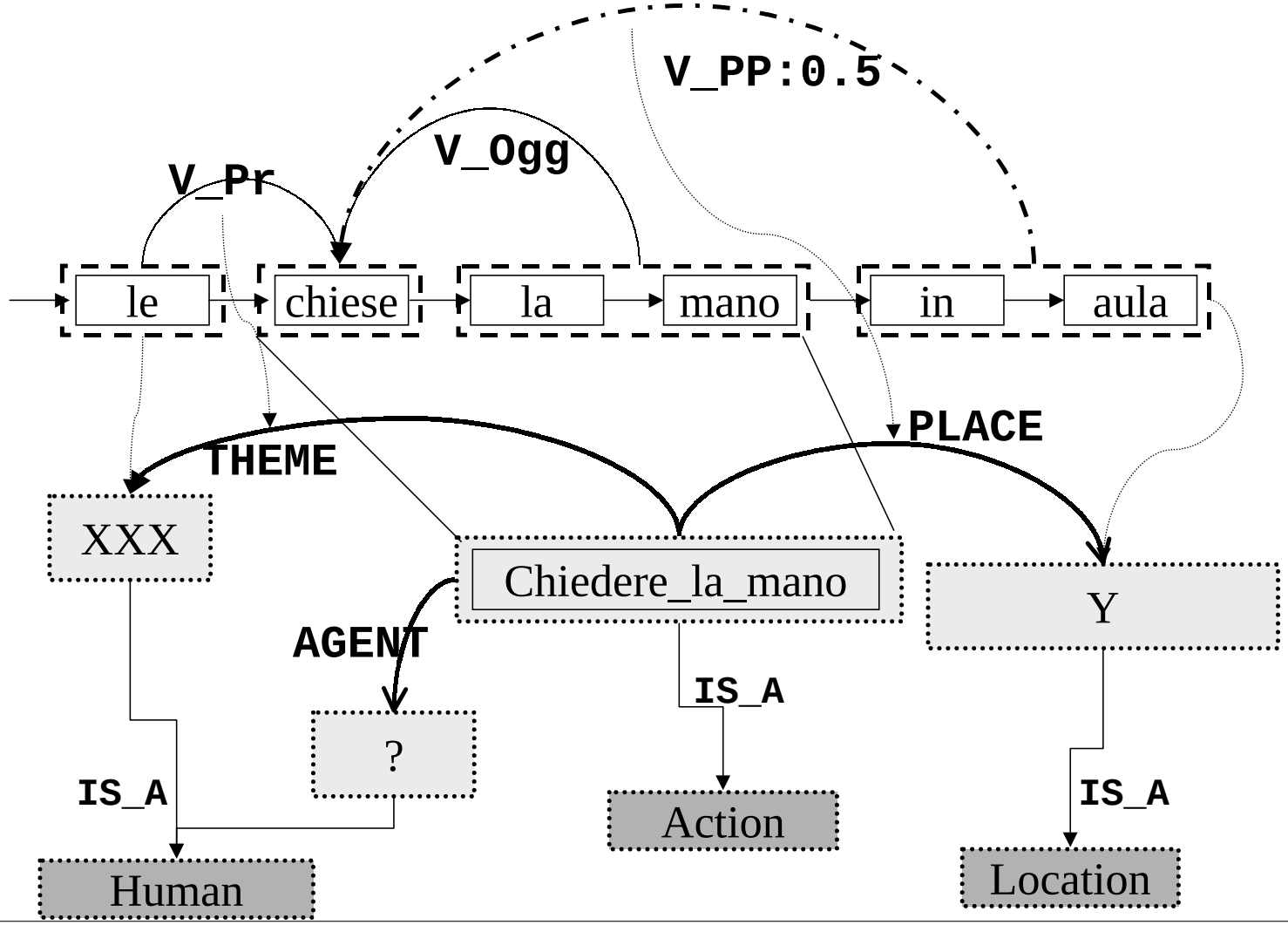
```
vp(Espr) --> v(X^Sem), np(X)
```

Analisi Semantica

Risultati:

- Una rappresentazione formale del significato (grafi, FL)
- Relazione (univoca) tra tokens e referenti
- Relazioni (di dominio) tra referenti
- Disambiguazione della struttura frasale

Analisi Semantica - Rappresentazione



Analisi Semantica
Osservazioni

- Nel processo di interpretazione i costituenti fanno riferimento a concetti del contesto

- I concetti del contesto possono rappresentare intere strutture (es. frasi transitive)

- I concetti non sono isolati ma sono organizzati in classi (e.g. *location*, *human*)

- La organizzazione dei concetti e' tutto cio' che il sistema conosce
  cioe' il suo mondo senza ulteriori livelli interpretativi

- Cio' che tale mondo *e'* (al di la' di un processo di interpretazione) e' detto *ONTOLOGIA*

Analisi Semantica
Osservazioni (2)

- Durante il processo di interpretazione un modello ontologico e' utilizzato per
  - Interpretare i concetti (invidualmente)
  - Disambiguare il senso degli elementi lessicali
  - Disambiguare le strutture grammaticali ambigue (es. PP)
  - Interpretare le relazioni grammaticali (cioe' concettualizzarle)

- Le proprieta' dei concetti nell'ontologia sostengono le inferenze necessarie alla disambiguazione semantica

Sommario (2)

- Sono stati discussi i principali aspetti della ambiguita' linguistica
- Sono stati analizzate le fasi principali dell'analisi di un sistema NLP
- Sono state suggerite le tecniche algoritmiche principali utilizzate in tali fasi
- E' stato evidenziato il ruolo del lessico nelle diverse fasi
- Sono state descritte le tecniche principali alla base della prassi corrente nella rappresentazione (computazionale) dell'informazione lessicale
- E' stato introdotta la nozione di ontologia nella prospettiva di un sistema di NLP

Bibliografia

- *Intelligenza Artificiale*, S. J. Russel, P. Norvig, Prentice Hall Int., Chapter 22.3-22.8, 23, 1998.

- *NLP In Prolog*, G. Gazdar, C. Mellish, Chapter 7, 8, 1998.

- *The Generative Lexicon*, J. Pustejovsky, Chapter 1, 2, 3, MIT Press, 1995.

- *An Introduction to lexical semantics*, P. Saint-Dizier, E. Viegas, in *Computational Lexical Semantics*, Cambridge University Press, 1995.