

# Indicizzazione Concettuale: Un Modo Migliore Per Organizzare La Conoscenza

**Corso di Intelligenza Artificiale:  
Trattamento Automatico di Linguaggio  
Naturale**

Barbara Tozzi  
Pisa 18 Giugno 2002  
Dipartimento di Informatica - Università di Pisa

Friday, February 16, 2007

1

## CONTENTS:

- Indicizzazione classica
- Indicizzazione concettuale
- Algoritmo di classificazione
- Tassonomia concettuale
- Sistema di indicizzazione concettuale e di recupero di brani
- Infrastruttura del sistema di recupero
- Struttura concettuale di una locuzione
- Sussunzioni concettuali
- WordNet e la sua Organizzazione
- Conclusioni
- Bibliografia

Friday, February 16, 2007

2

## INDICIZZAZIONE CLASSICA

### Indicizzazione Classica

- Elenco telefonico (in ordine alfabetico)
- Pagine gialle (elenco sulla base di categorie)
- Indice di un libro (ordine gerarchico: per capitoli\argomenti)

Friday, February 16, 2007

3

## PROBLEMI I

### Problemi Ordine Alfabetico:

- Giusta sequenza di parole (dipendente dalle tecniche usate dal catalogatore)
- Le parole cercate sono diverse da quelle di cui abbiamo bisogno (es.: "automobile cleaning" – "car washing")
- La parola cercata può essere catalogata con maggiori informazioni sotto un altro indice.

Friday, February 16, 2007

4

## PROBLEMI II

### Problemi Ordine Gerarchico

- Tipo di classificazione statica il cui aggiornamento è lento
- "n" scelte sull'albero gerarchico che nel peggiore dei casi devono essere controllate tutte separatamente.
- Parole che rimandano ad altre parole

Friday, February 16, 2007 5

## SINONIMI

Possibile soluzione :  
uso di un dizionario dei sinonimi

### Problemi:

1. In uno stesso gruppo di sinonimi alcuni hanno spesso significati diversi
2. Può portare a termini con significati troppo generici

Friday, February 16, 2007 6

## GENERICITA'

### POSSIBILE SOLUZIONE:

- La situazione migliora usando una relazione di "genericità"
- **Es.:** se cerco "veicolo a motore" posso trovare tutti i tipi di veicoli a motore, mentre se cerco "automobile" posso trovare auto e taxi ma non autobus o camion che invece rientrano nella categoria dei veicoli a motore.

Friday, February 16, 2007 7

## ONTOLOGIA

- Ontologia: in generale descrive i concetti di un dato dominio applicativo
- **Es.:** WordNet tenta di descrivere i concetti (sinonimi, contrari, e relazioni concettuali) contenuti nella lingua inglese attraverso uno Schema Ontologico

"An ontology is a specification of a conceptualization" [Gruber '95]

Friday, February 16, 2007 8

## CONCETTUALIZZAZIONE

- La struttura formale della realtà come percepita e organizzata da un agente

**Indipendente da:**

- Vocabolario usato
- Specifica situazione

• **ES.:**



Friday, February 16, 2007 9

## DA ONTOLOGIA A SCHEMA CONCETTUALE

| Ontologia  | Schema Concettuale   |
|--|--|
| – Semantica formale  | – Non sempre ha un a semantica formale                                     |
| – Si specificano tutti i vincoli rilevanti per caratterizzare il significato | – Si specificano solo i vincoli rilevanti per l'aggiornamento del database |

Friday, February 16, 2007 10

## INDICIZZAZIONE CONCETTUALE

- **DEF:** metodologia che permette ai sistemi indicizzanti di superare la struttura delle locuzioni nel materiale indicizzato
- **ES.:** "PIOVONO CANI E GATTI" (=locuzione)



- suddivisione automatica in una o più strutture concettuali che rappresentano il modo in cui gli elementi della locuzione sono uniti per costruirne il significato.

Friday, February 16, 2007 11

## INDICIZZAZIONE CONCETTUALE I

- **Es.:** nella ricerca di "car washing" attraverso delle relazioni di genericità posso arrivare ad associare "car" ad un concetto più generico come "automobile" e "washing" a "cleaning", arrivando così ad ottenere una nuova locuzione:  
"car washing" → "automobile cleaning"
- **CAR = SUBSUMEE**  
( + specifico di: )
- **AUTOMOBILE = SUBSUMERS**

Friday, February 16, 2007 12

## INDICIZZAZIONE CONCETTUALE II

MEZZI UTILIZZATI DA QUESTO SISTEMA

- **AXIOMS** ( assiomi ) = nozioni di base → ATOMICHE. (ASSIOMI specifici appartengono a campi specializzati: informatico, scientifico, commerciale, etc.)
- **LEXICON** (o dizionario) = insieme di ASSIOMI inseriti da operatori umani
- **AXIOMS SUBSUMPTION** = relazioni di inclusione tra le nozioni base conosciute ( AXIOMS)
- **ES.:** Car è un tipo di automobile.  
Cleaning è un tipo di washing

Friday, February 16, 2007 13

## INDICIZZAZIONE CONCETTUALE III

Consente di

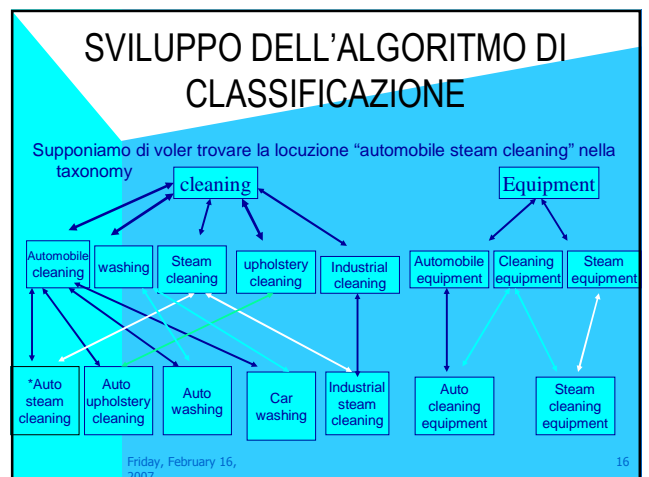
- Decidere formalmente quando una frase è più generale di un'altra
- Organizzare automaticamente la tassonomia concettuale
- Esplicito collegamento con i più specifici sottoconcetti
- Espansione della tassonomia

Friday, February 16, 2007 14

## ALGORITMO DI CLASSIFICAZIONE

- **MSS** = "most specific subsumers", algoritmo che serve a determinare assiomi che specificano (sotto-concetti o sottoclassi)
- **MSG** = "most specific subsumee", algoritmo che serve a determinare assiomi che generalizzano (super-classi)
- **CLASSIFICATION ALGORITHM = MSS + MSG :** pone una nuova parola o locuzione (cioè non esistente nel dizionario a sua disposizione) nella giusta posizione della TAXONOMY (=CLASSIFICAZIONE), pur non facendone fisicamente parte.

Friday, February 16, 2007 15



## RISULTATO DELLA RICERCA

- Supponiamo di applicare uno di questi algoritmi - MSS - ad un elenco indicizzato in ordine gerarchico come ad es. le pagine gialle darebbe questo risultato:
- **automobile cleaning**
  - automobile upholstery cleaning+
  - automobile washing-
  - car washing+
- **steam cleaning**
  - industrial steam cleaning+

Friday, February 16,  
2007

17

## TASSONOMIA CONCETTUALE

**CONCEPTUAL TAXONOMY**: organizzazione automatica di parole e/o locuzioni.

TASSONOMIA CONCETTUALE → MAPPA CONCETTUALE

- **SISTEMA DI RECUPERO**: estrae concetti dal testo e li organizza in una tassonomia concettuale
  - **vantaggio**: il tempo di ricerca nel materiale indicizzato è minimizzato

Friday, February 16,  
2007

18

## TECNOLOGIA DI INDICIZZAZIONE CONCETTUALE

Può essere pensata come combinazione di :

- **estrattore di concetto** (che identifica le parole e le locuzioni da indicizzare)
- un "**ASSIMILATORE DI CONCETTI**" che analizza la struttura e il significato di una locuzione-concetto per determinare dove posizionarla nella tassonomia
- Un **sistema concettuale di recupero** che usa la tassonomia concettuale per fare le connessioni tra le richieste e gli oggetti indicizzati dell'informazione
- Un **navigatore concettuale** che permette ad un utente di leggere la tassonomia concettuale e di muoversi avanti e indietro tra: concetti della tassonomia e le occorrenze di questi concetti nel materiale indicizzato

Friday, February 16,  
2007

19

## INFRASTRUTTURA DEL SISTEMA DI RECUPERO

- Manipolazione dei documenti
- analisi lessicale
- analisi della frase
- classificazione tassonomica
- lettura e recupero

Friday, February 16,  
2007

20

## MANIPOLAZIONE DEI DOCUMENTI

- **TOKENIZER** : segmenta una sequenza di caratteri in una sequenza di parole e simboli, segni di punteggiatura e indicatori di struttura
- **INTERPRETE DI MARKUP** : permette al sistema di trattare con sistemi di markup espliciti (HTML, SGML)
- **ANALIZZATORE DI MARKUP IMPLICITI** : tratta con la struttura implicita nel testo normale che non è stata segnata ma che è stata organizzata per essere letta da una persona

Friday, February 16,  
2007

21

## TOKENIZING

- **TOKENIZING**: metodologia di estrazione dei concetti da un testo
- **Per Estrarre i Concetti** : interpretare il testo come sequenza di parole (occasionalmente interrotte dalla punteggiatura)
- **TOKENIZER** :
  - segmenta una stringa di caratteri → una sequenza di token
  - problema : o'clock o'clock (tokenizer)  
o'clock o'clock (dizionario)
  - obiettivo : raggruppare i caratteri in unità (= WORD TOKEN) in modo da essere ritrovati nel dizionario.

Friday, February 16,  
2007

22

## TOKENIZING II

- **WORD TOKEN** : sequenza di caratteri e simboli speciali che saranno usati come chiave di ricerca nella ricerca lessicale
- **IMPORTANTE** : una volta trovati verranno aggiunti al suo lessico

Friday, February 16,  
2007

23

## ANALISI LESSICALE

Si può fare in due modi:

- **DIZIONARIO PRORIO E CONOSCENZE DI BASE DELLE INFORMAZIONI SEMANTICHE E SINTATTICHE DELLE PAROLE E LOCUZIONI**
- **INSIEME DI REGOLE MORFOLOGICHE** : utili per riconoscere
  - le regolari forme inflesse delle parole
  - parole derivate da altre parole a cui è stato aggiunto un suffisso o un prefisso
  - parole composte dall'unione di parole note
  - poter ricavare informazioni sintattiche e semantiche abbastanza attendibili sulle parole sconosciute

Friday, February 16,  
2007

24

## ANALISI LESSICALE II

### CONFRONTO CON IL DIZIONARIO

- Identificato un token in un testo → si cerca su di un dizionario per determinare qualcosa su di esso
- Dalla sua elaborazione → risposte codificate come coppie "caratteristica-valore" da inserire nel suo dizionario
- Mantenere un dizionario relativamente grande in cui aggiungiamo e aggiorniamo i tokens trovati con i loro significati+un dizionari specializzati di termini tecnici, scientifici, ecc.

Friday, February 16, 2007 25

## ANALISI LESSICALE III

### RISULTATI DEL CONFRONTO

Le voci che entrano a far parte del dizionario, contengono informazioni su:

- le possibili categorie sintattiche che una parola può avere
- le relazioni semantiche tra una parola e altri concetti
- le flessioni e le variazioni della pronuncia di una parola

Friday, February 16, 2007 26

## ANALISI LESSICALE IV

### L'AMBIGUITA' DEL LINGUAGGIO NATURALE

- Una parola può avere diversi significati
  - diverse categorie sintattiche ⇒
  - differenti comportamenti sintattici ⇒
  - differenti relazioni semantiche

MAY

- = MAGGIO ( mese )
- = May ( nome di persona )
- = potere ( verbo )

- Soluzione: nel dizionario vengono associati i significati delle parole alle caratteristiche sintattiche o semantiche

Friday, February 16, 2007 27

## ANALISI LESSICALE V

### ANALISI MORFOLOGICA

- Non tutte le parole possono appartenere ad un dizionario
- Quando la ricerca fallisce → ANALISI MORFOLOGICA

#### COMPITI DELL'ANALIZZATORE MORFOLOGICO

- determinare una categoria sintattica e le caratteristiche sintattiche che per mezzo di un programma (=PARSER) analizza la struttura concettuale delle locuzioni
- identificare la radice e qualche volta altre relazioni semantiche che determinano gli assiomi di sussunzione e altre relazioni da aggiungere alla tassonomia per supportare il recupero concettuale.

Friday, February 16, 2007 28

## ANALISI LESSICALE VI

### REGOLE DELL'ANALIZZATORE MORFOLOGICO

#### 600 REGOLE DIVISE IN TRE CATEGORIE

- **REGOLE FLESSIONALI** : indicano quale paradigma è usato per formare la loro flessione
- **REGOLE FLESSIONALI E DERIVAZIONALI** : parole conosciute basate sul loro spelling
- **REGOLE FLESSIONALI, DERIVAZIONALI E CATEGORICO-DEDUTTIVE** : parole con radice sconosciuta o parole con nessuna radice identificabile

#### CARATTERISTICHE DELL'ANALIZZATORE MORFOLOGICO

QUESTO SISTEMA HA LA CAPACITA' DI ANALIZZARE:

- parole la cui radice è nota
- parole la cui radice è sconosciuta  
-attraverso una lista di suffissi e prefissi, risale alla radice

Friday, February 16,  
2007

29

## ANALISI DELLA FRASE

- **ESTRATTORE DI LOCUZIONI** : identificare le locuzioni che devono essere indicizzate nel testo che sta esaminando
- **UN ANALIZZATORE GRAMMATICALE** : riconoscere la struttura grammaticale delle locuzioni e trasformarle in strutture concettuali che possono essere inserite in una tassonomia concettuale

Friday, February 16,  
2007

30

## TASSONOMIA CONCETTUALE

- **ASSIMILATORE CONCETTUALE** : aggiunge parole e locuzioni nella tassonomia concettuale
- **GLI ALGORITMI DI CLASSIFICAZIONE MSS e MSG**:  
posizionano efficientemente nuovi concetti nel giusto posto nello sviluppo di una tassonomia

Friday, February 16,  
2007

31

## TASSONOMIA CONCETTUALE II

### ASSIMILATORE DI CONCETTI

Per "assimilare" una locuzione identificata e indicizzata si deve:

- Determinare la struttura concettuale della locuzione
  - Tramite il parsing della struttura sintattica della locuzione e l'inserimento nella tassonomia concettuale
  - Attraverso gli algoritmi MSS e MGS
- Determinare quali altri concetti legati a queste locuzioni potrebbero essere assimilati nella tassonomia
  - e quali relazioni dovrebbero essere stabilite tra questi concetti

Friday, February 16,  
2007

32



### TASSONOMIA CONCETTUALE III

ESEMPIO

AUTOMOBILE UPHOLSTERY CLEANING  
 genera le locuzioni  
 → AUTO CLEANING (1)  
 → UPHOLSTERY CLEANING (2)

- le due locuzioni rappresentano concetti più generali
- possono a loro volta racchiudere concetti collegati a se stessi da classificare concettualmente sotto queste categorie più generali

ES.

```

  graph TD
    AC[Auto Cleaning] --> AUC[Auto Upholstery cleaning]
    AC --> CW[Car Washing]
  
```

Sono unite sotto lo stesso concetto di astrazione

Friday, February 16, 2007 33

### TASSONOMIA CONCETTUALE IV

ORGANIZZARE UNA TASSONOMIA

- L'organizzazione di una tassonomia è un tipo di "CONCEPTUAL SORTING"

Che produce una struttura parzialmente ordinata nella quale ogni concetto punta al suo più specifico subsumer ed è puntato dal suo più generale subsumee

- Per ottenere questa struttura: si parte da una tassonomia iniziale composta da relazioni di sussunzione tra gli assiomi, e si aggiungono concetti in maniera incrementale, ogni volta posizionando il concetto vicino al suo più specifico subsumer e al suo più generale subsumee

Friday, February 16, 2007 34

### TASSONOMIA CONCETTUALE V

SVILUPPARE UNA TASSONOMIA

- Usando l'algoritmo MSS per trovare i concetti dei subsumers più specifici
- Usando l'algoritmo MGS per trovare nella tassonomia concetti più generali
- Aggiungendo i legami "kind-of" alla struttura
  - dal concetto a ognuno dei suoi più generali subsumers
  - e da ognuno dei suoi più generali subsumers al concetto, e quindi eliminare i legami che nella struttura risultano ridondanti dalle relazioni di transitività "kind-of"

Friday, February 16, 2007 35

### STRUTTURA CONCETTUALE DI UNA LOCUZIONE

ANALIZZARE LA STRUTTURA CONCETTUALE DI UNA LOCUZIONE

- -Identificare i vari ruoli che i differenti costituenti della frase giocano in essa
- -Identificare l'elemento concettuale chiave

Friday, February 16, 2007 36

## STRUTTURA CONCETTUALE DI UNA LOCUZIONE

RAPPRESENTARE LA STRUTTURA CONCETTUALE DI UNA LOCUZIONE

- Attraverso una lista di coppie : "RELAZIONE :VALORE" (R:V)
  - "RELAZIONE" : il ruolo concettuale giocato dai vari componenti della locuzione
  - "VALORE" : la descrizione dei costituenti che ricoprono questi ruoli
- gli elementi di una locuzione possono a loro volta essere rappresentati come liste di coppie "relazione:valore" arrivando così a costruire una struttura gerarchica:

**STRUTTURA CONCETTUALE**

Friday, February 16, 2007 37

## STRUTTURA CONCETTUALE DI UNA LOCUZIONE II

- **COPPIE (R:V)=**
  - o esprimono relazioni di sussunzione
  - o modificano le restrizioni

**ES.:** STEAM CLEANING  
 tipo-di (kinf-of): cleaning  
 usando (using): steam

- la relazione tipo-di è usata per esprimere una relazione di sussunzione che il concetto in questione è un tipo di pulizia
- il modificatore "using: steam" esprime una restrizione sul tipo di pulizia
- queste relazioni sono **SEMANTICHE** e non sintattiche in quanto steam e cleaning **NON** sono semplici **PAROLE MA** esprimono i **CONCETTI** di steam e di cleaning

Friday, February 16, 2007 38

## STRUTTURA CONCETTUALE DI UNA LOCUZIONE III

- **DESCRIZIONI CONCETTUALI** : lista di coppie (R:V) che descrivono la struttura di concetto di locuzione
- **descrizioni concettuali**
  - atomiche
  - composite
- **atomiche**: nome di un oggetto o una categoria di oggetti come steam, cleaning, Pisa
- **composite**: descrizione composta da altre descrizioni e rappresentata da una lista di coppie (R:V)

Friday, February 16, 2007 39

## SUSSUNZIONI CONCETTUALI

- **SUBSUMPTION** una descrizione ne include un'altra se la forma della prima è più generale seconda

**X SUBSUMES Y** se:

- X è più generale di Y
- si deduce che X include Y attraverso catene di relazioni di inclusione
- si deduce che X include Y attraverso regole di "inclusione strutturale" che esaminano le strutture delle descrizioni e usano relazioni di inclusione tra le loro parti

Friday, February 16, 2007 40

## SUSSUNZIONI CONCETTUALI II

- **(1) X È PIU' GENERALE DI Y**  
Le inclusioni testano la base di un insieme di assiomi, sulle categorie di sussunzione e di membri che provengono dalle voci scritte in un dizionario o in una base di conoscenze semantiche scritte da un lessicografo, o da un progettatore ontologico, o derivate da regole di analisi morfologica
- **(2) LE RELAZIONI DI INCLUSIONE** derivano da una chiusura transitiva di inclusione di tipo (1) e (3)
- **(3) il criterio dell'INCLUSIONE STRUTTURALE** dipende dalle relazioni di inclusioni tra i costituenti di un concetto che possono essere dedotti dagli altri due metodi

Friday, February 16,  
2007

41

## SUSSUNZIONI CONCETTUALI III

- Le sussunzioni generalizzano sia **relazioni kind-of** che **relazioni instance-of**
- Ognuna di queste relazioni può essere asserita in una tassonomia
  - dall'estrazione di assiomi,
  - derivate da un dizionario
  - come effetto di una analisi morfologica

Friday, February 16,  
2007

42

## SUSSUNZIONI CONCETTUALI IV

### **RELAZIONI DI SUSSUNZIONE CHE HANNO MODELLI CARATTERISTICI DI ESPRESSIONI:**

- (1) an X is a Y (un uomo è una persona)
- (2) an X is a kind of Y (un leone è un tipo di animale)
- (3) X is a Y (l'Italia è una nazione)
- (4) X-ing è un tipo di Y-ing (washing è un tipo di cleaning)

Friday, February 16,  
2007

43

## SUSSUNZIONI CONCETTUALI IV

- (1) e (2) corrispondono a relazioni "kind-of":
  - categoria X è un sottotipo della categoria Y
- (3) corrisponde ad una relazione "instance-of"
  - dice che un'entità X è un'istanza di una categoria Y
- (4) asserisce che una azione X è un sottotipo di un'azione Y
  - "wash" è un kind-of di "clean"

Friday, February 16,  
2007

44

### SUSSUNZIONI CONCETTUALI V

- **KIND - OF:** relazione TRANSITIVA  
 -cioè se X è un tipo di Y  
 e Y è un tipo di Z } ⇒ X è un tipo di Z
- **INSTANCE - OF:** relazione INTRANSITIVA  
 -cioè se X instance-of Y  
 e Y instance-of Z } ↯ X non instance-of Z

ES.: il capitano Ahab è il capitano di una nave, fare il capitano è una professione, ma il capitano Ahab non è una professione

- **conclusion:** la sussunzione non è transitiva perché generalizza anche le relazioni instance.of oltre a quella di kind-of

Friday, February 16, 2007 45

### SUSSUNZIONI CONCETTUALI

ESEMPIO

- "AUTOMOBILE STEAM CLEANING" sarà analizzato come segue:
  - cleaning sussume [kind-of: cleaning, object: [kind-of: car] ]
  - kind-of: maintaining, object: [kind-of: vehicle] sussumono [ kind-of: washing, object: [kind-of: car] with: [kind-of: hose] ]

Friday, February 16, 2007 46

### SUSSUNZIONI CONCETTUALI

CONCLUSIONI

- **Conclusion:** un piccolo numero di *assiomi primitivi* kind-of e instance-of inducono automaticamente ad un numero di conseguenze per le sussunzioni delle descrizioni composite. In questo modo, una taxonomy di categorie atomiche, organizzata secondo le loro relazioni di sussunzione, può automaticamente portare relazioni di sussunzione per una tassonomia più grande di descrizioni composite, senza l'aggiunta di un richiesta di input da parte dell'uomo

Friday, February 16, 2007 47

### UN'APPLICAZIONE DELL'INDICIZZAZIONE CONCETTUALE

WordNet

IL DATABASE LESSICALE

Friday, February 16, 2007 48

## COSA E' WORDNET

- **WORDNET:** è un sistema di gestione di un dizionario lessicale basato sulla categorizzazione concettuale
- WordNet conta 129.625 lemmi organizzati in 99.759 synset(=INSIEME DI SINONIMI)
- Ogni categoria sintattica ( nomi, verbi, aggettivi e avverbi) è organizzata in insiemi di sinonimi che rappresentano un concetto lessicale e che sono collegati da diversi tipi di relazioni
- E' considerato la più importante risorsa disponibile per i ricercatori nei campi della linguistica computazionale, dell'analisi testuale, e di altre aree collegate

Friday, February 16,  
2007

49

## INNOVAZIONI DI WORDNET

Questo database non si pone in competizione con i dizionari tradizionali dal punto di vista linguistico, ma pur basandosi su di un classico vocabolario presenta delle innovazioni:

- la comprensione della differenza tra lemmi e significati
- le relazioni, che permettono navigabilità alle informazioni
- la struttura interna delle categorie sintattiche

Friday, February 16,  
2007

50

## LEMMI E SIGNIFICATI

- **LEMMA:** indica la forma scritta o il suono di una parola
- **SIGNIFICATO:** indica il concetto associato ad una parola
- il punto fondamentale della semantica lessicale è il riconoscimento che esiste un'associazione convenzionale fra la forma delle parole ( LEMMA) e il concetto/significato che esse esprimono

Friday, February 16,  
2007

51

## LEMMI E SIGNIFICATI II

### RELAZIONI

In WordNet sono presenti due famiglie di relazioni a seconda del tipo degli operatori, cioè ci sono relazioni che esprimono un legame tra singoli lemmi e altre tra significati

- **RELAZIONI SEMANTICHE:** stabiliscono un nesso tra significati
- ES.: relazioni di iponimia( = rapporto semantico tra un vocabolo di significato più specifico e ristretto: **iponimo**, e uno di significato più generico ed esteso: **iperonimo**) o di metonimia (= parte/tutto)
- **RELAZIONI LESSICALI:** stabiliscono un nesso tra singoli lemmi
- ES.: un contrario può essere valido per uno dei termini di un synset ma non per gli altri

Friday, February 16,  
2007

52

## LEMMA E SIGNIFICATO III

### ASSOCIAZIONE

#### Associazione tra lemma e significato:

- è di tipo multi-a-molti
- dà luogo alle proprietà di
  - **sinonimia:** proprietà di un concetto/significato di avere due o più parole in grado di esprimerlo  
ES.: liberty  $\approx$  freedom
  - **polisemia:** proprietà di una stessa parola di avere più significati

Friday, February 16, 2007

53

## SINONIMIA

- **LA SINONIMIA:** relazione lessicale alla base della struttura di WordNet
- I termini vengono raggruppati per insiemi di sinonimi
- esiste una relazione di sinonimia implicita tra tutte le coppie di elementi appartenenti ad ogni synset

Friday, February 16, 2007

54

## SINONIMIA

- **Leibniz:** due espressioni sono sinonime se la sostituzione di una per l'altra non cambia il valore di verità della frase nella quale è fatta la sostituzione
- $\Rightarrow$  i veri sinonimi sono molto rari
- $\Rightarrow$  due espressioni sono sinonime in un contesto linguistico C se la sostituzione di una per l'altra, nel contesto C, non cambia il valore di verità della frase. ES.: plank e board sono sinonimi solo in un contesto di falegnameria
- Questa definizione di sostituibilità implica che non ci possono essere sinonimi tra categorie sintattiche diverse.
- Questo è uno dei motivi per cui i synset in wordnet sono rigidamente separati per categoria sintattica

Friday, February 16, 2007

55

## IPONIMIA

- **Iponimia:** un concetto rappresentato da un gruppo di sinonimi è un iponimo di un concetto rappresentato da un altro insieme di sinonimi se vale la  
relazione: X è un (is-a, kind-of) Y
- **Troponimia:** relazione di iponimia tra verbi:  
il verbo X è un troponimo del verbo Y se  
fare X è fare Y in qualche maniera.

Friday, February 16, 2007

56

## IPERONIMIA

- **Iperonimia:** la relazione che lega un sovraordinato, cioè un concetto generale, ad un concetto più specifico che ne eredita le caratteristiche generali differenziandosi almeno per qualcosa. Dato un synset si può trovare la chiusura transitiva degli iperonimi, cioè trovare tutti i concetti più generali del concetto in esame

Friday, February 16, 2007

57

## IPONIMIA e IPERONIMIA

### IPONIMIA/IPERONIMIA:

- sono relazioni semantiche e sono l'equivalente della gerarchia di generalizzazione/specializzazione dei modelli relazionali
- è definita solo per i nomi ed i verbi ed è la relazione di gran lunga più frequente

Friday, February 16, 2007

58

## ORGANIZZAZIONE PER CATEGORIE SINTATTICHE

- Le parole in WordNet sono divise per categorie sintattiche
- **categorie:**
  - nomi
  - verbi
  - aggettivi ed avverbi
- **SVANTAGGI:** la maggioranza dei termini ha significati in diverse categorie sintattiche, quindi per cercare i significati di una parola bisogna entrare nell'indice di ogni categoria per vedere se è presente
- **VANTAGGI:** le differenze di organizzazione semantica sono rese manifeste a livello di struttura

Friday, February 16, 2007

59

## ORGANIZZAZIONE PER CATEGORIE SINTATTICHE II

### METODO DI CATEGORIZZAZIONE DELLE CATEGORIE SINTATTICHE

- **NOMI** gerarchia di argomenti  
I nomi sono organizzati in una gerarchia di specializzazione (iperonimi/iponimi). In WordNet un iponimo eredita dai termini sovraordinati tutte le loro caratteristiche, ma se ne differenzia almeno per una
- **VERBI** una varietà di relazioni di implicazione  
I verbi sono strutturati come una gerarchia di specializzazione (troponimia, modo particolare di compiere un'azione). Per i verbi sono molto importanti le relazioni di implicazione
- **AGGETTIVI E AVVERBI** iperspazio n-dimensionale
  - Gli aggettivi sono organizzati in gruppi che contengono un synset principale e dei synset satelliti.
  - I synset principali sono collegati ai synset satelliti da una relazione di similarità e i synset satellite mediano con i synset principali tramite relazioni di antinomia

Friday, February 16, 2007

60

## ALGORITMO PER TROVARE LE RELAZIONI TRA I TERMINI

### ALGORITMO PER TROVARE LE RELAZIONI TRA COPPIE DI TERMINI

- **In input** si prende il vettore dei termini e si suppone che il loro significato sia già definito
  - Si crea un Hash di synset (= vettore dei termini sinonimi)
    - per ogni termine viene trovato un synset
    - si cerca nel Hash se è già presente:
      - se non c'è si crea l'elemento Hash
      - si aggiunge al vettore corrispondente alla chiave
- (Per come è costruito WordNet un synset può essere usato come chiave di un Hash)

Friday, February 16, 2007

61

## ALGORITMO PER TROVARE LE RELAZIONI TRA I TERMINI II

- **TROVA LE RELAZIONI TRA GLI ELEMENTI, PER OGNI ELEMENTO DEL HASH:**
  - si scrivono le relazioni di sinonimia tra gli elementi associati alla chiave attuale per ogni elemento del vettore
  - si trovano gli iperonimi
    - per ogni iperonimo si fa una ricerca nel Hash
    - se c'è si scrivono le relazioni di specializzazione con tutti i termini del vettore
  - si trovano gli iponimi e i termini correlati
    - per ognuno di questi si cerca nel Hash
    - se ci sono si scrivono le relazioni di tipo Related Term, dove cioè l'aggregazione è considerata un legame più rispetto alla specializzazione

Friday, February 16, 2007

62

## ALGORITMO PER TROVARE LE RELAZIONI TRA I TERMINI III

- **ULTIMA FASE:**
  - Elimina le relazioni tra i termini della stessa tabella
  - Elimina relazioni duplicate
- **RISULTATO:**

Un vettore di coppie di termini con la relazione che li lega

Friday, February 16, 2007

63

## CONCLUSIONI

- Tutti questi sistemi estraggono automaticamente locuzioni dal testo per indicizzarle
- Queste locuzioni sono assimilate in una tassonomia concettuale che ne registra anche la posizione nel materiale indicizzato
- Materiale Indicizzato: riferito a numerosi termini come ad es.: i documenti tradizionali, le pag. WEB, gli e-mail, ecc.
- Le risultanti tassonomie concettuali sono poi usate per supportare un algoritmo di recupero dinamico di brani che trova, segna e classifica brani specifici nel materiale che probabilmente contiene informazioni rilevanti nella richiesta
- L'obiettivo di questi sistemi è quello di saper rispondere subito ad una informazione richiesta, minimizzando il bisogno dell'utente di dover riformulare la domanda per scoprire l'esatta terminologia per trovare quello che sta cercando

Friday, February 16, 2007

64



## References

- William A. Woods. 1997. "Conceptual Indexing: A Better Way to organize knowledge"
- Jacek Ambroziak and William A. Woods. 1998. "Natural Language Technology in Precision Content Retrieval"
- William.A.Woods, Ann.Houston, Robert.Kuhns,Paul.Martin,Stephen.Green. "Linguistic Knowledge can Improve Information Retrieval"
- "Knowledge Management Needs Effective Search Technology" by William.A.Woods
- libro di testo del corso
- William.A.Woods,Stephen.Green,Paul.Martin,Ann.Houston. "Aggressive Morphology and Lexical Relations for Query Expansion"

Friday, February 16,  
2007

65

## Reference II

- Sito ufficiale dell'Università di Princeton:  
"www.cogsci.princeton.edu/-wn/"
- Nicola.Guarino,Alessandro.Oltramari. "Conceptual Analysis of Lexical Taxonomy: The Case of WordNet Top Level"

Friday, February 16,  
2007

66