WordNet & wordnets

Dott.ssa Gloria Gagliardi
Ontologia

• Definizione (Intelligenza Artificiale e Linguistica Computazionale): “rappresentazione formale di una concettualizzazione di un dominio di interesse”

• Elementi costitutivi:
  - insieme di entità
  - serie di relazioni e funzioni tra le entità

• Struttura in cui è possibile fare inferenze

• In semantica lessicale:
  - Entità = lessemi oppure concetti
  - relazioni = relazioni semantiche
WordNet: a lexical database for English

- Progetto della Princeton University [1980s]
- George A. Miller – Christiane Fellbaum
- Versione attuale: 3.1

http://wordnet.princeton.edu

Idea di base: costruire un database semantico (un dizionario *machine readable*) in cui i concetti siano rappresentati e definiti mediante set di sinonimi (*synset*).
Theoretical foundation & Design (1)

- Elemento base: SYNSET, ovvero l'insieme delle parole che esprimono un dato concetto;

- Nozione di sinonimia debole: “sostituibilità limitata ad alcuni contesti” (vs. intercambiabilità completa)

- In aggiunta: glossa e frasi esplicative
Esempio: un synset del verbo “to wind”
Theoretical foundation & Design (2)

- WordNet non è però solo una lista di concetti: i synset sono legati tra loro mediante relazioni

- Relazioni lessicali (legano i singoli lemmi, es. Sinonimia) vs.
  Relazioni semantiche (organizzano i concetti-synset in una struttura formale, es. iponimia/iperonimia)
Struttura

- 4 reti semantiche separate (una per ogni “classe aperta”):
  - nomi
  - verbi
  - aggettivi
  - modificatori
  - avverbi
## WordNet 3.1 Statistics

| POS [part of speech] | words   | synset  |
|----------------------|---------|---------|
| NOME                 | 117798  | 82115   |
| VERBO                | 11529   | 13767   |
| AGGETTIVO            | 21479   | 18156   |
| AVVERBIO             | 4481    | 3621    |
| **Totale**           | **155287** | **117659** |
Nomi

**IPONIMIA**

(*hyponymy, class inclusion, subsumption, IS_A, IS-A-KIND-OF*)

Relazione di tipo semantico (riguarda i sensi delle parole, non semplicemente le forme lessicali);

Rappresentata nel database mediante un puntatore tra synset:

\[
\{\text{robin, redbreast}\} \rightarrow \{\text{bird}\} \rightarrow \{\text{animal, animate\_being}\}
\]

Caratteristiche:

- transitività;
- asimmetria
IPERONIMIA (hyponymy)

Rappresentata nel database mediante un puntatore tra synset:

\{animal, animate\_being\} \rightarrow \{bird\} \rightarrow \{robin, redbreast\}

Caratteristiche:
- transitività;
- asimmetria.

- Iponimia e Iperonimia organizzano WordNet in una gerarchia lessicale.
- Sistema di ereditarietà lessicale tra livelli
SuperSENSI

Supersensi o “Unique Beginners” [Miller, 1998]

I nomi sono divisi in 25 gerarchie lessicali corrispondenti a campi semantici differenti

Il synset “entity” è il nodo più alto della gerarchia.

{ entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving)) }
Nomi: Altre relazioni semantiche

- **MERONIMIA (meronymy)**
  [Cruse, 1986; Pustejovsky, 1991]
  Relazione semantica tra un oggetto e i suoi costituenti o parti proprie.
  "IS A PART OF"

- Relazione inversa: **OLONIMIA (holonomy)**
  "HAS A"

- Principalmente in `noun.body e noun.artifact`
Nomi: Altre relazioni semantiche

MERONIMIA

• Tipologie di meronimia in WordNet:
  - component/object (branch \#p \rightarrow tree);
  - member/collection (tree \#m \rightarrow forest);
  - stuff/object (aluminium \#s \rightarrow airplane).
Nomi: Altre relazioni semantiche

- **ANTONIMIA (antonymy)**

  Relazione di opposizione semantica, fondamentale nell'organizzazione del lessico nominale anche dal punto psicolinguistico.

  $\text{[\{man\} \rightarrow \{woman\}] ~ [\{woman\} \rightarrow \{man\}]}
  
  $\text{[\{victory\} \rightarrow \{defeat\}] ~ [\{defeat\} \rightarrow \{victory\}]$
Verbi

- Così come i nomi, anche i verbi sono raggruppati in synset.

- Principio globale di organizzazione dei synset verbali:
  LEXICAL ENTAILMENT (implicazione)
  es. to snore * → to sleep

- Caratteristiche:
  - relazione unilaterale (se V1 implica V2, V2 non può implicare V1. Se due verbi si implicano a vicenda sono sinonimi)
Verbi: tipi di entailment [Fellbaum, 1998]

Entailment

- Temporal inclusion
  - Troponymy (proper inclusion)
    - Troponymy (co-extensiveness)
      + Troponymy (co-extensiveness)
        + Temporal inclusion
          - Temporal inclusion
            - Backward presupposition
              Cause
                - Temporal inclusion
                  + Temporal inclusion
                    es. whisper/talk
                      es. snore/sleep
                        es. forget/know
                          es. show/see
Troponimia

TROPONIMIA (troponymy) [Fellbaum & Miller, 1990]

"V₁ is V₂ in some particular manner"

- Il troponimo V₁ implica sempre il verbo più generale V₂
  
es. to march/to walk

- V₁ e V₂ sono temporalmente coestensivi

- cfr. Iponimia
2 classi:
1) aggettivo desrittivo
(es. beautiful, big)
2) aggettivo relazionale
(derivazione dal nome, es. electrical/electricity)
Descriptive adjective (1)

- Organizzazione semantica molto diversa dalle gerarchie nominali.

Relazione principale: ANTONIMIA

“IS ANTONYMOUS-TO”

(la funzione principale di questi aggettivi è di esprimere il valore di un attributo, perciò tendono ad essere bipolari)

{heavy} !→ {light}  {light} !→ {heavy}
I synset simili tra loro sono raggruppati in cluster

{quick} &→ {fast}            {fast} &→ {quick}
il synset degli aggettivi “relazionali” è collegato nel DB al nome da cui deriva.

S: (adj) stellar, astral (being or relating to or resembling or emanating from stars) "an astral body"; "stellar light"

○ pertainym

○ W: (n) star [Related to: stellar, astral] ((astronomy) a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior)
Informazioni non presenti in WordNet

• Pronuncia del lemma;
• Informazioni morfologiche;
• Etimologie;
• Note d'uso
• The “tennis” problem [Fellbaum, 1998]: non ci sono in WordNet relazioni che indicano l'appartenenza di due parole ad uno stesso “topic” del discorso:
es. racquet – ball – net
• Identificazione delle “Basic Level Categories” (nomi)
L'informazione sintattica presente in WordNet è estremamente sintetica (pattern sintattici e restrizioni di selezione)
Applicazioni

- Annotazione semantica (es. “Semantic concordance”)

- Word Sense Identification – Word Sense Disambiguation

- Information Retrieval
The Global WordNet Association

http://www.globalwordnet.org/

- Bibliografia
- Atti delle conferenze GWA
- Lista “wordnets in the world”
WordNet per la Lingua Italiana

Sono stati sviluppati autonomamente due diversi WordNet per l'Italiano:

**ItalianWordNet (MultiWordNet)**
[Fondazione Bruno Kessler - Trento]
http://multiwordnet.fbk.eu/english/whatin.php

**ItalWordNet (EuroWordNet)**
[Istituto di Linguistica Computazionale – CNR di Pisa]
http://www.ilc.cnr.it/iwndb/iwndb_php/
MultiWordNet vs. EuroWordNet

Entrambi i progetti hanno sviluppato WordNet per le maggiori lingue Europee.

- **MultiWordNet**: costruzione delle singole ontologie direttamente sulla base del Princeton WordNet 1.6. Allineamento stretto dei database lessicali: i synset sono stati creati, quando possibile, in corrispondenza con l'ontologia originale e le relazioni semantiche sono state importate dai corrispondenti synset inglesi.

- **EuroWordNet**: costruzione completamente autonoma delle gerarchie lessicali delle diverse lingue, messe in corrispondenza tra di loro e con il Princeton WordNet solo in una seconda fase.
EuroWordNet (EWN)

Coordinatore: P. Vossen

http://www.illc.uva.nl/EuroWordNet/

Lingue:
Inglese, Spagnolo, Olandese, Francese, Tedesco, Ceco, Estone e ITALIANO (ItalWordNet - IWN)
EuroWordNet – Design (1)

- Ogni WordNet è strutturato secondo le linee guida del Princeton WordNet (Synset, relazioni semantiche).
- I database lessicali sono legati tra di loro mediante una lista non strutturata e *language-indipendent* di concetti, ILI (InterLingual Index);
- Ogni synset delle ontologie è connesso almeno ad un record dell'ILI;
- Attraverso ILI è possibile procedere dalle parole di una lingua a parole simili di qualsiasi altra lingua.
EuroWordNet – Design (3)

• ILI dà anche accesso ad una Top Ontology, una gerarchia condivisa di concetti indipendenti dalla lingua, e ad una Domain Ontology (ontologia di domini semantici).

• La Top Ontology offre una cornice semantica comune per tutte le lingue (63 distinzioni semantiche), mentre proprietà specifiche della lingua sono mantenute nei wordnet individuali: gli elementi della TO possono essere lessicalizzati in vari modi (oppure non essere lessicalizzati affatto!) nelle diverse lingue [Rodriguez et al. 1998];

• Attraverso l'ILI tutti i concetti del wordnet sono direttamente o indirettamente collegati alla TO.
Ontologie lessicali basate su domini specifici, integrate al wordnet “generico” mediante relazioni “plug-in”.

- Mariterm [Marinelli et al.]
- JurWordNet
- EconomicWordNet
- ArchiWordNet
Relazioni plug-in

- la sub-gerarchia terminologica del wordnet specialistico viene connessa al nodo del wordnet generico mediante relazioni plug-in:
  - PLUG-SYNONYMY
  - PLUG-NEAR-SYNONYMY
  - PLUG-HYPONYMY
- “Procedura di oscuramento”: nasconde eventuali nodi “doppioni”
- Le due ontologie (domain-specific e wordnet genale) pur essendo integrate mantengono la loro struttura originale, e sono consultabili indipendentemente.