

WORDNET

WordNet è un sistema di riferimento lessicale on-line il cui design è basato su conoscenze psicolinguistiche sulla memoria lessicale umana. Costituisce un database di termini e di significati di termini della lingua inglese, utilizzato come strumento per ricercare i sensi delle parole. I nomi, i verbi e gli aggettivi della lingua inglese sono organizzati in **insiemi di sinonimi (synset)**, ognuno dei quali rappresenta un concetto lessicale. Diverse relazioni collegano gli insiemi di sinonimi. Dal punto di vista della gestione architeturale, WordNet è organizzato secondo una struttura a grafo, i cui nodi rappresentano gli insiemi di sinonimi e gli archi sono le relazioni fra gli insiemi. Possiamo vedere wordNet come il primo tentativo della linguistica computazionale di rappresentare i sensi

WordNet ha la funzione di un dizionario: è possibile ricercare parole e accedere al loro significato, ma ci sono delle differenze. Come prima cosa WordNet distingue i termini del lessico in 4 categorie:

- **Nomi**
- **Verbi**
- **Aggettivi**
- **Avverbi**

In secondo luogo definisce un meccanismo di ricerca basato sui significati delle parole. I dizionari, di contro, eseguono un ordinamento alfabetico dei termini. Questo comporta allo sparpagliamento dei significati, così sinonimi come scrivania e tavolo si trovano in posizioni diverse del dizionario. L'idea alla base di WordNet è avere un dizionario veloce, grazie alle capacità computazionali delle macchine che organizzano le informazioni lessicali in termini di significati delle parole. Di conseguenza, laddove è possibile definire delle gerarchie è possibile usare alberi.

- I **sostantivi** sono **strutturati in forma gerarchica**: ad esempio, la scrivania è un elemento di arredo che è un artefatto...
- i **verbi** sono organizzati da una varietà di relazioni di implicazione;
- **Aggettivi e avverbi** sono organizzati in forma vettoriale.

Terminologia

Una parola è un'associazione tra un concetto lessicalizzato e un'espressione che svolge un ruolo sintattico. Tipicamente con il termine "*parola*" ci si riferisce sia all'espressione sia al suo significato. Per ridurre l'ambiguità distinguiamo tra "**word form**" ovvero il **termine**, e "**word meaning**", ovvero il **significato**.

Matrice lessicale

Immaginiamo di usare una tabella per rappresentare termini e significati. I termini sono riportati come colonne mentre i sensi come righe di una tabella.

Word Meanings	Word Forms				
	F ₁	F ₂	F ₃	. . .	F _n
M ₁	E _{1,1}	E _{1,2}			
M ₂		E _{2,2}			
M ₃			E _{3,3}		
⋮				⋮	
M _m					E _{m,n}

Ogni cella è un incrocio di riga e colonna che indica che un certo termine può essere usato per esprimere un certo significato.

I termini sulle stesse righe sono sinonimi, mentre i termini sulle stesse colonne sono polisemici.

Come si potrebbe rappresentare un senso in un computer? Il significato M1 può essere rappresentato dalla lista di termini associati: {E11, E12.} Si usa una struttura basata su **insiemi di sinonimi**, detti anche **synset**. Tuttavia i synset non sono in grado di spiegare il significato dei termini inclusi, dimostrano banalmente l'esistenza di un concetto. In questo modo, solo le persone che hanno già una certa conoscenza della lingua inglese e quindi hanno un certo background culturale sono in grado di riconoscere il significato dei termini in un synset.

L'elemento minimale per distinguere un senso da un altro è detto **glossa**. In alcuni casi i sinonimi non sono disponibili oppure abbiamo synset con un solo membro, così a fianco al termine si associa una descrizione usata per disambiguare. Ad esempio: quella tra parentesi è una glossa

- **house** (*a dwelling that serves as living quarters for one or more families*) "he has a house on Cape Cod"; "she felt she had to get out of the house"
- **house** (*the members of a religious community living together*)

Relazioni semantiche

WordNet è organizzato in relazioni semantiche. Poiché una relazione semantica è una relazione tra significati, e poiché i significati possono essere rappresentati da synset, viene naturale pensare alle relazioni semantiche come puntatori tra synset. Da un altro punto di vista possiamo vederle come archi di un grafo che connettono i nodi rappresentati da synset.

Ci sono 4 relazioni semantiche in Wordnet:

- **Sinonimia**: due termini sono sinonimi se la sostituzione di un per l'altro in una frase non ne modifica il significato effettivo, dato uno specifico contesto.

- **Antonimia:** due termini sono antonimi quando esprimono un senso opposto. Ad esempio: ricco e povero. Tuttavia non è solo una relazione di tipo semantico. Il significato dei termini *{rise, ascend}* e *{fall, descend}* può essere opposto ma non sono in relazione di antonimia. Infatti [rise, fall] e [ascend, descend] sono antonimi.
- **Iponimia:** due termini sono iponimi quando esprimono una relazione di sottoclasse. Li possiamo vedere come un nodo padre e un nodo figlio nell'albero di WordNet. L'iponimo, (sottoclasse) eredita tutte le caratteristiche della superclasse e definisce un qualche differenza col genitore.
- **Meronomia:** due termini sono in relazione di meronomia quando uno esprime una parte dell'altro. Esempio ruote e macchina.

Queste relazioni formano una complessa rete di synset. Conoscere dove si trova una parola nel grafo costituisce una parte importante per accedere al significato della parola stessa.

Gerarchia dei sostantivi

I sostantivi sono organizzati in modo gerarchico. Quando si cerca un sostantivo, WordNet restituisce un sostantivo padre, (elemento superordinato) più le caratteristiche distintive del sostantivo cercato. (ad esempio, una cattedra è un elemento da arredo che serve per fare lezione). Queste informazioni forniscono la base per l'organizzazione dei nomi in wordNet. La struttura è ad albero, ogni nodo-parola eredita tutte le caratteristiche dei nodi-parola genitori. Ci sono 3 tipi di caratteristiche ereditate:

- attributi (aggettivi)
- parti (sostantivi)
- funzioni (verbi)

WordNet tende a coprire tutta la lingua inglese. Ci sono 25 alberi iniziatori che fungono da elementi primitivi. Si possono capire le potenzialità di wordNet mettendolo a confronto con un semplice dizionario elettronico. Ad esempio supponiamo di ricercare il termine "albero":

albero: [àl-be-ro] s.m.

1 Ogni pianta perenne con fusto legnoso (tronco) ramificato: *a. da frutto; piantare un a.; chioma dell'a.*

2 È usato nella denominazione di molte varietà di piante || a. del latte, albero presente nella fascia tropicale americana, dalla cui corteccia incisa fuoriesce un lattice commestibile | a. del pane, albero tropicale con frutti ricchi di amido | a. di Giuda, pianta con fiori rossi a grappolo

Quali potrebbero essere le informazioni a cui siamo interessati ad avere e che questo strumento non fornisce?

- In primo luogo l'elemento superordinato non è definito, manca una definizione. Il senso di un termine radice dell'albero dei termini non c'è. Risalire verso il senso base comporta una ramificazione su altri termini.

- In secondo luogo mancano i termini coordinati.
- Infine mancano i riferimenti verso il basso da termine a termine.

WordNet risolve questi problemi attraverso la struttura ad albero.

Le informazioni di WordNet

Abbiamo introdotto WordNet come un database di termini e di sensi che, differentemente da un dizionario che realizza un ordinamento lessicografico dei termini, introduce questa nozione di synset. I synset sono insiemi di sinonimi che costituiscono i nodi di una rete, o albero, indotto dalla relazione isA. Sulla base di altre relazioni semantiche come iponimia, antonimia e meronimia i nodi modificano la precedente struttura ad albero per generarne una a grafo. (Quindi con la sola relazione isA WordNet è strutturato in alberi.) Infine abbiamo visto le differenze di uso tra un dizionario elettronico e WordNet. Abbiamo notato che il meccanismo definitorio dei dizionari è basato sulla dichiarazione, per ogni termine cercato, di una classe padre e di caratteristiche tipiche del termine cercato(???). Che cosa manca nel sistema di definizioni di un dizionario? Non è presente una specificazione compiuta degli elementi utilizzati, se si definisce l'albero come una pianta ci si pone il problema di capire che cosa vuol dire pianta. Questo risalire non è efficiente. Ogni volta che si risale la gerarchia si è costretti a disambiguare le superclassi. Non vengono specificati le informazioni sugli

iponimi, ovvero tutte le sottoclassi. (quali sono i tipi di albero?) Tutte informazioni che sono presenti nel dizionario ma sono sparpagliati. Nei dizionari ci sono dei cerchi, il fatto di avere delle definizioni circolari è accettabile da un essere umano ma non ad una macchina che deve elaborare il linguaggio naturale.

Proviamo a trovare il termine "tree" su WordNet.

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (frequency) {offset} <lexical filename> <lexical file number>
(gloss) "an example sentence"
Display options for word: word#sense number (sense key)

Noun

- {107}{13124818} <noun.plant>[20] [S: \(n\) tree#1 \(tree%1:20:00::\)](#) (a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms)
- {13935275} <noun.shape>[25] [S: \(n\) tree#2 \(tree%1:25:00::\)](#), [tree diagram#1 \(tree_diagram%1:25:00::\)](#) (a figure that branches from a single root) "genealogical tree"
- {11368155} <noun.person>[18] [S: \(n\) Tree#3 \(tree%1:18:00::\)](#), [Sir Herbert Beerbohm Tree#1 \(sir_herbert_beerbohm_tree%1:18:00::\)](#) (English actor and theatrical producer noted for his lavish productions of Shakespeare (1853-1917))

Verb

- {01938064} <verb.motion>[38] [S: \(v\) corner#2 \(corner%2:38:01::\)](#), [tree#1 \(tree%2:38:00::\)](#) (force a person or an animal into a position from which he cannot escape)
- {01619197} <verb.contact>[35] [S: \(v\) tree#2 \(tree%2:35:01::\)](#) (plant with trees) "this lot should be treed so that the house will be shaded in summer"
- {01147629} <verb.competition>[33] [S: \(v\) tree#3 \(tree%2:33:00::\)](#) (chase an animal up a tree) "the hunters treed the bear with dogs and killed it"; "her dog likes to tree squirrels"
- {00319912} <verb.change>[30] [S: \(v\) tree#4 \(tree%2:30:00::\)](#), [shoetree#1 \(shoetree%2:30:00::\)](#) (stretch (a shoe) on a shoetree)

La risorsa restituisce l'elenco dei synset associati alla risorsa cercata, prima sono elencati i sostantivi, poi i verbi. Ci sono un certo numero di informazioni correlate come:

- Fra parentesi tonde, il numero di volte in cui il termine cercato compare nel database di WordNet. Il primo elemento nella lista è il più frequente.
- Una glossa, ovvero delle definizioni abbastanza sintetiche per distinguere un termine da un altro.
- Fra parentesi graffe viene riportato l'id del nodo nel grafo di WordNet associato ad un certo synset.
- Esempi di uso come frasi dove è possibile vedere come viene usato il termine in una certa accezione cercato.
- Fra parentesi angolari vengono riportati i supersensi. Questi rappresentano gli iniziatori, ovvero le primitive semantiche, le radici dei sottoalberi in cui la risorsa è partizionata.

Inoltre è possibile navigare tutte le accezioni semantiche, quindi si possono esplorare le parti dell'albero.

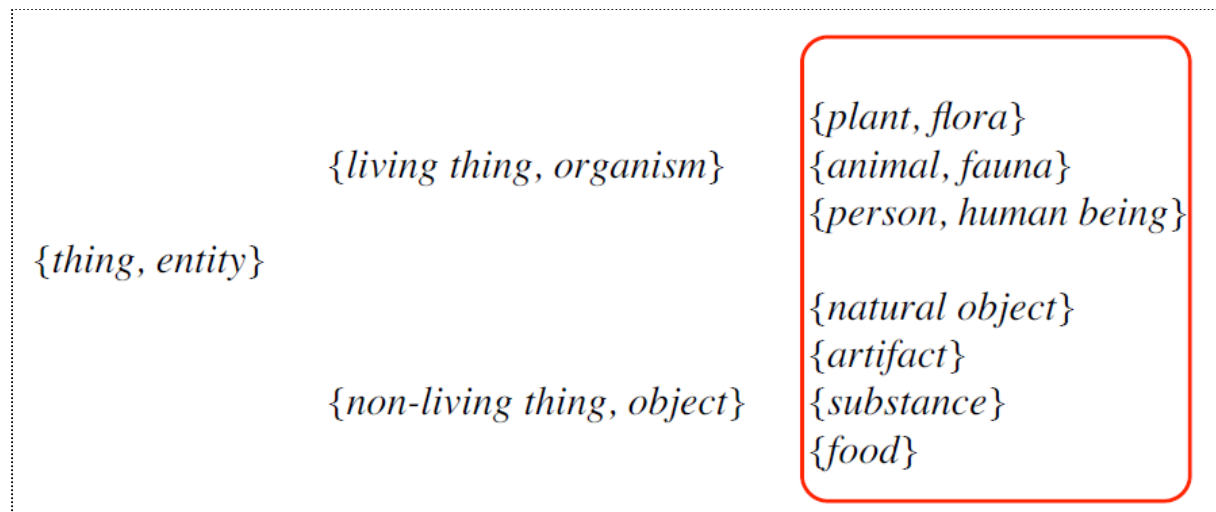
Organizzazione di wordNet

Da un punto di vista generale, come è articolato wordNet? Abbiamo visto che esistono dei supersensi. I supersensi sono degli iniziatori delle primitive semantiche, ovvero radici di altri sottoalberi in cui è partizionata la risorsa di WordNet e che idealmente ricoprono tutto il vocabolario della lingua inglese. Esistono 25 supersensi per i sostantivi e 15 per i verbi. I supersensi corrispondono a regioni semantiche distinte, ciascuna dotata di un proprio vocabolario. Nel partizionamento di tutti i termini basato su sensi, i termini possono essere spostati in vari campi semantici a seconda di quali sono le primitive semantiche (supersensi). Esempi di supersensi per i sostantivi: (i supersensi sono anch'essi dei synset)

- {act, action, activity}
- {animal, fauna}
- {artifact}
- {attribute, property}

E' una multigerarchia piatta che distingue elementi astratti da elementi concreti. Alcuni elementi colgono sia i processi (perdurant) sia elementi di concretezza (endurant). Presi assieme, questi 25 elementi forniscono una collocazione per tutti i termini della lingua inglese. I 25 supersensi possono essere ulteriormente gerarchizzati come figli di 7 iniziatori di primitive semantiche:

{plant, flora}, {animal, fauna}, {person, human, being}, {natural object}, {artifact}, {substance} {food}. Da questo punto in poi possono essere collocati gli altri elementi sottostanti.



Caratteristiche distintive

Quando si cerca un termine su WordNet, la risorsa restituisce una superclasse e delle caratteristiche distintive del termine cercato. Come funziona questo meccanismo di specificazione dei tratti distintivi? Esiste un livello intermedio di descrizione delle entità del mondo a cui la gran parte delle caratteristiche distintive sono associate e queste sono dette categorie di livello base oppure concetti generici. Al di sopra di questo livello di base le descrizioni sono brevi e generali e al di sotto sono più specifiche ovvero vengono aggiunti dei tratti distintivi. WordNet ci permette di estrarre informazioni relativamente agli **attributi** alle **parti** e alle **azioni** dei sostantivi.

Attributi e modificatori

Gli attributi sono aggettivi: linguisticamente rappresentano i modificatori dei sostantivi ovvero specificano ulteriormente il significato dei sostantivi. Le parti sono altri sostantivi che definiscono relazioni di meronimia e le azioni sono funzioni rappresentate da verbi. Ad esempio: il canarino è piccolo, giallo, e vola. In questo caso “volare” è un’azione compiuta dal sostantivo. Il canarino viene inserito nella risorsa non solo come elemento iponimo di uccello, da cui eredita tutti i tratti ma in più specificando dei vari elementi ovvero per esempio che ha un becco e delle ali costituite da penne. I valori degli attributi sono espressi tipicamente dagli aggettivi. Ad esempio: “Size(canary) = small”. Gli aggettivi modificano i sostantivi e i nomi funzionano da argomenti per gli attributi.

Parti e meronimia

La meronimia definisce una relazione parte-tutto tra sostantivi. E' considerata una relazione semantica al pari della sinonimia, iponimia, antonimia. Se A è un meronimo di B vuol dire che A descrive una parte di B, come ad esempio le ruote(A) e l'automobile(B). La relazione ha anche un inverso: B si dice olonimo di A. La relazione di meronimia è transitiva. In alcuni casi la transitività appare insensata. Ad esempio, dire che la maniglia è una parte della porta e che la porta è una parte della casa ha senso, ma dire che la maniglia è parte della casa non è proprio giusto.

Funzioni e azioni

Le funzioni denotano azioni compiute da sostantivi. Rappresenta una descrizione di qualcosa che le istanze di un concetto normalmente compiono. Ad esempio: il coltello ha la funzione di tagliare, la matita ha la funzione di scrivere, ecc... In alcuni casi, le funzioni sono un pò forzate: dire che un canarino ha la funzione di cantare non ha molto senso. In questo caso, più che parlare di funzioni si può parlare di *"ruoli semantici"*.