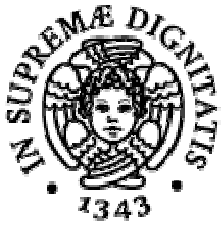


UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Revisone ed estensione di un corpus annotato con supersensi per la lingua italiana

Candidato: *Giulia Di Pietro*

Relatore: *Chiar.mo Professor Alessandro Lenci*

Correlatore: *Chiar.ma Professoressa Maria Simi*

Anno Accademico 2008-2009

Alla mia famiglia,

sole di ogni mio giorno.

1 INDICE

2 INTRODUZIONE

«Quando uso una parola», Humpty Dumpty disse in tono piuttosto sdegnato, «essa significa esattamente quello che voglio – né di più né di meno.» «La domanda è», rispose Alice, «se si può fare in modo che le parole abbiano tanti significati diversi.»

«La domanda è,» replicò Humpty Dumpty, «chi è che comanda – tutto qui.»[1]

Quando si parla a volte non si fa caso al fatto che capita di ripetere la stessa parola ma attribuirle un significato differente, e non si fa abbastanza caso a quanto semplice sia per il nostro interlocutore comprendere a quale dei tanti significati della parola ci stiamo riferendo, e quanto difficile sia spiegare come questo avviene.

Humpty Dumpty dice che è lui a comandare sulle parole, e loro si piegano al significato che vuole che esse prendano. In questo modo è l'essere umano a definire ogni significato, in base alle proprie conoscenze; ciò mette in luce come una parola sia segno che porta con sé dei significati che devono essere identificati dagli ascoltatori, nelle varie situazioni comunicative.

Al giorno d'oggi, grazie allo sviluppo crescente, e in continua evoluzione, di Internet, la maggior parte dell'informazione viene condivisa sotto forma testuale digitale. Se necessitiamo di qualche informazione basta scrivere le parole chiave su un motore di ricerca e, se queste sono state selezionate in maniera adeguata, troveremo in poco tempo l'informazione che cerchiamo. Però questo non sempre funziona, proprio per il motivo sopra citato: le parole sono segni, che portano con sé significati diversi.

Ma in questo modo, dato che la comunicazione avviene tra due macchine, chi attribuirà ad ogni singola parola il giusto valore semantico? A decifrare la correttezza dei valori di output saremo noi esseri pensanti, ma nel momento in cui viene dato l'input un computer riesce a decifrare il valore semantico che si nasconde dietro ad una parola?

[1] Lewis Carroll, *Attraverso lo specchio e quel che Alice vi trovò.* Trad. Silvio Spaventa Filippi. 1871.

Ovviamente no, se si lascia che la parola mantenga il suo ruolo primario di segno. Ma se si provasse a riconoscere il significato che una parola ha in un dato contesto, allora allora potremmo risolvere gran parte dei problemi di ambiguità semantica.

Ed è proprio da qui che è nato WordNet [2], dall'esigenza di fornire una rappresentazione computazionale del significato delle parole.

La caratteristica peculiare di WordNet è la sua struttura, basata su relazioni semantiche tra synset[3]. Le relazioni semantiche al suo interno sono strutturate gerarchicamente, in maniera tale da avere dei nodi figli che ereditano le proprietà dei nodi padri[4].

Dato il successo di WordNet – utilizzato per svariate applicazioni, quali il question answering, la traduzione automatica, il semantic web – nel 1996 è iniziato il progetto EuroWordNet[5], mirante alla costruzione di un database semantico multilingue delle lingue europee, al quale progetto ha partecipato anche l'Istituto di Linguistica Computazionale (ILC) del C.N.R. di Pisa. I WordNet all'interno di EuroWordNet sono collegati al WordNet di Princeton attraverso l'Inter-Lingual Index, ovvero un indice che permette l'allineamento tra i lessici delle varie lingue e WordNet.

Nell'ambito del progetto TAL, il database di WordNet italiano è stato ulteriormente espanso, con il nome di ItalWordNet, sempre seguendo la struttura e i principi architetturali di EuroWordNet.

All'interno del progetto TAL è stato sviluppato anche l'ISST[6], un corpus annotato con i sensi di ItalWordNet, per permettere l'addestramento di algoritmi che poi avrebbero proceduto in maniera automatica all'annotazione semantica di testi.

---

[2] WordNet è una rete semantica realizzata per l'inglese da George A. Miller e il suo gruppo a Princeton.
[3] Gruppi di parole legati da una relazione di sinonimia.
[4] es. Nella gerarchia di WordNet si avrà "Pettirosso" figlio del nodo "uccello", che è figlio del nodo "animale", percui un pettirosso erediterà le caratteristiche proprie di ogni "uccello", e in quanto "uccello" erediterà le caratteristiche che accomunano ogni "animale".
[5] Progetto finanziato dalla Comunità Europea all'interno del Programma "Language Engineering", protrattosi da marzo 1996 a luglio 1999.
[6] ISST, Italian Syntactic-Semantic Treebank.

Proprio da qui prende le mosse il progetto "*SemaWiki*[7]", al quale abbiamo preso parte io e il Dottor Dei Rossi, che mira all'annotazione semantica della parte italiana di Wikipedia.

Il dottor Dei Rossi, per prima cosa, ha provveduto, ad applicare alle parole contenute nell'ISST, alle quali erano applicati i sensi di ItalWordNet, i Supersense del WordNet di Princeton, sfruttando i collegamenti ILI tra ItalWordNet e WordNet .

Il mio compito, invece, è stato di revisionare il corpus nella fase successiva all'assegnazione dei Supersense – avvenuta in maniera automatica –, e assegnare un Supersense adeguato alle parole che ne erano sprovviste.

Nei capitoli che seguiranno verranno approfonditi i problemi della disambiguazione semantica e dei corpora annotati a livello semantico. L'ultima parte sarà occupata da un'analisi del processo di revisione ed estensione del corpus annotato che è oggetto del presente lavoro.

[7] Progetto il cui scopo è l'annotazione a livello semantico di testi all'interno del Wikipedia italiano.

3 LA DISAMBIGUAZIONE SEMANTICA

La disambiguazione semantica, per la maggior parte dei casi, non è fine a sé stessa, in quanto la maggior parte dei task della Linguistica Computazionale per essere portati a compimento hanno bisogno di risolvere la polisemia, ovvero il problema rappresentato dal fatto che molte parole hanno più di un significato, che cambia a seconda del contesto. Verrà subito alla mente com'è ovvio che per avere successo con le applicazioni di comprensione di linguaggio, come software per la comunicazione uomo-macchina, sia necessario risolvere in maniera soddisfacente la questione della disambiguazione.

Attualmente risulta necessaria per vari tipi di task e applicazioni, quali:

Traduzione automatica: la disambiguazione dei sensi di una parola è necessaria se vogliamo avere una traduzione appropriata di parole polisemiche; è il caso della parola *pesca*, la quale in base al contesto verrà tradotta in inglese o *peach*, o *fishing*.

Recupero d'informazione e navigazione di ipertesti: quando si cerca una parola chiave, sarebbe opportuno che non fossero ritenuti validi risultati la parola o le parole che compaiono in sensi inappropriati alla ricerca effettuata.

Information extraction: effettuando una buona disambiguazione si facilita l'information extraction, il cui compito è di derivare informazioni ben strutturate partendo da documenti *machine-readable* non strutturati.

Elaborazione del parlato: la disambiguazione semantica è richiesta anche per avere la corretta pronuncia delle parole con i sintetizzatori vocali. Ad esempio, "principi" o "ancora" sono parole che in base al significato che assumono cambiano l'accento. La questione assume maggiore rilevanza per la lingua inglese, dove alcune parole cambiano proprio assetto fonetico.

Il problema della disambiguazione semantica è stato addirittura definito "AI-complete", cioè, un problema che può essere risolto solamente risolvendo prima i principali problemi dell'AI (Artificial Intelligence), come la rappresentazione del senso comune e la conoscenza enciclopedica.

Nonostante ciò, sono stati pensati e messi in atto diversi metodi per risolvere questo task. Il problema più complesso è la *Word Sense Disambiguation*, la

quale mira ad attribuire ad ogni parola il suo senso specifico; un task correlato da esso è la *Named Entity Recognition*, la quale ha come obiettivo di identificare all'interno di un testo nomi propri di persone, organizzazioni e luoghi (e non solo); l'ultimo, e il più interessante problema nel contesto di questo lavoro, è la *SuperSenses Disambiguation*, la quale si può ritenere una semplificazione della WSD in quanto vuole attribuire alle parole non sensi specifici, bensì "supersensi", ovvero tipi semantici generici e più astratti rispetto ai sensi specifici.

3.1 WordSense Disambiguation

Quando negli anni '40 del secolo scorso si iniziò a parlare di traduzione automatica, una delle difficoltà che emersero fu proprio la disambiguazione semantica, e fu in quel frangente che venne definito il problema della Word Sense Disambiguation a livello computazionale. Ovviamente, non esistendo ancora risorse lessicali elettroniche – che specificassero i repertori di sensi di una parola e in formato machine readable – dalle quali attingere informazioni, l'annotazione avveniva solamente per via manuale. Dagli anni '80 in poi, quando l'*Oxford Advanced Learner's Dictionary of Current English* fu reso disponibile, l'operazione di annotazione subì una svolta, in quanto l'annotazione manuale venne rimpiazzata dall'annotazione automatica, sulla base della "conoscenza" estratta dalle risorse che erano nate e stavano nascendo.

L'approccio della Word Sense Disambiguation è tanto semplice a livello teorico quanto complicato a livello pratico. Il problema consiste nell'assegnare ad ogni parola un senso specifico appropriato al contesto. I metodi usati per l'assegnazione dei sensi sono molteplici. Ecco alcuni esempi.

*3.1.1 **Approcci basati su algoritmi di apprendimento supervisionati** (Supervised-Learning Approaches)*

L'approccio supervisionato alla disambiguazione semantica si può ritenere il più semplice a livello formale. L'approccio più comune utilizza il **classificatore di Bayes** (Domingos e Pazzani, 1997), e questo prevede una risorsa di apprendimento *(training-set)* che contenga al suo interno delle liste di parole *feature-encoded*, ovvero espresse nei loro contesti più comuni, alle quali è già stata assegnata l'etichetta *(label)* che ne identifica il senso appropriato. A partire da una simile risorsa, sarà poi un algoritmo che, basandosi su un modello statistico creato a partire dal training-set, assegnerà ad ogni nuova parola il senso corretto all'interno del contesto, tenendo in considerazione le feature.

Ad esempio, considerando la frase:

*"An electric guitar and **bass** player stand off to one side"*.

La parola *bass* in inglese identifica sia il *basso*, come strumento musicale, che un pesce, la *spigola*.

Rule		Sense
fish within window	⇒	**bass**^{1}
striped bass	⇒	**bass**^{1}
guitar within window	⇒	**bass**^{2}
bass player	⇒	**bass**^{2}
piano within window	⇒	**bass**^{2}
tenor within window	⇒	**bass**^{2}
sea bass	⇒	**bass**^{1}
play/V *bass*	⇒	**bass**^{2}
river within window	⇒	**bass**^{1}
violin within window	⇒	**bass**^{2}
salmon within window	⇒	**bass**^{1}
on bass	⇒	**bass**^{2}
bass are	⇒	**bass**^{1}

Figura 1

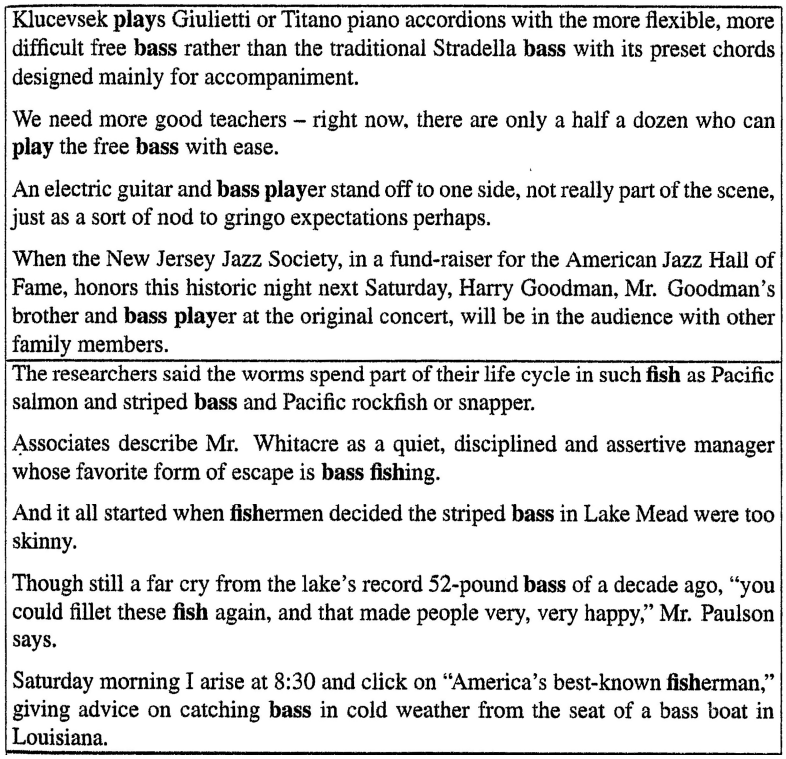
Seguendo la metodologia sopra descritta, per disambiguare *"bass"* esisterà un training set come quello della figura 1, il quale mostra (in parte) la lista di possibili occorrenze della parola *"bass"* operando una distinzione tra il senso *"spigola"* e il senso *"basso"*. La prima riga della tabella rende esplicito che se la parola *"fish"* occorre all'interno del contesto di input, allora il senso appropriato per la parola *"bass"* in quel dato contesto, sarà certamente *"spigola"*. La stessa cosa alla riga 3; se la parola che compare all'interno del contesto di input è "guitar", ovviamente il senso sarà quello di *"basso"*.

3.1.2 ***Approcci incrementali*** *(Bootstrapping Approaches)*

Questo metodo elimina uno dei problemi principali dell'approccio supervisionato, ovvero la necessità di basarsi su un training set. Questo infatti si basa su un numero ristretto di esempi sui sensi delle parole d'interesse. Questi esempi, una volta etichettati, saranno usati come *semi (seed)* per addestrare un classificatore iniziale. Successivamente, quest'ultimo verrà utilizzato per estrarre un training set più grande dal restante corpus non annotato. Ripetendo questo processo si otterrà maggiore precisione e maggiore copertura: per fare in modo di avere buoni risultati, il processo deve agire solo dove il classificatore iniziale ha una percentuale di accuratezza molto alta, e così facendo, ad ogni iterazione il training corpus avrà sempre più informazioni sulle quali basarsi, e la parte non ancora annotata andrà pian piano riducendosi. Questo processo iterativo può essere ripetuto fin quando non si raggiunge una percentuale di errore inferiore ad una certa soglia.

Per quanto riguarda la scelta dei *seed* iniziali, si può fare ricorso a diverse strategie. Quella che porta maggiori vantaggi, consiste nel creare a mano un piccolo set di esempi all'interno del corpus iniziale. Ovviamente in questo modo si ha la certezza che i *seed* siano corretti.

Una tecnica alternativa, chiamata vincolo **One sense per collocation** (Yarowsky, 1995), la quale da risultati molto buoni, mira a cercare costruzioni, all'interno di un corpus, che contengano parole o sintagmi, fortemente associate con il senso che ci interessa. Un esempio, potrebbe essere di voler avere un set di frasi seed per entrambi i sensi di "*bass*", pesce e strumento musicale. Potremmo operare dunque ritenendo che la parola *"fish"* abbia un forte legame con un senso di *bass* e che la parola *"play"* lo abbia con un altro senso di *bass*.


Klucevsek **play**s Giulietti or Titano piano accordions with the more flexible, more difficult free **bass** rather than the traditional Stradella **bass** with its preset chords designed mainly for accompaniment.

We need more good teachers – right now, there are only a half a dozen who can **play** the free **bass** with ease.

An electric guitar and **bass play**er stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.

When the New Jersey Jazz Society, in a fund-raiser for the American Jazz Hall of Fame, honors this historic night next Saturday, Harry Goodman, Mr. Goodman's brother and **bass play**er at the original concert, will be in the audience with other family members.

The researchers said the worms spend part of their life cycle in such **fish** as Pacific salmon and striped **bass** and Pacific rockfish or snapper.

Associates describe Mr. Whitacre as a quiet, disciplined and assertive manager whose favorite form of escape is **bass fish**ing.

And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny.

Though still a far cry from the lake's record 52-pound **bass** of a decade ago, "you could fillet these **fish** again, and that made people very, very happy," Mr. Paulson says.

Saturday morning I arise at 8:30 and click on "America's best-known **fish**erman," giving advice on catching **bass** in cold weather from the seat of a bass boat in Louisiana.

Figura 2

Come mostra la figura 2, questi sono i risultati parziali della ricerca delle stringhe *"fish"* e *"play"* all'interno di un corpus nel quale compare la parola *bass*.

Questo metodo da risultati molto buoni, tanto che Yarowsky (1995) registra una precisione media del 96,5%.

*3.1.3 **Approcci basati su dizionario** (Dictionary-Based Approaches)*

La prima implementazione degli approcci basati sull'utilizzo dei dizionari è dovuta a Lesk (1980). Con questo metodo, per effettuare la disambiguazione di una data parola all'interno di un contesto, bisogna effettuare una ricerca di questa in un dizionario. Le parole che compariranno nelle definizioni dei vari sensi verranno confrontate con le parole che compaiono nelle definizioni delle rimanenti parole all'interno del contesto. Se le intersezioni degli insiemi di parole estratte dalle definizione non sono vuote, i sensi corrispondenti delle

parole saranno scelti come sensi appropriati per il contesto. Riporto un esempio dello stesso Lesk, che mira a dare un senso appropriato a *"cone"*, pigna, nella locuzione *"pine cone"*, partendo dalle seguenti definizioni di *pine* e *cone*:

pine
Kinds of evergreen tree with needle-shaped leaves;
Waste away trhough sorrow or illness.
cone
Solid boby wich narrows to a point;
Something of this shape whether solid or hollow;
Fruit of certain evergreen trees.

In questo esempio, il metodo di Lesk selezionerebbe il terzo senso della parola *cone*, in quanto sia tra le definizioni di *pine* e le definizioni di *cone* compaiono due parole in comune, ovvero *evergreen* e *tree*.

Uno dei principali problemi di questo approccio è che spesso le definizioni che si trovano nei dizionari sono relativamente brevi, e potrebbero non avere sufficiente materiale per la creazione di classificatori adeguati.

Un modo per rimediare a questo problema è di inserire all'interno del classificatore parole attinenti al contesto, ma che non compaiono nelle definizioni dei sensi delle parole.

3.2 Named Entity Recognition[8]

Il termine "Named Entity", fu coniato nel Novembre del 1995, in occasione del MUC-6, ovvero la sesta edizione della *Message Understanding Conferences*. Uno degli obiettivi della conferenza era quello di proporre lo studio di nuovi metodi di estrazione di informazione da testi non strutturati, come ad esempio articoli di giornale. Ciò evidenzia che la Named Entity Recognition non si sia sviluppata partendo dal problema della disambiguazione semantica, bensì abbia un ruolo molto importante all'interno dell'Information Extraction, del quale appunto è un sottotask. Oggetto della Named Entity Recognition è dunque l'estrazione di entità nominali all'interno di un testo, e in particolar modo si

[8] Riconoscimento di entità nominali

occupa dell'individuazione di nomi propri o espressioni linguistiche che designano entità individuali.

3.2.1 Le categorie delle Named Entities

Lo scopo della Named Entity Recongnition è proprio quello di individuare e classificare unità informative all'interno del testo secondo categorie predefinite. Le entità vengono messe in evidenza con il formato XML descritto nelle MUC, ovvero "ENAMEX" per i nomi, "NUMEX" per le entità numeriche, "TIMEX" per le entità temporali. Ad esempio, se teniamo in considerazione la frase

Jim bought 300 shares of Acme Corp. in 2006

il codice XML con le entità annotate potrebbe essere:

<ENAMEX TYPE="PERSON">**Jim**</ENAMEX> bought <NUMEX TYPE="QUANTITY">**300**</NUMEX> shares of <ENAMEX TYPE="ORGANIZATION">**Acme Corp.**</ENAMEX> in <TIMEX TYPE="DATE">**2006**</TIMEX>.

Le categorie di base, in questo caso, sono:

Names (enamex);

Organization;

Person;

Location;

Times (timex) ;

Date;

Time;

Numbers (numex);

Money;

Percent.

Esistono poi delle categorie che possono anche essere ritenute, forse più propriamente sottocategorie, ovvero:

Distance;

Speed;

Age;

Weight;

City;

Country;

State/Province;

River.

Ovviamente, il numero di categorie può essere aumentato o diminuito in base alle necessità di un dato progetto di ricerca. Se ad esempio dovessimo analizzare un testo per operare una classificazione geografica potrebbe essere necessario classificare ogni entità della categoria *Location* come particolare tipo di località.

Alcuni studi recenti non pongono limiti a quelle che potrebbero essere le categorie delle entità, lasciando largo spazio a dettagli sempre maggiori, introducendo così categorie quali *"museum","river"* ed *"airport"*, come *"product"* ed *"event"*, e come *"animal"*, *"religion"* e *"color"*. All'interno di un progetto di ricerca si è cercato di individuare i tipi di nomi che compaiono più frequentemente in articoli di quotidiani. Il numero di categorie risultante è stato di ben 200 (Sekine & Nobata, 2004).

Per quanto questo possa sembrare un grande passo avanti nella ricerca sulla *Named Entity Recognition*, bisogna notare che l'indiscriminata introduzione di nuove categorie di entità può creare grossi problemi, in quanto

3.2.2 I metodi per operare

Il principale metodo per attuare la Named Entity Recognition è basato sull'utilizzo di risorse di conoscenza, i Gazetteer. Esiste anche un altro metodo, fondamentalmente statistico, che fa utilizzo di corpora annotati per addestrare algoritmi d'apprendimento.

3.2.2.1 Metodo basato su risorse di conoscenza (Knowledge-Based Method)

Questo metodo utilizza risorse esplicite, come regole o gazetteers, le quali spesso sono create manualmente. Si può fare una distinzione tra due tipi di gazetteers. Da una parte vi sono i gazetteer, all'interno dei quali si trovano parole chiave che indicano la possibile presenza di una entità nominale, come ad esempio la parola *"Ms."* (l'italiano "Sig.ra"), la quale prevede quasi certamente che dopo vi sia una *Person Entity*.

Word + {City, Town} → Word = location:city

es: Cape City, Campbell Town, NY City, etc…

Word + {Street, Road, Avenue, Boulevard}

→ Word = location:street

es: Portobello Road, Fifth Avenue, etc…

Dall'altra parte vi sono *entity gazetteer*, i quali contengono essi stessi delle entità – molto spesso nomi propri – che serviranno ad assegnare la giusta categoria alle parole riscontrate nel testo che sono contenute all'interno del gazetteer.

List of Nation	
Location	*Nation*
Italy, France, Spain, Portougal...	

List of Currency		
Measure	*Money*	*Quantity*
Euro, Dollar, Pound...		

3.3 SuperSense Disambiguation

Fino ad ora abbiamo visto come sia possibile in parte risolvere il problema dell'ambiguità semantica attraverso la *Word Sense Disambiguation* e la *Named Entity Recognition*. La SuperSense Disambiguation, a differenza di entrambe – che mirano rispettivamente all'assegnazione di un senso specifico l'una ad ogni

parola e l'altra ai nomi propri riscontrati in un testo – punta all'assegnazione non di un senso, bensì di un supersenso, ovvero una categoria semantica più generale rispetto agli specifici sensi di una parola.

A tal proposito, diventa necessario parlare di WordNet, il più importante database semantico-lessicale per la lingua inglese.

3.3.1 WordNet

Lo sviluppo di WordNet iniziò nel 1985. Lo scopo di George Miller e di altri professori e ricercatori della Princeton University era quello di strutturare la conoscenza attraverso una rete semantica. Proprio nel 1985 alcuni psicologi cognitivi e linguisti computazionali stavano cercando di modellare il significato delle parole in termini di *reti semantiche*, diagrammi nei quali i nodi rappresentavano i significati e le frecce rappresentavano le relazioni tra i significati. Fu proprio a partire da questa teoria che nacque l'architettura di WordNete.

Secondo la teoria sopra descritta devono essere esplicitate le relazioni tra i significati delle parole, per fare in modo di agire non sulle singole parole bensì sui significati di queste. Le prime e più semplici relazioni però, devono essere effettuate tra singole parole, in modo che si possa riconosce quali parole hanno medesimo significato e quali significato opposto: queste sono la sinonimia[9] e l'antonimia (synomymy e antonymy). Le relazioni sinonimiche non sono però da immaginare come nodi collegati da frecce; sarebbe più opportuno immaginare gruppi di parole che hanno lo stesso significato e che prendono il nome di synset (synonym set). Dato ciò, si capisce che se ci troviamo di fronte ad una parola polisemica questa comparirà in synset diversi.

Le relazioni semantiche quindi verranno definite direttamente tra synset. Quelle possibili all'interno di WordNet sono le seguenti:

Iperonimia *(hypernyms):* secondo la quale Y è un iperonimo di X se ogni X è una specie di Y;

ad es. Pasto è un **iperonimo** di breakfast;

[9] In WordNet sono ritenuti sinonime quelle parole che possono essere sostituite senza cambiare la verità della frase.

***Iponimia** (hyponyms):* secondo la quale Y è un iponimo di X se ogni Y è una specie di X;

ad es. Pranzo è un **iponimo** di pasto;

***Olonimia** (holonym):* secondo la quale Y è un olonimo di X se ogni X è una parte di Y;

ad es. Corpo è un **olonimo** di braccio;

***Meronimia** (meronym):* secondo la quale Y è un meronimo di X se ogni Y è una parte di X;

ad es. Braccio è un **meronimo** di corpo;

Possiamo dunque immaginare WordNet come un grande albero, dove, partendo da sensi generali si arriva alle specifiche voci di un dizionario qualunque.

Questi che ho chiamato "sensi generali" sarebbe meglio definirli come classi, o *Unique Beginners*, ovvero delle macroclassi dalle quali prendono il via le gerarchie semantiche.

Per i nomi sono:

1 Tops	8 communication	15 object	22 relation
2 act	9 event	16 person	23 shape
3 animal	10 feeling	17 phenomenon	24 state
4 artifact	11 food	18 plant	25 substance
5 attribute	12 group	19 possession	26 time
6 body	13 location	20 process	
7 cognition	14 motive	21 quantity	

Per i verbi sono:

1 body	6 consumption	11 perception
2 change	7 contact	12 possession
3 cognition	8 creation	13 social
4 communication	9 emotion	14 stative
5 competition	10 motion	15 weather

Ognuna di queste classi specifica a quale grande gruppo appartiene un synset.

Cosa da specificare è che WordNet, a livello strutturale, è organizzato come un comune dizionario, il quale ha proprio la particolarità di contenere non solo nozioni comuni a qualunque altro dizionario, ma anche informazioni semantiche, come appunto l'appartenenza a queste classi.

In concreto, se ci trovassimo a dover analizzare la frase

*"An electric guitar and **bass** player stand off to one side"*

e sia di nostro particolare interesse la parola *bass*, cercando la parola all'interno di WordNet avremmo tale esito:

- <noun.attribute>S: (n) **bass** (the lowest part of the musical range)
- <noun.communication>S: (n) **bass**, bass part (the lowest part in polyphonic music)
- <noun.person>S: (n) **bass**, basso (an adult male singer with the lowest voice)
- <noun.food>S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- <noun.food>S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- <noun.communication>S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- <noun.artifact>S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- <noun.animal>S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes).

Queste racchiuse tra uncinate, sono le categorie alle quali appartengono i synset della parola *"bass"*.

Le classi di cui ho parlato fino ad ora sono quelle che Massimiliano Ciaramita, conducendo ricerche sull'argomento, ha definito *Supersenses* (Ciaramita e Johnson, 2003).

Uno dei lavori di cui si è occupato (Picca, Gliozzo, Ciaramita, 2008), è stata la creazione di un *SuperSense Tagger* per la lingua italiana, ovvero uno strumento che associa in maniera automatica un SuperSenso ad ogni nome, verbo, aggettivo e avverbio all'interno di un dato testo, partendo da un corpus annotato utilizzato per l'addestramento dello stesso.

Attualmente i *Supersenses* di WordNet sono i più utilizzati in ambito di ricerca poiché hanno una struttura consolidata e permettono di effettuare una più

accurata disambiguazione lessicale, oltre al fatto che sono presi come punto di riferimento in quanto sono ritenuti uno standard *de facto*.

4 CORPORA E SUPERSENSI

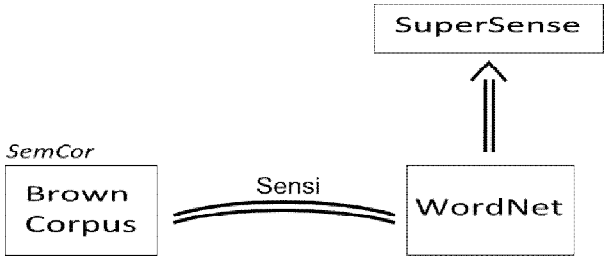
Come accennato, per permettere l'annotazione automatica di un testo con i Supersensi tramite SuperSense Tagger, è necessario avere un corpus già annotato con i SuperSensi corretti, che servirà per l'addestramento del SuperSense Tagger.

I principali corpora annotati con i Supersensi – che descriverò sotto – sono il SemCor Corpus, costituito da una parte del Brown Corpus, il MultiSemCor Corpus, un corpus parallelo tra l'italiano e l'inglese – che è quello utilizzato da Picca e altri per l'addestramento del loro SuperSense Tagger per l'italiano – e l'ISST Corpus, la cui revisione ed estensione è l'oggetto del presente lavoro.

4.1 SemCor Corpus

Il *SemCor Corpus*, creato dall'Università di Princeton, rappresenta un sottoinsieme del Brown Corpus, e contiene circa 700.000 parole dell'inglese contemporaneo. Sebbene stiamo parlando di un Corpus annotato a livello semantico, non tutte le parole al suo interno sono annotate con un senso: soltanto poco più di 200.000 parole sono annotate con i sensi di WordNet.

Per meglio comprendere come sia stato annotato il SemCor Corpus e perché viene citato tra i Corpora annotati con i Supersense, ecco il seguente schema.

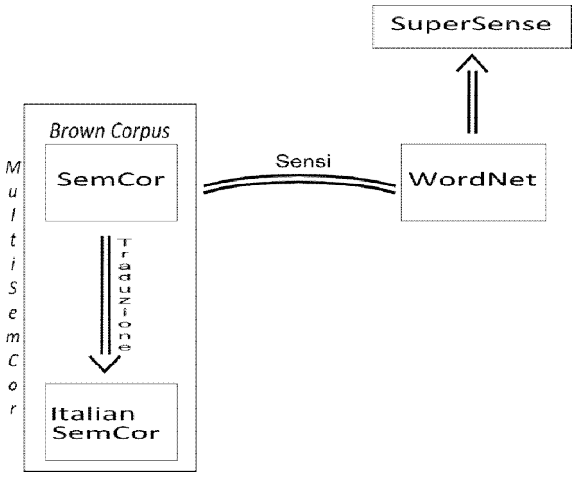


Il SemCor Corpus è stato annotato con i sensi di WordNet. Dai sensi di questo, si può risalire ai Supersensi, in modo tale che ogni parola del SemCor Corpus che possiede un senso, potrebbe risalire anche al proprio Supersenso, sfruttando l'organizzazione gerarchica di WordNet.

4.2 MultiSemCor Corpus

Il *MultiSemCor Corpus*, creato dall'Università di Princeton, è un corpus parallelo delle lingue inglese e italiano – allineato a livello di parola –, creato a partire dal SemCor Corpus.

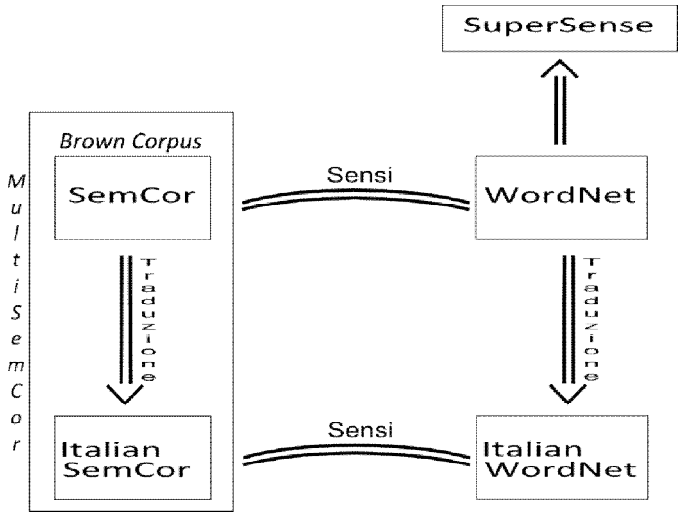
Questo corpus parallelo, realizzato dall'*ITC-irst*[10], può essere visto come una traduzione in italiano di una risorsa inglese. Inizialmente si è partiti traducendo le parole all'interno dei testi di SemCor, poi si è proceduto con l'allineamento automatico dei testi italiani e inglesi a livello di frase e di parola, e successivamente si è fatto ereditare alle parole in italiano il senso presente nell'annotazione delle parole in inglese.



La lettura di questo schema parte dal SemCor Corpus. Di questo è stata fatta una traduzione, con allineamento a livello di parola, in lingua italiana. Il corpus che ne è derivato è l'Italian SemCor. L'insieme dei due corpora, inglese e italiano, è chiamato MultiSemCor. Importante è il fatto che, con l'allineamento a livello di parola è stata fatta ereditare alle parole italiane l'annotazione presente nel SemCor inglese, che puntava ai sensi di WordNet. Dunque, anche l'Italian SemCor è annotato in maniera analoga al SemCor Corpus.

[10] *Irst, istituto per la ricerca scientifica e tecnologica.* Centro di ricerca pubblico della provincia autonoma di Trento.

In un secondo momento, a seguito di studi condotti da Picca[11], all'interno del MultiSemCor Corpus, per la parte italiana, sono stati applicati i sensi di ItalianWordNet, ovvero un database semantico-lessicale – creato anch'esso dall'ITC-irst – avuto dalla traduzione del WordNet di Princeton.



Questo schema, più complesso, mostra il quadro completo. Il WordNet di Princeton è stato tradotto in italiano, come anche il SemCor Corpus. Abbiamo dunque in lingua italiana due risorse originariamente in inglese. Gli studi di Picca hanno voluto che i sensi applicati all'Italian SemCor fossero quelli di ItalianWordNet, non più quelli ereditati dalla traduzione di SemCor. L'ItalianWordNet non possiede però una traduzione dei SuperSensi. Come nel caso inglese, però è possibile associare a ogni senso annotato sul corpus un supersenso utilizzando la struttura gerarchica della risorsa lessicale.

I risultati ottenuti dal SuperSense Tagger in questione addestrato con il MultiSemCor in italiano non sono stati pienamente soddisfacenti: il valore

[11] *Supersense Tagger for Italian.* (2008), Davide Picca, Alfio Massimiliano Gliozzo, Massimiliano Ciaramita.

medio di precisione è circa del 63%, mentre la controparte inglese raggiunge il 77%.

Questo probabilmente è dovuto alla non ottima qualità dei dati di addestramento (Italian SemCor), frutto di una traduzione che non è mai stata revisionata manualmente.

4.3 ISTT

L'*ISST (Italian Syntactic-Semantic Treebank)* è una Treebank (ovvero un corpus annotato a livello sintattico), realizzata tra il 1999 e il 2001 all'interno del progetto di ricerca SI-TAL, che contiene anche un livello di annotazione semantica.

Essa è costituita da una porzione specialistica di ambito economico-finanziario, che comprende articoli estratti da *Il Sole 24 Ore*, e da un corpus bilanciato costituito da articoli estratti da varie riviste, per un totale di 305.547 parole (tokens), così come segue:

Partizione corpus	**Fonte**	**Origine**	**Tokens**
Finanziario	*Il Sole-24 Ore*	Giornata del 25/5/1994	89.941
Bilanciato	*La Repubblica*	Articoli di vario argomento usciti tra il 1985 e il 1988	59.945
		Giornata del 15/7/1995	77.808
Bilanciato	*Il Corriere della Sera*	Giornata del 7/8/1995	57.938

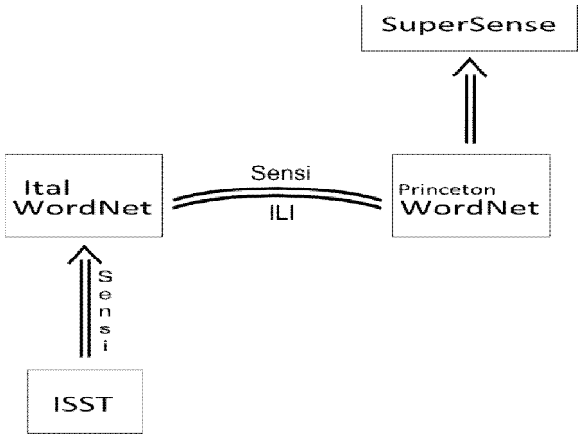
Partizione corpus	Fonte	Origine	Tokens
Bilanciato	*Periodici:* ▪ *Casaviva* ▪ *Centocose* ▪ *Epoca* ▪ *Espansione* ▪ *Grazia* ▪ *Panorama* ▪ *Starbene* ▪ *Storia Illustrata* ▪ *Zerouno*	Selezione di articoli usciti nell'anno 1988	19.915

L'annotazione a livello semantico di questo corpus, secondo le richieste del progetto, doveva essere di 80.000 parole (tokens). Dato che il corpus conteneva circa 142.000 occorrenze – tra sostantivi, verbi e aggettivi – si fece una selezione delle unità da annotare, per mantenersi nei limiti imposti dagli obiettivi progettuali.

Inizialmente il corpus fu annotato con sensi specifici di ItalWordNet.

Successivamente, il dottor Dei Rossi ha lavorato affinchè l'ISST potesse essere annotato con i SuperSensi di WordNet. Dato che ItalWordNet, non possiede i SuperSensi, è stato necessario ricavare i Supersensi da WordNet. Per far ciò, il dottor Dei Rossi ha creato un algoritmo per la navigazione della struttura gerarchica di ItalWordNet (secondo le relazioni di iperonimia) fino al raggiungimento dei synset ai quali era stato associato un ILI, quell'indice che collega alcuni sensi di ItalWordNet con i sensi di WordNet.

Una volta che quindi si è trovata la corrispondenza tra i sensi italiani e quelli inglesi, i SuperSensi di WordNet sono stati applicati alle corrispondenti voci della risorsa italiana.



4.3.1 I Supersense

I Supersense dei quali stiamo parlando sono in totale 45, e prevedono l'assegnazione di un SuperSense per aggettivi, nomi, verbi e avverbi. Questi sono:

File Number	Name	Contents
0	adj.all	all adjective clusters
1	adj.pert	relational adjectives (pertainyms)
2	adv.all	all adverbs
3	noun.Tops	unique beginner for nouns
4	noun.act	nouns denoting acts or actions
5	noun.animal	nouns denoting animals
6	noun.artifact	nouns denoting man-made objects
7	noun.attribute	nouns denoting attributes of people and objects
8	noun.body	nouns denoting body parts
9	noun.cognition	nouns denoting cognitive processes and contents
10	noun.communication	nouns denoting communicative processes and contents
11	noun.event	nouns denoting natural events
12	noun.feeling	nouns denoting feelings and emotions
13	noun.food	nouns denoting foods and drinks
14	noun.group	nouns denoting groupings of people or objects
15	noun.location	nouns denoting spatial position
16	noun.motive	nouns denoting goals
17	noun.object	nouns denoting natural objects (not man-made)
18	noun.person	nouns denoting people
19	noun.phenomenon	nouns denoting natural phenomena
20	noun.plant	nouns denoting plants
21	noun.possession	nouns denoting possession and transfer of possession
22	noun.process	nouns denoting natural processes
23	noun.quantity	nouns denoting quantities and units of measure

24	noun.relation	nouns denoting relations between people or things or ideas
25	noun.shape	nouns denoting two and three dimensional shapes
26	noun.state	nouns denoting stable states of affairs
27	noun.substance	nouns denoting substances
28	noun.time	nouns denoting time and temporal relations
29	verb.body	verbs of grooming, dressing and bodily care
30	verb.change	verbs of size, temperature change, intensifying, etc.
31	verb.cognition	verbs of thinking, judging, analyzing, doubting
32	verb.communication	verbs of telling, asking, ordering, singing
33	verb.competition	verbs of fighting, athletic activities
34	verb.consumption	verbs of eating and drinking
35	verb.contact	verbs of touching, hitting, tying, digging
36	verb.creation	verbs of sewing, baking, painting, performing
37	verb.emotion	verbs of feeling
38	verb.motion	verbs of walking, flying, swimming
39	verb.perception	verbs of seeing, hearing, feeling
40	verb.possession	verbs of buying, selling, owning
41	verb.social	verbs of political and social activities and events
42	verb.stative	verbs of being, having, spatial relations
43	verb.weather	verbs of raining, snowing, thawing, thundering
44	adj.ppl	participial adjectives

Analizziamoli nello specifico.

0 - adj.all
A questa classe appartengono tutti gli aggetivi comuni, quali *grande*, *bello*, *simpatico*.

1 - adj.pert
A questa classe appartengono gli aggettivi che hanno relazioni con una parola, dalla quale in realtà derivano, come *"scolastico"* che deriva da *scuola* o *"marittimo"* che deriva da *mare*.

2 - adv.all
A questa classe appartengono tutti gli avverbi, in quanto altra classe per questi non c'è.

3 - noun.Tops
A questa classe appartengono tutti i nomi che compaiono come classe, ad esempio il nome *"gruppo"* o *"animale"* è un **noun.Tops**, in quanto esistono le classi *noun.group* e *noun.animal*.

4 - noun.act
A questa classe appartengono tutti quei nomi che identificano lo svolgimento di un'azione, come la parola *"corsa"* o *"liberazione"*.

5 - noun.animal
A questa classe appartengono tutti i nomi che identificano gli animali, come *"cane"*, *"gorilla"*, *"coniglio"*.

6 - noun.artifact
A questa classe appartengono tutti i nomi che indicano oggetti creati per mano dell'uomo, come *"edificio"* o *"fontana"* o *"bomba"*.

7 - noun.attribute
A questa classe appartengono tutti i nomi che identificano attribuzioni di qualcosa a individui o oggetti, come *"ricchezza"* o *"pigrizia"*.

8 - noun.body
A questa classe appartengono tutti i nomi che indicano parti del corpo, come *"braccio"* o *"occhio"*.

9 - noun.cognition
A questa classe appartengono tutti quei nomi che indicano processi cognitivi, ovvero che mirano alla conoscenza di qualcosa, come *"pensiero"*.

10 - noun.communication
A questa classe appartengono tutti quei nomi di oggetti che permettono la comunicazione, come *"libro"*, *"radio"* o *"giornale"*.

11 - noun.event
A questa classe appartengono tutti quei nomi che identificano eventi, come *"trionfo"* o *"incidente"*.

12 - noun.feeling
A questa classe appartengono tutti quei nomi che identificano sentimenti ed emozioni, come *"amore"* o *"paura"*.

13 - noun.food
A questa classe appartengono quei nomi che indicano cibo e bevande, come *"miele"* o *"aranciata"*, oppure momenti in cui questi si consumano, come *"cena"* o *"merenda"*.

14 - noun.group
A questa classe appartengono quei nomi che identificano delle associazioni o organizzazioni, dei gruppi o delle comunità, come ad esempio *"chiesa"* e *"ONU"*. A questa classe sono state associate pure le squadre sportive, anche nazionali come *"Italia"* o *"Francia"*.

15 - noun.location
A questa classe appartengono quei nomi che indicano luoghi, come *"Pisa"* o *"Roma"* e come *"via"* o *"piazza"*.

16 - noun.motive
A questa classe appartengono quei nomi che identificano uno scopo, come *"ragione"*, *"causa"* o *"motivo"*.

17 - noun.object
A questa classe appartengono tutti quei nomi che indicano oggetti naturali, non creati dall'uomo, come *"pietra"* o *"mare"*.

18 - noun.person
A questa classe appartengono tutti i nomi propri di persona, come "Mario" o "Giovanni". A questa classe si fanno appartenere anche i cognomi.

19 - noun.phenomenon
A questa classe appartengono i nomi che identificano fenomeni naturali, quali *"nebbia"* o *"fulmine"*.

20 - noun.plant
A questa classe si fanno appartenere tutti i nomi di piante o vegetali, come *"pino"*, *"polline"* o *"basilico"*.

21 - noun.possession
A questa classe appartengono tutti quei nomi che indicano possedimenti, come *"finanziamento"* o *"tassa"*.

22 - noun.process
A questa classe appartengono quei nomi che identificano lo svolgersi di un processo, come *"crescita"* o *"sviluppo"*.

23 - noun.quantity
A questa classe appartengono quei nomi che indicano una quantità e unità di misura, e anche i nomi che indicano valute. Alcuni esempi potrebbero essere numeri o *"metri"* o *"dollari"*.

24 - noun.relation
A questa categoria appartengono quei nomi che identificano relazioni tra persone o cose, come "pezzo" o "resto".

A questa classe appartengono anche i punti cardinali, in quanto ogni punto esiste se in relazione con gli altri.

25 - noun.shape
A questa classe appartengono tutti quei nomi che indicano una qualche forma, come *"colonna"* o *"piano"* o *"curva"*.

26 - noun.state
A questa classe appartengono i nomi che indicano lo stato di un individuo o di una situazione, come *"morte"* o *"crisi"*.

27 - noun.substance
A questa classe appartengono quei nomi che indicano una sostanza, come ad esempio *"oro"* o *"gas"*.

28 - noun.time
A questa classe appartengono quei nomi che identificano il tempo o relazioni temporali, come *"notte"* o *"settembre"* o come *"ore"*.

29 - verb.body
A questa classe appartengono quei verbi che identificano azioni che rigurdano il corpo umano, come *"dormire"* o *"respirare"*.

30 - verb.change
A questa classe appartengono quei verbi che indicano il cambiamento di qualcosa, come ad esempio *"accendere"* o *"chiudere"*.

31 - verb.cognition
A questa classe appartengono quei verbi che identificano processi che coinvolgono la mente e il pensiero, come *"immaginare"* o *"dubitare"*.

32 - verb.communication
A questa classe appartengono tutti quei verbi che abbracciano la sfera della comunicazione, come *"parlare"* o *"cantare"* o anche *"leggere"*.

33 - verb.competition
A questa classe appartengono tutti quei verbi che esprimono avversità o che indicano attività atletiche, come *"espugnare"* o *"sparare"*, o come *"correre"*, se espresso in senso agonistico.

34 - verb.consumption
A questa classe appartengono tutti quei verbi che identificano l'assunzione di cibi, liquidi o sostanze da parte di esseri viventi, come *"mangiare"*, *"sorseggiare"* o *"fumare"*.

35 - verb.contact
A questa classe appartengono tutti quei verbi che indicano un contatto, come *"avvolgere"* o *"sfiorare"*.

36 - verb.creation
A questa classe appartengono tutti quei verbi che indicano la creazione – o distruzione – di qualcosa, o qualunque processo creativo, come anche *"dipingere"*.

37 - verb.emotion
A questa classe appartengono tutti quei verbi che esprimono sentimenti o emozioni, come *"amare"* o *"deludere"*.

38 - verb.motion
A questa classe appartengono tutti quei verbi che indicano lo spostamento di un corpo, come *"camminare"*, *"volare"* o *"muovere"*.

39 - verb.perception
A questa classe appartengono quei verbi che riguardano la percezione, come *"vedere"* o *"sentire"*.

40 - verb.possession
A questa classe appartengono quei verbi che indicano lo scambio di possedimenti, come *"finanziare"*, *"pagare"* o *"investire"*.

41 - verb.social
A questa classe appartengono quei verbi che indicano azioni, attività ed eventi che si svolgono nella società, quali *"presentare"*, *"organizzare"* o *"emarginare"*.

42 - verb.stative
A questa classe appartengono quei verbi che indicano lo stato di essere o avere, e sono proprio questi i verbi che rientrano in questa categoria.

43 - verb.weather
A questa classe appartengono quei verbi che indicano situazioni metereologiche, quali *"piovere"*, *"nevicare"* o *"tuonare"*.

44 - adj.ppl
A questa classe, messa in fondo, appartengono quegli aggettivi detti participiali, in quanto hanno la stessa forma di un participio, ma non derivano da nessun verbo, come ad esempio l'aggettivo *"preoccupante"*.

5 REVISIONE ED ESTENSIONE DEL CORPUS ANNOTATO

Per la buona riuscita del progetto *"SemaWiki"* era necessario avere una buona risorsa semantica di riferimento. Proprio per questo, si è voluto che la maggior parte dei dati all'interno dell'ISST corpus fossero annotati con i SuperSensi adeguati. È proprio all'interno di questo quadro che si colloca il lavoro che ho svolto.

Il mio compito è stato di controllare che i SuperSensi assegnati in maniera automatica a seguito dell'allineamento – grazie all'algoritmo di mapping realizzato dal dottor Dei Rossi – fossero corretti e di annotare manualmente tutte le parole all'interno dell'ISST che non avevano un supersenso assegnato. Data la natura polisemica di alcuni lemmi, è stata necessaria una disambiguazione degli stessi in base al contesto nel quale si trovavano. I lemmi in questione comprendevano sia verbi che nomi.

5.1 Problemi riscontrati

I problemi riscontrati non sono stati pochi, e soprattutto sono stati di varia natura. Alcuni erano dovuti ad errori commessi durante il lavoro precedente al mio arrivo, altri erano dovuti alla risorsa stessa.

5.1.1 PoS Tagging Errato

Il primo problema riscontrato con la risorsa è stato a livello di PoS[12]. A causa dell'individuazione di categorie grammaticali errate i supersensi suggeriti nell'operazione di mapping non sempre erano congruenti con le categorie corrette. Questo è capitato molto spesso con i nomi propri, per i quali erano stati indicati PoS come se nomi propri non fossero. Ad esempio, incontrando la parola *"Cavalli"* o *"Chiesa"* la categoria grammaticale annotata indicava che

[12] PoS, Part of Speech Tagging; è un tipo di annotazione che serve ad identificare le categorie grammaticali di una parola.

fossero entrambi nomi comuni. Dato ciò, l'annotazione semantica suggerita per l'uno, era *noun.animal*, e per l'altro era *noun.artifact*:

Cavalli	Nome, plurale	noun.animal
Chiesa	Nome, singolare	noun.artifact

In determinate situazioni però questo non era sempre corretto. Capitava che questi nomi non fossero nomi comuni, bensì nomi propri, per i quali l'annotazione corretta era *noun.person*.

Cavalli	Nome proprio, singolare	noun.person
Chiesa	Nome proprio, singolare	noun.person

In questi casi la soluzione adottata è stata di annotare con il supersense corretto, anche se creando un certo disallineamento di contenuti tra l'annotazione grammaticale e l'annotazione semantica, in quanto non sono stati corretti i PoS, per i quali, ad esempio, *"Cavalli"* è ancora un *nome plurale* ma semanticamente è un *noun.person*.

5.1.2 SuperSensi, o troppi o pochi

Uno dei problemi più frequenti è stato il seguente. Considerando le categorie semantiche a disposizione – quelle sopra descritte –, non sempre è stato semplicissimo scegliere quella appropriata al contesto. Questo perché o i supersensi che rispondevano adeguatamente a quella parola erano troppi, o perché erano troppo pochi, oppure, alle volte, non si è riuscito a trovare il supersenso che si adattasse perfettamente a una determinata parola.

Simili problemi sono sorti davanti alla parola *"ricerca"*, la quale andrebbe benissimo come <*noun.act*> (che denota un'azione) o <*noun.cognition*> (che denota processi cognitivi); oppure davanti alla parola *"dieta"* (nel suo significato di particolare regime alimentare), per la quale non esiste un supersenso veramente adeguato, ma bisognerebbe un po' forzarne il significato, magari legandolo all'ambito in cui si può trovare il concetto di dieta, ovvero <*noun.food*>, anche se non propriamente corretto.

Alcuni esempi si trovano in queste frasi:

1. Era in stato di ***shock*** *(*noun.state o noun.feeling);

2. Eppure il figlio di ***Dio*** non esitò a mettersi alla sua scuola (noun.person o noun.artifact);

3. […] gridando slogan a favore della Costituzione, per la libertà di stampa e di ***manifestazione*** (noun.act o noun.communication);

4. Ma i drammatici momenti di Hiroshima avevano comunque un che di sacro nel ***film*** trasmesso ieri dalla tv (noun.communication o noun.artifact);

5. […] ma per il periodo che va dall'83 all'86 le ***previsioni*** vogliono tagli del 22% nel nostro paese (noun.communication o noun.cognition);

6. […] è il risultato di una mediazione tra le diverse ***anime*** del Polo (noun.person o noun.attribute);

7. Ora su un tratto di mare del litorale ***sud*** sono comparsi liquami e rifiuti (noun.location o noun.relation);

8. Le ultime elezioni politiche hanno provocato una vera e propria ***rivoluzione*** (noun.act o noun.event).

5.1.3 MultiWord Expressions

Un altro problema riscontrato durante l'annotazione si è presentato nel dover dare una corretta categoria semantica a parole come *"croce"*, ma nella situazione in cui la parola che la segue è *"rossa"*. Si capisce infatti che l'espressione croce rossa ha un significato che non è riconducibile alla semplice somma del significato di croce e di rosso prese singolarmente. Di fatti, singolarmente andrebbero così annotate:

croce <*noun.artifact*>

rossa <*adj.all*>

anzichè

croce_rossa <*noun.group*>

Il problema nasce dal fatto che le parole sono state mantenute a livello di tokenizzazione.

Essendo questo un problema cruciale, si è operato secondo i seguenti criteri: inizialmente si proceduto solo a segnare i casi problematici, procedendo con l'annotazione. In un secondo momento si sono aggregate le varie parti delle MultiWords con la notazione BIO, una notazione convenzionale per segnalare che queste due parole sono legate e sono l'una di seguito all'altra, e di seguito a questa è stato posto il supersenso adeguato, precisamente come segue:

croce <*B-noun.group*>
rossa <*I-noun.group*>

Nel caso in cui la MultiWord Expression comprendesse tre tokens il terzo sarebbe annotato con una O.

5.1.4 Verbi Aspettuali

Un ulteriore problema è sorto nel momento in cui mi sono trovata a dover annotare i cosiddetti *verbi aspettuali*, ovvero quei verbi che specificano in che fase, nello svolgimento di un'azione, ci troviamo; ad esempio, *stare per*, *continuare*, ecc. Trovandosi a dover scegliere una categoria per questi verbi che dipendono totalmente dal contesto, quella più indicata potrebbe risultare verb.stative – come anche suggerisce il WordNet di Princeton –. Ma dato che tra i verbi stativi si trovano il verbo essere o avere, che indicano uno stato effettivo dell'essere, questa categoria non si adatta ad essere assegnata a un verbo come continuare".

Una soluzione potrebbe essere, ad esempio, la creazione di una categoria *ad hoc* per tali verbi. Un'altra possibile soluzione – quella adottata – è di non assegnare a questi verbi nessuna categoria semantica, dato che comunque sono verbi che specificano la fase dello svolgimento di un'azione, non dando ulteriori informazioni semantiche al verbo che segue.

Ad esempio, nelle locuzioni *"continuare a parlare"* o *"stare per uscire"* si capisce che il senso che assumono è rispettivamente di *"parlare"* e *"uscire"*. *"Continuare a"* e *"stare per"* non caratterizzano particolarmente il significato delle locuzioni.

5.1.5 Verbi Supporto

Molto simile al problema delle Multiwords, è l'annotazione dei verbi che singolarmente hanno un significato, ma associati a determinate parole comunicano un concetto totalmente diverso. Questo è il caso di verbi come *"dare"*, il quale, ad esempio, nella sua accezione più comune potrebbe essere ritenuto come un *verb.possession*, in quanto indica il trasferimento di un qualcosa da un individuo ad un altro *(dare qualcosa a qualcuno)*.

Nel caso in cui, però, quel qualcosa risulti essere una mano, sorge qualche perplessità. Il significato della costruzione *"dare una mano"*, ovviamente, non vuol lasciar intendere che dare possa essere un *verb.possession*, bensì risulterebbe più adatto un *verb.social*. Nel caso di *"dare l'allarme"* o *"dare ragione"* la categoria semantica d'appartenenza del verbo dare cambia ancora, e rispettivamente sarebbero *verb.communication* e *verb.cognition.*

Questo, comunque, non è solo il caso del verbo *dare, ma di un numero più ampio di verbi che hanno un comportamento simile, come prendere, mandare, fare, etc.* Per fare qualche esempio: *"prestare attenzione"*, *"correre il rischio"*, *"portare avanti"*, *"mettere in scena"*, *"togliere la vita"*, *"prendere sonno"*, *"mandare in onda"*, *"aprire le danze"*, *"dare fiducia"*, *"tirare le somme"*, *"portare a termine"* e tanti altri.

La soluzione adottata per tali verbi è stata la stessa che per le MultiWord, ovvero è stata utilizzata la notazione BIO davanti al supersense adatto all'espressione.

5.1.6 Verbi Modali

I verbi modali, ovvero quei verbi che generalmente esprimono desiderio, proposito, possibilità, permesso, capacità o necessità, hanno creato pure qualche problema non indifferente. Il motivo è che questi verbi non esprimono un concetto semantico compiuto, ma semplicemente reggono e caratterizzano il verbo che li segue.

Questo è il caso di verbi come *"dovere"*, *"potere"* o *"volere"*, i quali per ovvi motivi non possono propriamente appartenere a nessuna classe semantica tra quelle sopra descritte.

Proprio per questo motivo a questi verbi non è stato assegnato nessun supersenso.

5.1.7 Metafore

Un altro problema l'ho riscontrato dovendo annotare semanticamente delle parole o locuzioni metaforiche. Questo perchè, trovandosi davanti locuzioni quali *"spezzare il cuore"*, *"mangiare la foglia"*, *"tirare le cuoia"*, non è così semplice trovare un modo di operare. Inizialmente si pensava che le occorrenze di metafore fossero poche, dunque poco rilevanti, e per questo si è proceduto annotando singolarmente ogni parte della locuzione, percui *"spezzare"* è stato annotato con *verb.change* e *cuore* come *noun.body*. In un secondo momento, quando ci si è accorti che le occorrenze non erano poi così poche, e che comunque bisognava già gestire il problema delle multiword o dei verbi supporto, si è optato per la stessa soluzione, ovvero la notazione BIO.

5.1.8 Supersense Verbali

Dopo aver parlato dei problemi riscontrati sia con i nomi che con i verbi, vorrei affrontare un problema che, in realtà, sta proprio alla base della struttura stessa del task del SuperSense tagging. Assegnare una categoria semantica ai nomi può rivelarsi relativamente semplice: inizialmente si hanno alcune difficoltà a individuare il senso di alcune parole o l'interpretazione adeguata delle stesse categorie, ma dopo una prima fase di addestramento si riesce ad assegnare ad ogni parola la categoria semantica corretta. Diverso invece è il caso dei verbi.

I verbi sono quella parte variabile del discorso che indicano lo svolgimento di un'azione. Il problema sorge con alcuni verbi che esprimono azioni o eventi di natura generale o complessa. Questo è il caso, per esempio, del verbo *"cercare"*. Se pensiamo un attimo a cosa accade mentre si svolge l'azione *"cercare"*, ci figureremo nella mente un individuo che si muove nello spazio *(<verb.motion>)* alla ricerca di qualcosa, magari chinandosi per terra o allungandosi per vedere meglio se la cosa che cerca si trova in quel posto, e nel mentre fruga *(<verb.contact>)* tra gli oggetti, spostandoli, per meglio vedere tra le cose che potrebbero intralciare la vista *(<verb.perception>)*.

Quante azioni, dunque, si svolgono nel dietro le quinte di *"cercare"*?

Nel dover assegnare una – e una sola – categoria semantica ad un verbo, si dovrebbe semplicemente tener conto di quale sia la più adatta e pregnante tra quelle a disposizione. Ma si capisce che, in situazioni del genere, il compito diventa molto più complicato, per diversi motivi. Uno di questi potrebbe essere il fatto che non c'è una regola che stabilisce quale sia la categoria semantica di alcuni verbi (es. *cercare*), dunque tra le azioni che comprendono determinati verbi non ce n'è nessuna che si adatti meglio delle altre.

Un altro problema rilevante dell'annotazione semantica per i verbi è legato al fatto che la polisemia verbale è tipicamente maggiore rispetto a quella nominale. Ciò vuol dire che i significati dei verbi variano in base al contesto più di quanto lo facciano i significati dei nomi. Si pensi a tutti gli esempi di problemi riscontrati. Se per i problemi inerenti ai nomi si è sempre trovata una soluzione, per i verbi non è stato sempre così facile trovarla, tanto che in alcuni casi si è preferito non assegnare alcun supersenso come è stato fatto, ad esempio, per il verbo *"dare"*, il cui senso è totalmente dipendente dal contesto. Il verbo *"prendere"*, ad esempio, ha più sensi di quelli che si immaginino. Il supersenso principale che gli si potrebbe attribuire è sicuramente di *verb.contact*, in quanto questo verbo è più comunemente utilizzato con il significato di *"afferrare"*. Ma prendere vuol dire pure ricevere, *verb.possession*, se si pensa allo stipendio, o conquistare, *verb.competition*, se si pensa ad una città, o ancora assumere qualcosa, *verb.consumption*, se si pensa ai medicinali, e ancora, se si pensa agli stati dell'animo o ai sentimenti, *verb.emotion*, come alla paura, all'odio, o al coraggio. Sono ben cinque i supersensi possibili per questo verbo.

5.1.9 Prospettive semantiche

Il punto di vista di chi ha il compito di assegnare un Supersenso adatto, è forse ciò che crea maggiori problemi. In alcuni casi, non è il contesto a fare la differenza, ma la sensibilità linguistica. Parole come *"lenticchia"* o *"pomodoro"* – e tanti altri alimenti –, possono essere ritenuti sia *"noun.plant"* che *"noun.food"*. Sta all'annotatore decidere, e operare con coerenza.

Ad esempio, nella precedente annotazione i nomi dei continenti erano stati annotati con il Supersenso *"noun.object"* – che identifica gli oggetti non creati dall'uomo –, e non *"noun.location"*, che costituirebbe comunque una valida alternativa. Anche giornale o libro, andrebbero benissimo entrambi sia come *"noun.communication"*, perchè permettono la comunicazione, sia come *"noun.artifact"*, perchè di fatto sono oggetti creati dall'uomo.

Sono tantissimi gli esempi che si potrebbero fare, perchè ogni parola può avere classificazioni semantiche diverse a seconda della prospettiva che si assume per caratterizzarne il significato.

5.2 Miglioramenti

Al termine del lavoro che ho svolto i miglioramenti sono stati notevoli, sia dal punto di vista della copertura, sia dal punto di vista del miglioramento che ha avuto il SuperSense Tagger che si addestra sui dati di questo corpus. Inizialmente la situazione era quella mostrata nella seguente tabella

	Con Supersense	Ambigui	Senza Supersense
Noun	43908	3492	38266
Verb	10088	1351	29260
Adj	3219	118	16492
Adv	0	0	13418
TOT	**57215**	**4961**	**97436**

Successivamente al lavoro svolto i risultati sono stati i seguenti

	Con Supersense	Ambigui	Senza Supersense
Noun	68876	0	12688
Verb	27635	194	1550
Adj	17465	0	4787
Adv	9366	0	0
TOT	**123342**	**194**	**19025**

Si può notare come la situazione sia notevolmente migliorata, soprattutto per quanto riguarda l'assegnazione di un SuperSenso alle parole che ne erano

sprovviste. Si è passati da 57.215 parole annotate con SuperSenso corretto a 123.342 parole annotate.

Nella colonna *"Ambigui"*, compaiono quelle parole alle quali inizialmente non era assegnato alcun supersenso, ed erano annotate con *"Ambiguo"* proprio perchè il supersenso da assegnare doveva essere definito in base al contesto in cui compariva quella data parola. Questi attualmente sono solamente 194 occorrenze dello stesso lemma, ovvero *"dare"*, che compare in certi casi di dipendenza. Per i token lasciati invece senza Supersenso – che sono passati da 97.436 a 19.025 –, c'è la possibilità di annotarli in via semiautomatica.

Questo per quanto riguarda il miglioramento a livello di copertura della Treebank.

Per quanto riguarda, invece, la precisione del SuperSense Tagger che si basa sui dati di questo corpus, siamo passati da

```
accuracy: 86.89%;
precision: 80.65%;
recall: 72.82%;
F: 76.53
```

a

```
accuracy: 89.46%;
precision: 79.92%;
recall: 78.30%;
F:  79.10
```

Per notare l'effettivo miglioramento del SuperSense Tagger, a seguito di un addestramento con il Corpus revisionato ed esteso, si può fare riferimento a *F-measure* – la media armonica di precision e recall– il quale è migliorato di più del 2%.

Per quanto questo miglioramento possa sembrare minimo, non ci si deve lasciar ingannare, perchè alzare la media ponderata dell'accuratezza di una risorsa presa per intero, magari ai primi passi è facile, ma lavorando con grosse quantità di dati anche il 2% è comunque un incremento significativo.

Inoltre, un vantaggio apportato dal lavoro che ho svolto è che, se precedentemente si allenava il SuperSense Tagger su dati non corretti e si procedeva con test su dati ugualmente non corretti, adesso ci si può vantare di avere una risorsa ben, e quasi completamente, annotata.

6 CONCLUSIONE

Abbiamo visto che esistono diversi metodi e approcci per la disambiguazione semantica. Uno è la *Word Sense Disambiguation*, che può utilizzare una risorsa di apprendimento creata manualmente, un dizionario, oppure può basarsi su un algoritmo che apprende e incrementa la propria efficacia ad ogni iterazione. Un altro metodo è la *Named Entity Recognition*, la quale agisce solo sui nomi propri, e utilizza o corpora di addestramento sui quali viene calcolata la percentuale di probabilità che determinate situazioni si ripetano, o su regole per le quali viene riconosciuta una determinata entità nominale. L'ultimo metodo che abbiamo visto per la disambiguazione semantica è la SuperSense Disambiguation, la quale necessita di un corpus interamente annotato con i supersensi – estratti da una risorsa di riferimento, come WordNet –, e di un SuperSense Tagger, ovvero un algoritmo che apprende la distribuzione dei diversi supersensi a seconda del contesto. Approfondendo questo ultimo tipo di disambiguazione semantica, abbiamo visto quali sono i più importanti corpora annotati con i supersensi – col fine di essere utilizzati come risorse di apprendimento –, ovvero il SemCor Corpus, costituito da una piccola parte del Brown Corpus e annotato con i sensi di WordNet, il MultiSemCor Corpus, costituito da due corpora paralleli, uno in lingua inglese, ovvero il SemCor Corpus, e uno in lingua italiana, esito della traduzione del SemCor Corpus. L'ultimo corpus che è stato illustrato è l'ISST, al quale ho lavorato apportando correzioni, ampliandolo dal punto di vista dell'annotazione semantica con i supersensi. Facendo questo lavoro, ho avuto diversi problemi e difficoltà, come il fatto che per alcuni nomi il PoS era errato e ha comportato problemi durante il mapping per il suggerimento dei supersensi, il fatto che alcune parole non si rispecchiano in nessun supersenso tra quelli possibili, il problema delle MultiWord Expression *(croce rossa)* e dei verbi supporto, il problema dato dai verbi aspettuali che indicano una fase dell'azione che si svolge, i verbi modali (dovere, potere, volere), e infine si è parlato in generale delle difficoltà di annotazione riscontrate per i supersensi verbali. Infine, si è mostrato di quanto la risorsa sia migliorata, in numero di token annotati, e in precisione.

Per conlcudere, annotare un corpus con i supersensi si è rivelato più complesso di quanto poteva sembrare inizialmente. I problemi riscontrati sono stati molteplici e di diversa natura. La teoria ci insegna che annotare con i supersensi è più semplice in quanto questi, in confronto ai sensi specifici, semplificano il lavoro di riconoscimento delle categorie e dell'annotazione. La pratica ci

insegna che annotare un corpus con i supersensi presenta comunque molte difficoltà, soprattutto per quanto riguarda l'annotazione verbale.

«Alice era così impacciata che non disse nulla, e dopo un minuto Humpty Dumpty ricominciò:

Alcune di esse sono intrattabili... specialmente i verbi sono orgogliosissimi... con gli aggettivi si può fare ciò che si vuole, ma non con i verbi... Però io so maneggiarle tutte quante.»

7 BIBLIOGRAFIA

Ping Chen, Wei Ding, Chris Bowes, David Brown. *A Fully UnsupervisedWord Sense Disambiguation Method UsingDependency Knowledge.* 2009.

Massimiliano Ciaramita, Yasemin Altun. *Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger.* 2006.

Massimiliano Ciaramita, Yasemin Altun. *Named-Entity Recognition in Novel Domains with External Lexical Knowledge.* 2005.

Massimiliano Ciaramita, Mark Johnson. *Supersense Tagging of Unknown Nouns in WordNet.* 2003.

Cristian D'Andrea, Massimo Ruffolo. *XONTO-LING: uno Strumento per l'Acquisizione di Proprietà Linguistiche da Documenti Testuali.* 2009.

Marco Ernandes. *Information Extraction*. 2005.

Christiane Fellbaum. *WordNet, an electronic lexical database.* 1998.

Julio Gonzalo, Irina Chugur and FeHsa Verdejo. *Sense clusters for Information Retrieval: Evidence from Semcor and the EuroWordNet InterLingual Index.* 2002.

Nancy Ide, Jean Véronis. *Word Sense Disambiguation: The State of the Art.* 1999.

Daniel Jurafsky, James H. Martin. *Speech and language processing.*

Davide Picca, Adrian Popescu. *Using wikipedia and supersense tagging for semi-automatic complex taxonomy construction.* 2007.

Davide Picca, Alfio Massimiliano Gliozzo, Massimiliano Ciaramita. *Semantic Domains and Supersense Tagging for Domain-Specific Ontology Learning.* 2007.

Davide Picca, Alfio Massimiliano Gliozzo, Massimiliano Ciaramita. *Supersense Tagger for Italian.* 2008.

Hinrich Shütze. *Automatic Word Sense Discrimination.* 1998.

Antonio Toral, Rafael Muñoz. *A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia.* 2006.

David Yarowsky. *Unsupervised Word Sense Disambiguation Rivaling Supervised Method.* 2002.

GuoDong Zhou, Jian Su. *Named Entity Recognition using an HMM-based Chunk Tagger.* 2002.

8 SITOGRAFIA

http://catalog.elra.info
Sito dell'Associazione Europea delle Risorse del Linguaggio.

http://ilc.cnr.it
Sito dell'Istituto di Linguistica Computazionale di Pisa "Antonio Zampolli"

http://itc.it/irst
Sito dell'Istituto di Ricerca Scientifica e Tecnologica della provincia autonoma di Trento.

http://wordnet.princeton.edu
Sito dal quale è consultabile WordNet.

9 RINGRAZIAMENTI

In questo piccolo spazio vorrei ringraziare le persone che mi sono state veramente vicino durante il lavoro svolto e la stesura di queste poche pagine.

Il primo ringraziamento va di certo a Stefano, il dottor Dei Rossi, senza il quale la qualità di questa tesi non sarebbe mai potuta essere la stessa. Lo ringrazio per la sua totale disponibilità in ogni momento.

Un secondo ringraziamento va a tutti i miei colleghi ed amici, i quali hanno sempre saputo sostenermi in ogni attimo di questo, a volte difficile, cammino. Senza di loro niente avrebbe avuto lo stesso valore.

E, dulcis in fundo, un ringraziamento di cuore a chi ha sopportato i miei sfoghi e le mie ansie, a chi non si è mai tirato indietro ogni volta che ne ho avuto bisogno, a chi mi è stato vicino sin dall'inizio fino alla fine. Grazie Fra.