Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecnica di disambiguazione di termini mediante l'uso di dizionari bilingua

Davide Alagna, Giovanni Pilato, Filippo Sorbello, Giorgio Vassallo

RT-ICAR-PA-03-16 **dicembre 2003**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: www.icar.cnr.it
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: www.na.icar.cnr.it
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: www.pa.icar.cnr.it

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecnica di disambiguazione di termini mediante l'uso di dizionari bilingua

Davide Alagna[2], Giovanni Pilato[1],
Filippo Sorbello[2], Giorgio Vassallo[2]

Rapporto Tecnico N.16: **RT-ICAR-PA-03-16**

Data: **dicembre 2003**

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sezione di Palermo Viale delle Scienze edificio 11 90128 Palermo

[2] Università degli Studi di Palermo Dipartimento di Ingegneria Informatica Viale delle Scienze 90128 Palermo

Indice

Introduzione

Nell'ambito dell'Intelligenza Artificiale (IA), la disambiguazione dei sensi di una parola (Word Sense Disambiguation, WSD) in maniera automatica è stata da sempre oggetto di interesse. E' infatti dal 1950 che si incominciano ad utilizzare i primi strumenti di trattamento computerizzato del linguaggio. La scelta del senso più appropriato per una parola ambigua in una frase non è da considerarsi come fine a se stessa ma, piuttosto, come "compito intermedio" (Wilks and Stevenson [1]), necessario per portare a termine altri compiti per il trattamento del linguaggio naturale.

Essa è essenziale per applicazioni di comprensione linguistica quali, ad esempio, comprensione di messaggi o comunicazione uomo-macchina, ed è utile in alcuni casi per applicazioni il cui scopo non è di comprensione linguistica, come nel caso di traduzione automatica (dall'inglese "Machine Translation", MT), recupero automatico di informazioni (dall'inglese "Information Retrival", IR), analisi tematica e di contenuto, analisi grammaticale, elaborazione del parlato ed elaborazione del testo.

Negli ultimi anni le soluzioni di disambiguazione automatica si sono moltiplicate grazie alla disponibilità di una grande quantità di testi in formato elettronico e al corrispondente sviluppo di metodi statistici per identificare e utilizzare le informazioni inerenti alle regolarità dei dati.

Attualmente sono stati identificati altri problemi riconducibili a questi metodi, quali ad esempio la disambiguazione di parti del discorso e l'allineamento di traduzioni parallele, e il problema della WSD ha assunto un ruolo centrale, citato frequentemente come uno dei più importanti problemi nella ricerca del trattamento del linguaggio naturale.

Il presente progetto si colloca nell'ambito del Web Semantico – il cui obiettivo è quello di dare una struttura alla moltitudine di risorse e documenti presenti in rete per facilitarne il reperimento e, inoltre, di agevolare la condivisione di informazioni – per il quale, senza la WSD, sarebbe aleatorio trovare una giusta collocazione di un termine nel proprio contesto di appartenenza.

Obiettivo del presente lavoro è la definizione di una tecnica per la disambiguazione automatica di termini lessicali componenti una query, allo scopo di estendere la ricerca in Internet. Dal momento che, infatti, la maggior parte dei documenti reperibili in Internet è presente in lingua inglese, si è realizzato uno strumento di supporto che, sulla base di una query (interrogazione) in lingua italiana, ne determini l'esatto senso dei termini che la compongono e la estenda, generando una serie di query, semanticamente inerenti con la prima, in lingua inglese.

Per realizzare ciò si è fatto uso di una base di dati lessicale, WordNet, di un motore di ricerca, Google, e di un sito di traduzioni in linea [2]. Si è proceduto pertanto alla creazione di una lista di sostantivi di base italiano-inglese, in cui ricercare i termini presenti nella query. Successivamente si è effettuata una prima disambiguazione dei termini presenti nella query mediante interrogazione a Google, scegliendo tra le varie

traduzioni possibili di ogni parola quella più opportuna. Infine è stato sfruttato WordNet per ottenere informazioni specifiche sui sensi di ogni singolo termine della query e poterlo classificare in base al senso corretto mediante ulteriore interrogazione a Google.

Il lavoro descritto nel presente elaborato risulta organizzato come segue:

- Nel **Capitolo 1** verranno descritti i principali lavori esistenti nell'ambito della WSD, classificati in base alle risorse utilizzate;
- Nel **Capitolo 2** verranno descritti gli strumenti adoperati in questo studio;
- Nel **Capitolo 3** verrà descritta la soluzione proposta;
- Nel **Capitolo 4** verranno esaminati i risultati sperimentali ottenuti;
- Nel **Capitolo 5** verranno illustrate le conclusioni e gli eventuali sviluppi successivi.

Capitolo 1
Tecniche di disambiguazione

In generale la disambiguazione dei sensi di una parola (Word Sense Disambiguation, *WSD*) implica l'associazione di una data parola in un testo o discorso con una definizione o significato (senso) che si distingue da altri significati che si possono attribuire potenzialmente a quella parola. Il compito prevede essenzialmente due passi:

1. la determinazione di tutti i diversi sensi per ogni parola rilevante per il testo o discorso considerato;

2. un modo per assegnare ogni occorrenza di una parola al senso appropriato.

Per il primo passo molti dei lavori più recenti sulla WSD si basano su sensi pre-definiti, che comprendono:

- una lista dei sensi come quelli che si trovano nei comuni dizionari;

- un gruppo di caratteristiche, categorie o parole associate (ad es. sinonimi, come in una raccolta di vocaboli);

- una voce in un dizionario che include traduzioni in un'altra lingua.

La definizione esatta di un senso è, comunque, un argomento di notevole dibattito. La varietà di approcci per definire i sensi ha destato recente interesse riguardo la comparabilità di molto del lavoro di WSD e, vista la difficoltà del problema della definizione dei sensi, nessuna soluzione definitiva si pensa potrà essere trovata a breve termine [3] [4].

Per il secondo passo, l'assegnazione di parole ai sensi viene compiuta affidandosi a due maggiori fonti di informazione:

- il **contesto** della parola da disambiguare, in senso ampio: esso include le informazioni contenute nel testo o discorso in cui appare la parola, insieme con le informazioni extra linguistiche sul testo quali la situazione, ecc.;

- le **fonti di conoscenza esterne**, comprese le risorse lessicali, enciclopediche, come anche le fonti di conoscenza create a mano, che forniscono dati utili per associare le parole con i sensi.

Tutto il lavoro di disambiguazione comporta l'associare il contesto della parola da disambiguare o con l'informazione di una fonte di conoscenza esterna (knowledge-driven WSD, disambiguazione guidata dalla conoscenza) o con l'informazione proveniente dai contesti di esempi precedentemente disambiguati della parola, derivanti da corpora (data-driven o corpus-based WSD, disambiguazione guidata dai dati o basata su corpora). Qualsiasi varietà di metodi associativi viene utilizzata per determinare la

migliore corrispondenza tra il contesto in esame e una di queste fonti di informazione, allo scopo di assegnare un senso a ciascuna occorrenza di una parola.

Nel presente capitolo vengono descritti alcuni dei metodi sviluppati e i lavori più rilevanti per ciascuno di essi.

1.1 Introduzione

I primi metodi per la disambiguazione automatica dei sensi furono realizzati nell'ambito della traduzione automatica (MT). Inizialmente lo scopo della MT era poco rilevante poiché essa si occupava essenzialmente di traduzioni di testi tecnici e riguardanti sempre particolari domini. Weaver nel suo *Memorandum* [5] tratta il ruolo del dominio nella disambiguazione dei sensi sostenendo un'idea che sarà ripresa parecchi decenni più tardi, e cioè che "*in matematica, per prendere quello che probabilmente è l'esempio più facile, entro un contesto generale di un articolo matematico, si può dire quasi certamente che ogni parola ha uno ed un solo significato.*"

Partendo da tale affermazione, molti sforzi sono stati fatti nei primi tempi della MT per sviluppare dizionari specialistici o "micro-glossari" contenenti esclusivamente il significato di una data parola rilevante per testi di un particolare dominio del discorso. Molti ricercatori hanno tentato di ideare una "interlingua", basata su principi matematici e logici, che potesse risolvere il problema della disambiguazione facendo corrispondere le parole in un qualsiasi linguaggio a una rappresentazione semantico-concettuale. Tra queste soluzioni, quello di Masterman [6] ha portato alla nozione di "rete semantica".

La prima base di conoscenza implementata automaticamente è stata costruita dal *Roget's Thesaurus*.

1.2 Metodi basati sull'Intelligenza Artificiale

I metodi basati sull'Intelligenza Artificiale cominciano a fiorire nei primi anni sessanta e cominciano ad affrontare il problema della comprensione linguistica. Come risultato la WSD, nel campo dell'Intelligenza Artificiale, è stata applicata nel contesto di più ampi sistemi designati per la piena comprensione del linguaggio. Tali sistemi sono fondati sulla teoria della comprensione della lingua umana, che hanno tentato di modellare e, spesso, hanno implicato l'uso della conoscenza dettagliata della sintassi e della semantica per eseguire il loro compito, poi sfruttato per la WSD.

1.2.1 Metodi simbolici

Come precedentemente accennato, le reti semantiche furono sviluppate alla fine degli anni cinquanta e furono subito applicate al problema della rappresentazione dei significati di una parola.

Masterman [6], lavorando nell'area della traduzione automatica, utilizza una rete semantica per ottenere la rappresentazione di frasi in una interlingua formata da concetti linguistici fondamentali; le distinzioni di senso sono rese implicitamente scegliendo rappresentazioni che riflettono gruppi di nodi fortemente connessi nella rete.

Basandosi su ciò Quillian [7] ha costruito una rete che include collegamenti fra parole (tokens) e concetti (types), in cui i collegamenti sono etichettati con varie relazioni semantiche o indicano semplicemente associazioni tra parole. Tale rete è creata a partire da definizioni prese da un dizionario, ma viene arricchita dalla conoscenza umana con criteri non automatici. Quando due parole sono presenti nella rete, il programma di Quillian simula l'attivazione graduale di "nodi concetto" lungo un percorso di collegamenti originati da ogni parola in ingresso per mezzo di un "evidenziatore di percorso" (marker passing); la disambiguazione viene compiuta poiché è probabile che solo un nodo associato con una data parola in ingresso sia coinvolto nel cammino più diretto tra le due parole in ingresso.

Ulteriori approcci basati sull'IA sfruttano l'uso di modelli (frame) contenenti informazioni riguardo le parole e i loro ruoli e relazioni ad altre parole in singole frasi. Altri, basati sul caso, si fondano sulla semantica preferenziale per la comprensione del linguaggio.

Boguraev [8] proseguendo in tale ambito, mostra che la semantica preferenziale è inadeguata a trattare i verbi polisemici, e migliora i metodi precedenti integrando la disambiguazione semantica con la disambiguazione strutturale. Come altri sistemi di questo periodo, questi sistemi sono basati sulla frase (sentence-based) e non tengono in considerazione i fenomeni che si manifestano ad altri livelli del discorso, quali l'informazione sull'argomento e sul dominio. Il risultato è che alcuni tipi di disambiguazione sono difficili o impossibili ad ottenersi.

Un approccio alquanto differente per la comprensione linguistica, che contiene una componente sostanziale di discriminazione dei sensi, è l'"*Analizzatore Esperto di Parole*" (dall'inglese "Word Expert Parser") (Adriaens [9]). Questo approccio deriva dalla teoria non convenzionale che la conoscenza umana sulla lingua è organizzata principalmente sulla conoscenza delle parole piuttosto che delle regole. Il sistema modella ciò che esso intuisce possa essere il processo di comprensione del linguaggio umano: un coordinamento di scambio di informazioni fra "esperti di parole" (word experts) sulla sintassi e la semantica, dal momento che ciascuno determina il proprio coinvolgimento nell'ambito in questione. Il maggiore inconveniente di questo sistema è che, dovendo ogni "esperto di parole" contenere una "rete di discriminazione" (discrimination net) per tutti i sensi delle parole, ciascun "esperto di parole" deve essere estremamente ampio e complesso, il che implica uno sforzo più grande della disambiguazione dei sensi.

1.2.2 Metodi di connessione

I modelli a diffusione di attivazione – in letteratura noti come "spreading activation models" – in cui i concetti in una rete semantica sono attivati dall'uso e l'attivazione si diffonde ai nodi connessi, derivano dai lavori psicolinguistici degli anni sessanta e settanta, che misero in evidenza l'importanza del "semantic priming" – un processo in cui l'introduzione di un certo concetto influenza e facilita l'elaborazione di altri concetti introdotti in seguito e correlati semanticamente – nella disambiguazione compiuta dall'uomo.

Similmente a quanto si verifica per gli esseri umani, l'attivazione dei nodi è indebolita mentre si diffonde, ma certi nodi possono ricevere attivazione da parecchie sorgenti e possono essere progressivamente rinforzati.

Le difficoltà legate alla creazione manuale delle fonti di conoscenza richieste per i sistemi basati sull'Intelligenza Artificiale hanno ristretto il campo di indagine e, di conseguenza, le procedure di disambiguazione utilizzate in questi sistemi sono per la maggior parte utilizzate su insiemi molto piccoli e con testi limitati, rendendo impossibile determinare la loro efficacia su testi veri e propri.

1.3 Metodi basati sulla conoscenza

I metodi basati sull'Intelligenza Artificiale degli anni settanta e ottanta, interessanti da un punto di vista teorico, non lo erano da un punto di vista pratico per la comprensione linguistica, poiché riguardavano soltanto dei domini molto limitati.

Il lavoro sulla WSD raggiunge un punto di svolta negli anni ottanta quando risorse lessicali su larga scala, quali dizionari, thesauri e corpora, divengono ampiamente disponibili. Si è tentato di estrarre automaticamente la conoscenza da queste fonti e, più recentemente, di costruire ampie basi di conoscenza in maniera non automatica.

1.3.1 Dizionari in formato elettronico

I dizionari in formato elettronico ("Machine-Readable Dictionaries", MRD) sono diventati una fonte di conoscenza molto usata per compiti di trattamento del linguaggio naturale durante gli anni ottanta. Un importante campo di attività riguardava tentativi di estrarre automaticamente basi di conoscenza lessicale e semantica dai dizionari in formato elettronico. Questi lavori contribuirono in modo significativo agli studi sulla semantica lessicale, ma lo scopo iniziale, cioè l'estrazione automatica di

ampie basi di conoscenza, non fu completamente raggiunto. L'unica base di conoscenza lessicale ampiamente disponibile al momento (WordNet[1]), fu creata manualmente.

Le difficoltà di estrarre semplici relazioni come l'iperonimia sono dovute in parte all'incongruenza nei dizionari, così come al fatto che i dizionari sono creati per uso umano e non per lo sfruttamento elettronico. Malgrado i difetti, i dizionari in formato elettronico forniscono una fonte di informazione immediatamente disponibile sui sensi di una parola, e perciò divennero uno strumento cardine della WSD.

Lesk [10] crea una base di conoscenza che associa a ciascun senso in un dizionario una "etichetta" composta dalla lista di parole che appaiono nella definizione di quel senso. La disambiguazione viene compiuta selezionando il senso della parola in esame la cui etichetta contiene il maggior numero di sovrapposizioni con le etichette delle parole vicine nel suo contesto. Il metodo raggiunge una disambiguazione corretta con una percentuale del 50-70%, utilizzando un insieme di distinzioni di senso abbastanza fine come quelli che si trovano in un tipico dizionario. Tale metodo è molto sensibile all'esatta espressione di ogni definizione: la presenza o l'assenza di una data parola può radicalmente alterare i risultati; comunque il metodo di Lesk è servito da base per la maggior parte dei lavori successivi di disambiguazione basati sui MRD.

Wilks *et al.* [11] hanno cercato di migliorare la conoscenza associata ad ogni senso calcolando la frequenza delle co-occorrenze delle parole in "testi di definizione", da cui derivarono molte misure del grado di relazione tra parole. Questa metrica è quindi utilizzata con l'aiuto di un metodo vettoriale che relaziona ogni parola e il suo contesto. Negli esperimenti su una singola parola (bank), il metodo ha raggiunto una precisione del 45% nell'identificazione del senso, e una precisione del 90% nell'identificazione delle parole omografe. Il metodo di Lesk è stato esteso creando una rete neurale dai testi di definizione del *Collins English Dictionary* (*CED*), in cui ogni parola è collegata ai propri sensi, a loro volta collegati alle parole nelle loro definizioni, che sono a loro volta collegate ai loro sensi, etc.

Parecchi autori hanno tentato di migliorare i risultati usando campi supplementari di informazione nella versione elettronica del *Longman Dictionary of Contemporary English* (*LDCOE*), in particolare i "box codes" e i "subject codes" forniti per ogni senso. I "box codes" contengono primitive come ASTRATTO, ANIMATO, UMANO, etc. e codificano restrizioni di tipo su nomi e aggettivi. I "subject codes" usano un'altra serie di primitive per classificare i sensi delle parole per argomento (ECONOMICO, INGEGNERISTICO, etc.).

Le incongruenze nei dizionari, precedentemente accennate, non sono l'unica e forse neanche la più importante fonte della loro limitazione per la WSD. Mentre i dizionari forniscono un'informazione dettagliata a livello lessicale, essi non presentano l'informazione pragmatica che serve per la determinazione del senso. Non è quindi

[1] La descrizione di WordNet è trattata nel Capitolo 2.1.

sorprendente che i corpora siano diventati fonte primaria di informazione (si veda in proposito il paragrafo 1.4).

1.3.2 Raccolte di vocaboli

Le raccolte di vocaboli, meglio note come thesauri, forniscono informazioni sulle relazioni tra parole, principalmente sinonimia. Il *Roget's International Thesaurus*, che fu tradotto in formato elettronico negli anni cinquanta ed è stato usato in una varietà di applicazioni che includono traduzione automatica, recupero di informazioni e analisi dei contenuti, fornisce anche una esplicita gerarchia concettuale costituita da otto livelli progressivamente raffinati. In genere, ogni occorrenza di una stessa parola sotto differenti categorie del thesaurus rappresenta sensi differenti di quella parola; cioè, le categorie corrispondono grossolanamente ai sensi delle parole. Un insieme di parole nella stessa categoria è semanticamente collegato.

Tra i vari progetti sviluppati, maggior interesse ha destato quello di Yarowsky [12]. Egli deriva classi di parole cominciando con le parole presenti nelle categorie comuni del *Roget's* (IV ed.). Per ciascuna parola di una categoria viene estratto un contesto di 100 parole da un corpus, e viene utilizzata una statistica basata sull'informazione mutua per identificare le parole che più probabilmente co-occorrono con i membri della categoria. Le classi risultanti sono usate per disambiguare nuove occorrenze di una parola polisemica: il contesto di 100 parole della occorrenza polisemica è esaminato per le parole nelle varie classi, e viene applicata la regola di Bayes per determinare la classe che più probabilmente risulta essere quella della parola polisemica. Poiché la classe rappresenta per Yarowsky un particolare senso di una parola, l'assegnazione a una classe identifica il senso. Egli ottiene un'accuratezza pari al 92% su una triplice distinzione di senso e nota che questo metodo è efficace per estrarre informazioni su un argomento, che, a sua volta, è di grande aiuto per disambiguare sostantivi.

Come per i dizionari in formato elettronico, un thesaurus è una risorsa creata per gli esseri umani, e pertanto non è una fonte di informazione perfetta sulle relazioni tra parole. I livelli più alti della sua gerarchia concettuale sono talmente ampi da essere di scarsa utilità per stabilire categorie semantiche significative. Ciò nonostante, i thesauri forniscono una ricca rete di associazioni di parole e un insieme di categorie semantiche potenzialmente importanti per degli studi linguistici; tuttavia, sia il *Roget's* che gli altri thesauri non sono stati ampiamente utilizzati per la WSD.

1.3.3 Lessici computazionali

A metà degli anni ottanta vennero fatti molti sforzi per cominciare a costruire manualmente basi di conoscenza su larga scala (ad esempio WordNet, Cyc, ACQUILEX). Esistono due approcci fondamentali per la costruzione di lessici semantici:

- **enumerativo**, in cui i sensi sono forniti esplicitamente;
- **generativo**, in cui l'informazione semantica associata a parole date è sottospecificata e regole generative vengono usate per derivare precise informazioni di sensi.

1.3.3.1 Lessici enumerativi

Tra i lessici enumerativi, WordNet è, al momento, la risorsa più utilizzata per la disambiguazione dei sensi in inglese. Sono in via di sviluppo versioni di WordNet per parecchie lingue europee.

WordNet combina le caratteristiche di molte altre risorse comunemente sfruttate nel campo della disambiguazione: esso comprende definizioni per sensi individuali di parole al suo interno come in un dizionario; definisce insiemi di sinonimi di parole – *synset*, synonym sets – che rappresentano un unico concetto lessicale e li organizza in una gerarchia concettuale come un thesaurus, e include altre relazioni tra parole secondo varie relazioni semantiche[2]. In tal modo esso costituisce il più ampio insieme di informazioni lessicali in un'unica risorsa.

Alcune delle prime soluzioni nelle quali si è sfruttato WordNet per la disambiguazione di sensi sono nel campo del recupero di informazioni.

Resnik [13] studia una misura della similarità semantica tra parole nella gerarchia di WordNet. Egli calcola il contenuto informativo condiviso delle parole, che è una misura della specificità del concetto che classifica le parole nella gerarchia IS-A di WordNet. Quanto più specifico è il concetto che classifica due o più parole, tanto più si suppone esse siano semanticamente correlate. Resnik contrappone il suo metodo di calcolo delle similarità a quelli che calcolano la lunghezza di percorso, sostenendo che i collegamenti nella tassonomia di WordNet non rappresentano distanze uniformi. Il metodo di Resnik, applicato usando le distinzioni di senso a grana fine di WordNet e confrontato con i parametri di esperti umani, raggiunge una precisione a livello umano. Questo lavoro considera soltanto sostantivi.

WordNet non è una risorsa perfetta per la WSD. Il problema più frequentemente citato è la grana fine delle distinzioni di senso di WordNet, che sono spesso ben al di là di ciò che può essere necessario in molte applicazioni di trattamento del linguaggio.

[2] Le relazioni tra parole sono descritte nel Capitolo 2.1.2.

1.3.3.2 Lessici generativi

La maggior parte dei lavori sulla WSD è basata sulle distinzioni enumerative dei sensi così come trovate nei dizionari. Secondo Buitelaar [14], la disambiguazione di sensi in un contesto generativo inizia con una etichettatura semantica che mira ad una complessa rappresentazione della conoscenza, la quale riflette tutti i sensi sistematicamente correlati di una parola. Dopodiché il calcolo semantico può generare una interpretazione dipendente dal discorso che contiene più precise informazioni di senso sull'occorrenza.

Successivamente Viegas *et al*. [15] descrivono un approccio similare alla WSD. Essi accedono ad un grande lessico sintattico e semantico che fornisce informazioni dettagliate sulle restrizioni selettive, e ricercano una ontologia ampiamente connessa per determinare quali sensi della parola in esame meglio soddisfino queste restrizioni, riportando un indice di successo del 97%.

1.4 Metodi basati su corpus documentali

Un corpus fornisce una raccolta di esempi che permettono lo sviluppo di modelli linguistici numerici. Sin dalla fine del diciannovesimo secolo l'analisi manuale di corpora ha permesso lo studio di parole e di grafemi e l'estrazione di liste di parole e collocazioni per lo studio dell'acquisizione o dell'insegnamento delle lingue.

Nell'area della WSD, Black [16] sviluppò un modello basato su alberi decisionali usando un corpus di 22 milioni di token (letteralmente "gettoni"), dopo avere etichettato a mano in base ai sensi circa 2000 linee di concordanza per cinque parole campione.

Tuttavia, malgrado la disponibilità di corpora sempre più ampi, due maggiori ostacoli impediscono l'acquisizione di conoscenza lessicale dai corpora: le difficoltà di etichettare manualmente in base ai sensi un corpus di addestramento e l'esiguità dei dati.

1.4.1 Etichettatura automatica in base al senso

Dal momento che l'etichettatura manuale di un corpus in base al senso è estremamente costosa, sono disponibili soltanto pochi corpora di questo tipo. Molti studi sono stati fatti per creare dei corpora etichettati in base al senso. Miller *et al*. [17] hanno intrapreso l'etichettatura manuale di 1000 parole dal *Brown Corpus* con i sensi di WordNet.

Numerosi sforzi sono stati fatti per etichettare automaticamente in base al senso un corpus di addestramento tramite metodi di *bootstrap* (basati su sequenze di istruzioni). Hearst [18] propone un algoritmo che include una fase di addestramento durante la quale ogni occorrenza di un insieme di sostantivi da disambiguare è etichetta

manualmente in base al senso in parecchie occorrenze. L'informazione statistica estratta dal contesto di queste occorrenze viene successivamente utilizzata per disambiguare altre. Se un'altra occorrenza può essere disambiguata con certezza, il sistema acquista automaticamente informazioni statistiche aggiuntive da queste occorrenze appena disambiguate, migliorando così progressivamente la sua conoscenza. Hearst indica che un insieme iniziale di almeno 10 occorrenze è necessario per il procedimento e che 20 o 30 occorrenze sono necessario per una elevata precisione. Questa strategia complessiva è più o meno quella di molti lavori successivi basati su sequenze di istruzioni.

Gale *et al.* [19] propongono l'uso di corpora bilingue per evitare l'etichettatura manuale di dati di addestramento. La loro idea è che sensi diversi di una data parola spesso sono tradotti diversamente in un'altra lingua (ad esempio, *pen* in inglese si traduce come *penna* o *recinto* in italiano). Usando un corpus parallelo allineato, la traduzione di ogni occorrenza di una parola, come ad esempio la parola inglese *sentence*, può essere usata per determinarne automaticamente il senso. Questo metodo ha dei limiti, dal momento che molte ambiguità rimangono nella lingua in esame; inoltre, i pochi disponibili corpora paralleli su larga scala sono molto specializzati, il che devia la rappresentazione del senso. Dagan *et al.* [20] propongono un metodo simile ma, al posto di un corpus parallelo, usano due corpora monolingue e un dizionario bilingue. Questo risolve in parte i problemi della disponibilità e specificità del dominio, dal momento che i corpora monolingue sono molto più facili a ottenersi dei corpora paralleli.

Magnini e Strappavara [21], lavorando nell'ambito dei testi paralleli, propongono un algoritmo per la disambiguazione del dominio di una parola (Word Domain Disambiguation, WDD), una variante della WSD dove le parole nel testo sono contrassegnate con una etichetta di dominio piuttosto che rispetto al senso. Essi creano dei codici chiamati "Subject Field Codes" (SFC), simili alle etichette di campo utilizzate nei dizionari (ad es. MEDICINA, ARCHITETTURA), per raggruppare parole rilevanti per uno specifico dominio. In questa maniera viene creato un FACTOTUM SFC, che include due tipi di insiemi di sinonimi:

- *Generici*, cioè difficili da classificare in un particolare SFC;
- *Interrompenti i sensi*, cioè che appaiono frequentemente in differenti contesti.

Essi utilizzano la versione inglese di WordNet, e la sua relativa versione italiana MultiWordNet, entrambi arricchiti con etichette di dominio, come principale risorsa di informazioni.

L'algoritmo viene sdoppiato in due sotto-algoritmi per il calcolo della frequenza delle etichette di dominio: il primo incentrato sulla frequenza nel testo, l'altro sulla frequenza dei sensi nelle parole. Ciascun sotto algoritmo, a sua volta, consta di due passi. Prima vengono considerate tutte le parole nel testo e, per ogni etichetta di dominio consentita dalla parola, il punteggio dell'etichetta viene incrementato di uno. Nel secondo passo, ogni parola viene riconsiderata, e l'etichetta di dominio (o le

etichette, a seconda di quante migliori soluzioni sono richieste) con il migliore punteggio viene selezionata come risultato della disambiguazione.

Usando le due versioni allineate di WordNet, Magnini e Strappavara calcolano l'intersezione tra gli insiemi di sinonimi accessibili da un testo inglese attraverso la versione inglese di WordNet e gli insiemi di sinonimi accessibili dalla traduzione parallela in italiano attraverso la versione italiana di WordNet. L'intersezione tra insiemi di sinonimi massimizza il numero di insiemi di sinonimi significanti per i due testi e, allo stesso tempo, esclude insiemi di sinonimi il cui significato non è pertinente al contesto.

Essendo i due WordNet allineati, (cioè condividendo gli offset dei relativi insiemi di sinonimi), l'intersezione può essere determinata direttamente.

Gli esperimenti sono stati condotti su un corpus di 168 articoli paralleli (ogni articolo avente cioè sia la versione italiana sia quella inglese) riguardanti vari argomenti (politica, economia, medicina etc.). e aventi lunghezza media di circa 256 parole. In questo modo si è ottenuta per il primo sotto-algoritmo una precisione dell'83% per l'italiano e dell'85% per l'inglese, e una precisione dell'86% per entrambi i linguaggi mediante il secondo sotto-algoritmo.

1.4.2 Metodi per il superamento della esiguità dei dati

Il problema della esguità dei dati, comune a molti lavori basati su un corpus, è abbastanza rilevante per il lavoro di WSD. In primo luogo, un'enorme quantità di testi è richiesta per assicurare che tutti i sensi di una parola polisemica siano rappresentati, data la grande disparità nella frequenza tra i sensi.

I metodi basati su "raffinamento" sono utilizzati per aggirare il problema di eventi che si verificano poco frequentemente, e in particolare per assicurare che eventi non osservati non vengano considerati aventi probabilità nulla.

I modelli basati su classi cercano di ottenere le migliori stime combinando osservazioni di classi di parole considerate appartenenti ad una categoria comune. Brown *et al.* [22] propongono dei metodi che derivano le classi da proprietà distribuzionali del corpus stesso, mentre altri autori usano fonti di informazione esterna per definire le classi. I metodi basati su classi rispondono in parte al problema della esiguità dei dati, ed eliminano il bisogno di dati pre-etichettati. Tuttavia vi è una certa perdita di informazione con questi metodi poiché l'ipotesi che tutte le parole di una stessa classe si comportino in modo simile è troppo forte.

I metodi basati sulla somiglianza sfruttano la stessa idea di raggruppare osservazioni per parole simili, ma senza raggrupparli nuovamente in classi fisse. Ogni parola ha un insieme di parole simili potenzialmente differente. Come molti metodi basati sulle classi, i metodi basati sulla somiglianza sfruttano una somiglianza metrica tra modelli di co-occorrenza.

1.5 Metodi ibridi

Vari studi sono stati fatti nella WSD sfruttando insieme più risorse quali Internet, corpus semanticamente etichettati (ad es. SemCor), informazioni lessicali ed euristiche etc.

Moldovan e Mihalcea [23] descrivono un metodo per disambiguare i sensi delle parole usando inizialmente Internet per raccogliere informazioni statistiche su occorrenze tra coppie di parole, e successivamente WordNet per valutare le densità semantiche tra le stesse coppie. Allo scopo di ottenere la disambiguazione hanno sviluppato un algoritmo che si articola nei seguenti passi:

1. Si selezionano due parole con i loro sinonimi (così come definiti in WordNet, cioè parole appartenenti allo stesso synset) e si richiede ad un motore di ricerca su Internet (è stato utilizzato AltaVista) il numero di occorrenze di query formate dalla combinazione di tutte le coppie di sinonimi dei vari sensi, allo scopo di determinare la coppia di sensi più comune. Ciò viene effettuato per coppie verbo–sostantivo, aggettivo–sostantivo e avverbio–verbo;

2. Successivamente, le coppie verbo–sostantivo vengono disambiguate prendendo i primi *t* possibili sensi delle parole (così come classificati dall'algoritmo iniziale) e calcolando la "densità concettuale" delle coppie esaminando i gloss forniti da WordNet nella sottostruttura gerarchia. Ciò classifica ogni coppia di sensi riferendosi al"noun-context" (contesto nominale) del verbo e confrontandolo con il sostantivo dato (e la sua sottogerarchia nella struttura ad albero di WordNet).

Tale metodo presenta alcune limitazioni:

- I gloss di WordNet non sono etichettati semanticamente e quindi risulta difficile la loro analisi (etichettare semanticamente una parola significa associare ad essa il rispettivo "Part Of Speech" cioè uno tra: sostantivo, verbo, aggettivo o avverbio);

- L'algoritmo fa affidamento sull'uso della ricerca su Internet per la disambiguazione iniziale, che è un metodo statistico non efficiente;

- L'algoritmo è stato completamente provato solo per coppie di parole, e benché gli autori presentino un esempio di un'applicazione a una frase più lunga, il loro metodo non è stato provato sistematicamente.

Successivamente all'algoritmo sopra descritto, Mihalcea e Moldovan [24] hanno presentato un metodo di disambiguazione iterativo. L'algoritmo presentato disambigua una frase applicando una serie di procedure in successione, in modo tale che ogni procedura successiva venga applicata solo a parole non precedentemente disambiguate. Inizialmente viene svolta una prima elaborazione del testo, applicando un algoritmo di "POS tagging", ovvero una etichettatura semantica, assegnando ad ogni parola il

rispettivo "Part Of Speech (sostantivo, verbo, avverbio o aggettivo) in base al contesto della frase. Successivamente vengono identificati i complessi nominali (es. "pipeline companies") che sono trattati come singolo concetto in WordNet. Quindi vengono marcate alcune parole "certe" in modo tale che, per esse, non venga tentata la disambiguazione: queste parole attualmente includono i verbi "be", "have" e "do". A questo punto vengono eseguite in sequenza le seguenti procedure, allo scopo di spostare progressivamente le parole da un insieme di parole non disambiguate ad un insieme di parole già disambiguate:

1. Riconoscimento di parole appartenenti ad un insieme denominato "Name Entity". Questo insieme di parole contiene nomi di persone, organizzazioni e località;

2. Riconoscimento di parole che hanno soltanto un senso e che quindi vengono marcate con quel senso;

3. Ricerca, mediante un "corpus" semanticamente etichettato (SemCor), delle occorrenze di coppie di parole successive nel testo, W_0-W_1 e W_1-W_2, da cui si determina un senso per W_1 (congiunzioni ed articoli vengono ignorati);

4. Determinazione del "noun-context" (contesto nominale) per ogni sostantivo. Questo è dato da tutti i concetti nei synset degli iperonimi e dai nomi presenti in una finestra di 10 parole in SemCor. Si calcola il numero di parole comuni tra il noun context ed il testo originale dal quale proviene il sostantivo per determinare il miglior senso;

5. Ricerca di parole che hanno una "distanza di connessione" (secondo WordNet) pari a 0 con parole già disambiguate (due parole hanno distanza pari a 0 se appartengono allo stesso synset);

6. Ricerca di due parole non disambiguate che sono semanticamente connesse con una distanza di connessione pari a 0;

7. Ricerca di parole aventi una distanza di connessione (usando WordNet) di 1 con parole già disambiguate (ovvero se sono legate da relazione di iperonimia o iponimia);

8. Ricerca di due parole non ancora disambiguate che sono semanticamente connesse con una distanza di connessione pari a 1.

Questa procedura ottiene una disambiguazione del 55% di verbi e nomi con una accuratezza del 92%. Problemi di questo algoritmo sono:

- circa la metà delle parole resta da disambiguare;

- l'algoritmo non è capace di affrontare aggettivi e avverbi;

- l'algoritmo si basa sull'analisi di SemCor, che ne riduce la sua efficienza;

- l'algoritmo richiede il modulo riconoscitore "Name Entity", che riduce ulteriormente l'efficienza e produce errori.

Lytinen *et al*. [25] presentano un metodo per confrontare un quesito fornito dall'utente con un quesito similare già archiviato usando misure di similarità che sfruttano la "Word Sense Disambiguation". Allo scopo di disambiguare un insieme di parole viene utilizzato WordNet per ricercare il synset che minimizza la somma di tutte le distanze tra coppie di synset. La distanza tra synset è data dal percorso più breve in WordNet tra synset, dove la lunghezza del percorso è data dalla somma delle distanze tra i due synset e un iperonimo comune. Per cercare il miglior synset viene usato un algoritmo che:

1. Trova la minor distanza semantica tra tutte le coppie nell'insieme di termini da disambiguare;

2. Assegna il corrispondente senso alla parola che produce la distanza minore;

3. Trova la distanza minore tra i rimanenti termini ambigui e ciascuna delle parole già disambiguate;

4. Assegna il corrispondente senso alla parola che produce la distanza minore e ripete i passi 3 e 4 finché tutti i termini sono stati disambiguati.

L'algoritmo presentato ha riportato nei test un'assegnazione del senso di circa il 50% dei termini, con una accuratezza del circa 60%. Problemi col presente algoritmo sono:

- la misura della distanza semantica adottata non tiene conto del fatto che alcune parti della struttura ad albero di WordNet sono maggiormente dettagliate di altre. In tale maniera la lunghezza ottenuta produce un indice di similarità non accurato;

- i risultati ottenuti non indicano una procedura di disambiguazione molto robusta.

Capitolo 2
Richiami sugli strumenti adoperati

In questo capitolo vengono descritte le risorse utilizzate per pervenire al raggiungimento degli obiettivi stabiliti. In particolare verranno analizzate la struttura di WordNet e le caratteristiche di Google, il motore di ricerca adoperato.

2.1 WordNet

WordNet [26] è una base di dati lessicale sviluppata alla Princeton University da un gruppo guidato da George Miller [27] [28]. Esso può essere definito come un dizionario basato su principi psicolinguistici.

La differenza più evidente tra WordNet e un dizionario standard è il fatto che WordNet divide il lessico in cinque categorie: sostantivi, verbi, aggettivi, avverbi e parole funzione (function words). Al momento WordNet contiene solo sostantivi, verbi, aggettivi e avverbi. L'insieme relativamente piccolo delle parole funzione è omesso secondo l'ipotesi che esse sono probabilmente immagazzinate separatamente come parte della componente sintattica del linguaggio. La comprensione che le categorie sintattiche differiscono nella organizzazione soggettiva è emersa da studi di associazione di parole.

Il prezzo da imporre per tale categorizzazione sintattica in WordNet è una inevitabile quantità di ridondanza che i dizionari convenzionali evitano. Ma il vantaggio è che differenze fondamentali nell'organizzazione semantica di queste categorie sintattiche possono essere chiaramente viste e sistematicamente utilizzate.

I sostantivi sono organizzati nella memoria lessicale come gerarchie topiche, i verbi sono organizzati tramite una varietà di conseguenti relazioni, e aggettivi e avverbi sono organizzati come iperspazi N-dimensionali.

La più ambiziosa caratteristica di WordNet, tuttavia, è il tentativo di organizzare l'informazione lessicale in termini dei significati delle parole (word meaning) piuttosto che delle loro rappresentazioni ortografiche (word form).

WordNet ricopre la maggior parte dei sostantivi, verbi, aggettivi ed avverbi della lingua inglese. Le parole sono organizzate in insiemi di sinonimi chiamati *synset*. Ogni synset rappresenta un concetto. Attualmente WordNet 1.7.1 possiede una larga rete di 146.350 parole, organizzate in 111.223 synset.

Vi è un vasto insieme di più di 400.000 collegamenti di relazione tra parole, tra parole e synset e tra synset. La relazione semantica di base tra parole, codificate in WordNet, è la relazione di sinonimia. I synset sono legati da relazioni di antonimia, iperonimia/iponimia (is-a) e meronimia/olonimia (part-whole). La Figura 1 mostra una gerarchia semantica di WordNet:

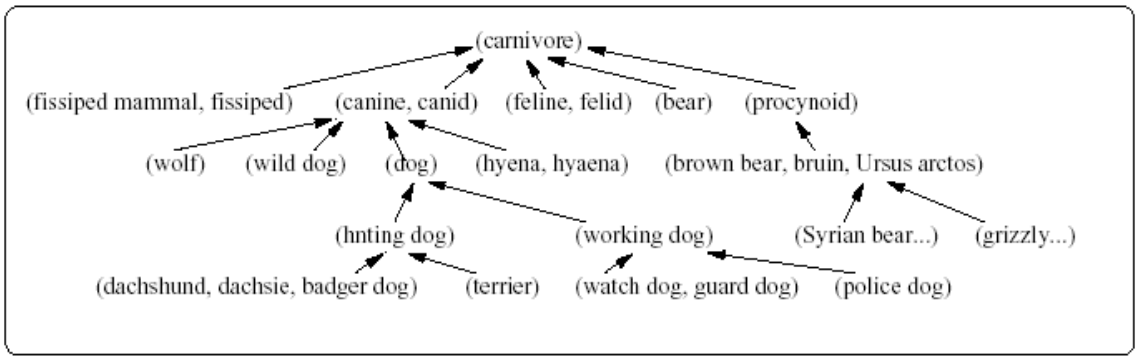

Figura 1 - Una gerarchia di WordNet

2.1.1 La matrice lessicale

La semantica lessicale comincia col riconoscimento che una parola è un'associazione convenzionale tra un concetto lessicalizzato e un'espressione che gioca un ruolo sintattico. Poiché la parola "*word*" è comunemente usata per riferirsi sia all'espressione sia al suo concetto associato, allo scopo di ridurre possibili ambiguità, si userà "word form" per riferirsi all'espressione fisica, e "word meaning" per riferirsi al concetto lessicalizzato che può essere usato per esprimere una forma. Pertanto si può dire che il punto di partenza per la semantica lessicale può essere la corrispondenza tra forme e significati.

Come si vede dalla Figura 2, le "word form" fanno riferimento alle colonne, mentre le "word meaning" alle righe:

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					$E_{m,n}$

Figura 2 - La matrice lessicale

Un elemento di una cella della matrice implica che la word form della colonna corrispondente può essere utilizzata (in un appropriato contesto) per esprimere il significato relativo alla riga.

Così, l'elemento $E_{1,1}$ implica che la word form F_1 può essere usata per esprimere la word meaning M_1. Se vi sono due elementi nella stessa colonna, la word form presenta polisemia; se vi sono due elementi nella stessa riga, le due word forms sono

sinonimi (relativamente a un contesto). Come si può facilmente vedere, F_1 e F_2 sono sinonimi, e F_2 presenta polisemia.

2.1.2 Relazioni e loro significato

WordNet organizza le relazioni in due principali categorie: semantiche e lessicali. Poiché una relazione semantica è una relazione tra significati – come specializzazione (iponimia) e generalizzazione (iperonimia) – e poiché i significati possono essere rappresentati tramite synset, è naturale pensare alle relazioni semantiche come puntatori tra synset. Caratteristica delle relazioni semantiche è la reciprocità: se vi è una relazione semantica R tra i significati $\{x, x', ...\}$ e i significati $\{y, y', ...\}$, allora vi è anche una relazione R′ tra $\{y, y', ...\}$ e $\{x, x', ...\}$.

Una relazione lessicale esprime invece un legame tra word form, come la sinonimia, l'antinomia etc.

Di seguito vengono descritti i vari tipi di relazioni e le loro proprietà.

2.1.2.1 Sinonimia

In base a quanto detto finora, appare ovvio che la più importante relazione per WordNet è la somiglianza di significati, poiché la capacità di giudicare tale relazione tra word form è un prerequisito per la rappresentazione dei significati in una matrice lessicale. Secondo una definizione (generalmente attribuita a Leibniz) due espressioni sono sinonimi se la sostituzione di una per l'altra non cambia la verità di una frase nella quale è applicata la sostituzione.

Si noti che la definizione di sinonimi in termini di sostituzione rende necessario suddividere WordNet in sostantivi, verbi, aggettivi e avverbi. Sarebbe a dire, se i concetti sono rappresentati da synset, e se i sinonimi devono essere intercambiabili, allora parole in differenti categorie sintattiche non possono essere sinonimi poiché non sono intercambiabili. La sinonimia può essere anche pensata come l'estremità di un continuo lungo il quale la similarità di significato può essere graduata.

E' conveniente assumere che la relazione di similarità semantica sia simmetrica: se *x* è simile a *y*, allora *y* è egualmente simile a *x*. La graduabilità della similarità semantica è presente ovunque, ma è più importante per comprendere l'organizzazione dei significati aggettivali e avverbiali.

2.1.2.2 Antinomia

L'antinomia è una relazione lessicale tra word form, la quale indica che una è il contrario dell'altra, ma ciò non è sempre vero. Per esempio, i termini *rich* (ricco) e *poor*

(povero) sono antonimi, ma dire che una persona non sia ricca non implica dire che sia povera. Essa fornisce un principio di organizzazione centrale per gli aggettivi e gli avverbi in WordNet.

2.1.2.3 Iponimia e iperonimia

A differenza di sinonimia e antinomia, che sono relazioni lessicali tra word form, l'iponimia/iperonimia è una relazione tra word meaning. Un concetto rappresentato mediante il synset {*x, x′, ...*} è detto essere un iponimo del concetto rappresentato dal synset {*y, y′, ...*} se madrelingua inglesi accettano la frase costruita come *An x is a (kind of) y* (*Un x è un (tipo di) y*). La relazione può essere rappresentata includendo in {*x, x′, ...*} un puntatore al suo ordine superiore, ed includendo in {*y, y′, ...*} puntatori ai suoi iponimi.

L'iponimia è transitiva e simmetrica, e poiché normalmente vi è un solo ordine superiore, essa genera una struttura semantica gerarchica.

2.1.2.4 Meronimia e olonimia

La meronomia/olonimia è una relazione semantica e corrisponde alla relazione *part-whole* (parte-tutto). Un concetto rappresentato dal synset {*x, x′, ...*} è un meronimo di un concetto rappresentato dal synset {*y, y′, ...*} se madrelingua inglesi accettano la frase costruita come *An y has an x (as a part)* (*Un y ha un x (come parte)*) o *An x is a part of y* (*Un x è una parte di y*). La relazione di meronimia è transitiva e asimmetrica, e può essere utilizzata per costruire una gerarchia di parte. La sua relazione duale è l'olonimia.

2.1.2.5 Relazioni morfologiche

Le relazioni morfologiche tra word form sono una classe importante di relazioni lessicali. WordNet si basa su un vocabolario di lingua inglese, la cui morfologia flessiva è relativamente semplice e si manifesta come declinazione per i sostantivi (ad esempio distinzione tra singolare e plurale) e coniugazione per i verbi. Le relazioni lessicali risultanti dalla morfologia flessiva sono incorporate nell'interfaccia di WordNet, non nel database centrale.

2.1.3 Le descrizioni in WordNet

Quasi tutti i synset in WordNet sono accompagnati da una piccola descrizione chiamata *gloss*. Un gloss consta di una definizione, dei commenti e degli esempi. Per esempio, il gloss del synset {dog, domestic dog, Canis familiaris} è (a member of the genus Canis (probably descended from the common wolf) that has been domesticated

by man since prehistoric times; occurs in many breeds; "the dog barked all night"). Esso possiede una definizione (a member of the genus Canis), un commento (probably descended from the common wolf) e un esempio (the dog barked all night), come mostrato in Figura 3.

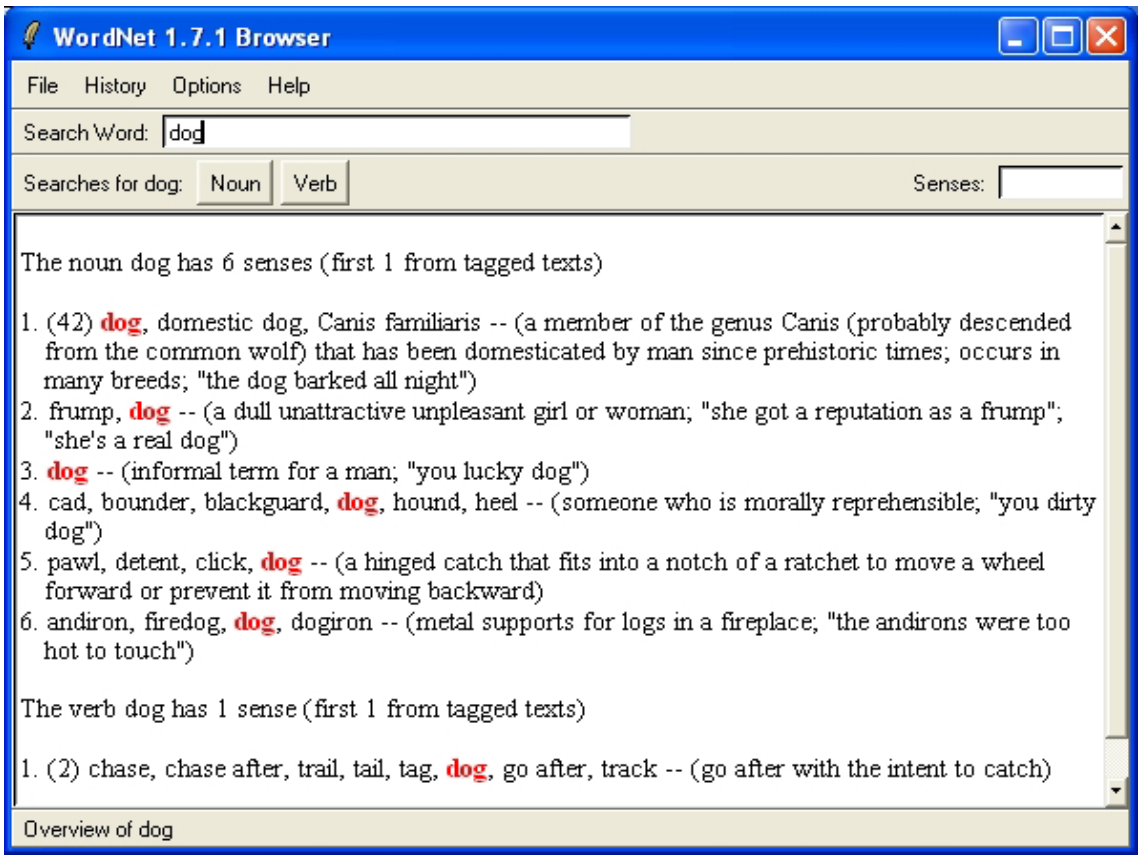

Figura 3 - Panoramica della parola "dog" in WordNet

Alcuni gloss possono contenere definizioni multiple o commenti multipli.

2.1.4 I sostantivi in WordNet

Nelle moderne teorie psicolinguistiche, per definire un sostantivo si usa solitamente un termine di ordine superiore, cioè di significato più generale, insieme a caratteristiche che lo distinguono. Tale metodo è anche alla base dell'organizzazione dei file comprendenti i sostantivi in WordNet.

La relazione di ordine superiore (iponimia) genera una organizzazione gerarchica semantica che è riprodotta nei file dei sostantivi mediante l'uso di puntatori etichettati tra i synset. Tale gerarchia è limitata in profondità, e raramente va oltre il dodicesimo livello.

Le caratteristiche distintive sono immesse in modo tale da creare un sistema di eredità lessicale, un sistema in cui ogni parola eredita le caratteristiche distintive di tutti i suoi termini di ordine superiore.

Piuttosto che costruire una singola gerarchia ad albero, i sostantivi sono stati suddivisi in un insieme di principi semantici chiamati *beginners* (principianti), allo scopo di scegliere un numero ristretto di concetti generici e trattare ciascuno come l'unico beginner di una gerarchia separata.

Ciò offre come vantaggio il fatto che, poiché le caratteristiche che distinguono ciascun beginner unico sono ereditate da tutti i suoi iponimi, un vocabolo eredita le caratteristiche del proprio beginner; inoltre si ha una notevole riduzione delle dimensioni dei file su cui i lessicografi devono lavorare, ed è possibile assegnare la scrittura e la modifica di differenti file a differenti lessicografi.

WordNet 1.7.1 utilizza 25 beginners per i sostantivi, come mostrato in Figura 4.

{act, action, activity}	{natural object}
{animal, fauna}	{natural phenomenon}
{artifact}	{person, human being}
{attribute, property}	{plant, flora}
{body, corpus}	{possession}
{cognition, knowledge}	{process}
{communication}	{quantity, amount}
{event, happening}	{relation}
{feeling emotion}	{shape}
{food}	{state, condition}

Figura 4 - Lista dei beginners per i sostantivi in WordNet

2.2 Il motore di ricerca Google

Google [29] è uno dei motori di ricerca attualmente più utilizzati in rete. Sviluppato all'università di Stanford da Sergey Brin e Larry Page, ha debuttato on-line a fine settembre 1999 dopo oltre un anno di test.

Vi sono molte caratteristiche che fanno di Google uno dei più potenti motori di ricerca, la più importante delle quali è quella di selezionare i risultati di ricerca valutando l'importanza di ogni pagina web con metodi matematici, in base ad un controllo di oltre 500 milioni di variabili e di 2 miliardi di termini. Questa tecnologia, chiamata **PageRank** ed attualmente in fase di brevetto, controlla non solo il contenuto della pagina web, ma verifica anche altri eventuali siti che hanno un *link* (collegamento) verso la pagina: in base alla quantità ed al tipo di link, la pagina riceve una valutazione più o meno alta.

Ciascuna delle pagine web indicizzate dallo spider di Google viene catalogata secondo due aspetti. Da un lato sulla base del numero di occorrenze di un certo termine, dall'altro da un punto di vista puramente topografico.

Ogni pagina web, infatti, è raggiungibile perché esistono altre pagine web che presentano un link ad essa. Da un punto di vista topografico, quindi, una pagina web può essere osservata come una struttura con un certo numero di entrate (le pagine che la collegano) e con un certo numero di uscite (le pagine collegate).

Maggiore è il numero di link entranti, più alta sarà l'autorevolezza di quella pagina e, allo stesso tempo, maggiore sarà la probabilità che i siti che questa collega siano anch'essi autorevoli. Naturalmente, più la pagina è collegata, maggiori saranno le possibilità che essa venga visitata e, contemporaneamente, più una pagina viene visitata, maggiori saranno le possibilità che le pagine da lei collegate siano a loro volta visitate. Da un punto di vista matematico questa visita è una catena di Markov. Si dà un insieme di stati e ad ogni istante di tempo ciascuno di questi ha una certa probabilità di transitare ad un altro stato. Questa probabilità è il *ranking* (classifica) effettuato da Google.

Vengono, pertanto, compiuti due tipi di ricerca incrociati. Il primo, più tradizionale, serve per individuare un certo numero di pagine che contengono la parola ricercata, il secondo per individuare, tra queste, quale sia più utile al navigatore.

Il PageRank rappresenta l'indicatore generale dell'importanza che Google attribuisce ad una determinata pagina web, indipendentemente dalla specifica interrogazione che genera l'elenco. L'ordine in cui vengono visualizzati i risultati dipende, quindi, dalle caratteristiche delle pagine stesse, ossia dai dati del web che Google analizza utilizzando complessi algoritmi che studiano la struttura dei link.

Naturalmente, una pagina "importante" non è di grande interesse per l'utente se non contiene il termine da lui ricercato. Per questa ragione Google utilizza sofisticate tecniche di analisi del testo per trovare pagine che siano nello stesso tempo importanti e attinenti dal punto di vista dell'interrogazione. Per analizzare una pagina, ad esempio, Google considera quello che di questa pagina dicono altre pagine contenenti link che rinviano ad essa.

Tali tecniche innovative rendono del tutto impossibile l'interferenza dell'uomo sui risultati delle ricerche, diversamente dagli altri motori di ricerca. Google è progettato in modo tale che non si possa "comprare" un PageRank più alto o alterare l'elenco dei risultati per ragioni di ordine commerciale.

Oltre alla suddetta tecnica di valutazione dell'importanza delle pagine, Google restituisce esclusivamente pagine che contengono tutti i termini di ricerca inseriti dall'utente, o nel testo della pagina o nei link che rimandano a quella pagina. Google analizza anche la vicinanza tra questi termini all'interno della pagina e dà la priorità a determinate pagine sulla base della vicinanza tra i vari termini di ricerca all'interno della stessa. Vengono privilegiate, in altre parole, quelle pagine in cui i termini risultano molto vicini tra di loro, in modo da minimizzare il tempo necessario per scartare i risultati irrilevanti.

Altre caratteristiche di Google sono la sua velocità di esecuzione, con un tempo medio dichiarato per ogni ricerca di 0,29 secondi, e la grandezza del suo archivio che ha quasi 3 miliardi di pagine censite a cui vengono aggiunte 1,5 milioni di nuove pagine al giorno.

Google inoltre supporta la ricerca tramite operatori booleani complessi, impostando inizialmente una ricerca con l'operatore booleano AND inserito a priori. Pertanto esso ricerca documenti che contengano tutte le parole inserite nel campo di ricerca. E' anche possibile utilizzare l'operatore OR per ricercare pagine che includano almeno una delle parole inserite nel campo di ricerca. Inoltre Google "preferisce" i siti che hanno le parole scelte vicine tra loro.

Altra ragione per cui si è deciso di utilizzare Google è la possibilità di interfacciarsi ad esso. Lo staff di Google ha reso infatti disponibile una serie di API [30] per la realizzazione di programmi che possano interrogare il motore di ricerca. In questo modo è possibile sviluppare applicazioni (con Visual Studio Net, Java, ecc.) che si connettano a Google, eseguano una query ed estraggano i risultati. Le Google API sono attualmente in fase beta e sono quindi soggette ad alcune limitazioni:

- possono essere eseguite al massimo 1000 interrogazioni al giorno;
- sono restituiti al massimo 100 risultati per interrogazione;
- è possibile accedere solo ai primi 1000 risultati di una interrogazione;
- non è garantita la continuità del servizio.

Per potere utilizzare le API di Google, è necessario registrarsi per ottenere una License-key (chiave di licenza) soggetta ai termini del contratto.

Capitolo 3
Soluzione proposta

Nel presente capitolo vengono analizzate nel dettaglio le varie fasi del processo. Il modello sviluppato può essere schematizzato come mostrato in Figura 5, nella quale si possono distinguere due fasi fondamentali: la creazione di un dizionario contenente le traduzioni dei termini da disambiguare, la quale viene eseguita solo una volta, e il processo vero e proprio di disambiguazione dei termini lessicali presenti nella query.

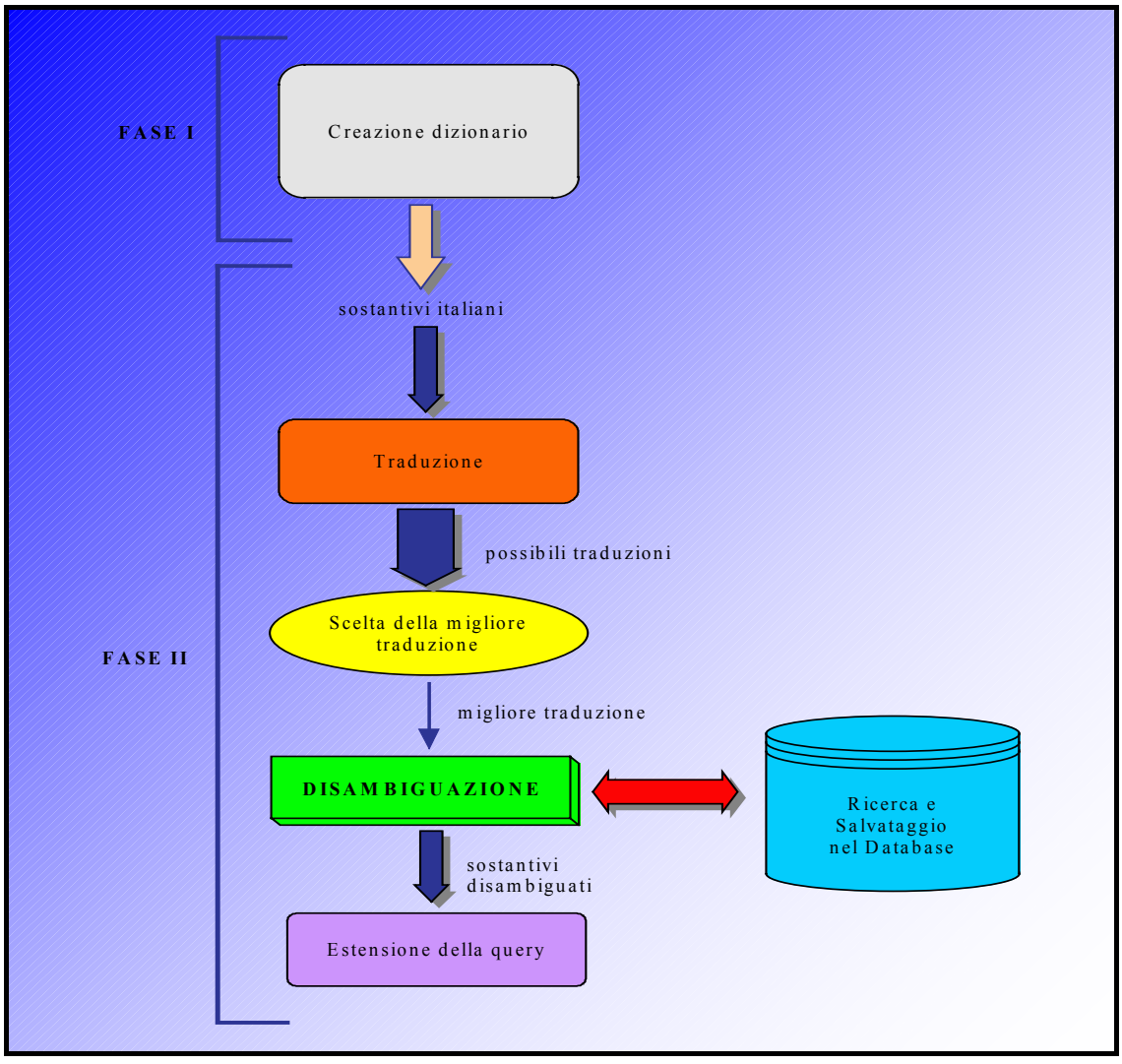

Figura 5 - Schema del modello sviluppato

3.1 Struttura dell'algoritmo

Il metodo presentato sfrutta il vantaggio del contesto della frase. Similmente a quanto visto in [23], le parole vengono accoppiate a due a due e si tenta di disambiguare

una parola nel contesto dell'altra. Questo è reso possibile tramite una ricerca in Internet con delle query generate mediante l'utilizzo di WordNet. I sensi di ogni parola verranno ordinati semplicemente mediante il calcolo di un valore detto "distanza di Tanimoto". In tal modo tutte le parole vengono processate e tutti i sensi di queste ultime vengono ordinati.

Come già mostrato in Figura 5, il modello sviluppato si articola in 7 passi fondamentali:

1. La **creazione del dizionario**, eseguita solo una volta;
2. La **traduzione** dei sostantivi dall'italiano all'inglese;
3. La **scelta della migliore traduzione** tra tutte quelle possibili;
4. La **ricerca nel database** di tutte le possibili coppie di parole che si possono ottenere dai sostantivi tradotti;
5. La **disambiguazione** della migliore traduzione. Questa fase include l'utilizzo di WordNet e del motore di ricerca per ottenere i migliori sensi di ciascuna parola. In questa fase vengono scelti i sensi più opportuni per i sostantivi inglesi.
6. L'**estensione della query**. Anche in questa fase viene utilizzato WordNet per ottenere i sinonimi dei sostantivi inglesi disambiguati.
7. L'eventuale **salvataggio** nel Database.

3.1.1 Creazione del dizionario

Il primo passo è stato quello di generare un dizionario contenente i principali sostantivi italiani e le rispettive traduzioni in lingua inglese. La realizzazione di tale dizionario è stata fatta utilizzando la tecnica dello *scraping* (letteralmente "raschiatura") di pagine in formato HTML. Questa tecnica consiste nell'estrarre una serie di informazioni utili dal codice HTML di una pagina web mediante tecniche di *pattern-matching*, cioè mediante confronto con un modello (pattern) – ad esempio una stringa – definito in origine.

A tal proposito, disponendo di una lista dei principali sostantivi italiani, si è sfruttato un sito di traduzioni in linea [2] per ottenere le traduzioni di ciascun sostantivo. Per ogni singolo sostantivo italiano, viene creato l'indirizzo Internet della pagina corrispondente alla richiesta di traduzione della parola in esame al sito di traduzioni in linea. Pertanto, supponendo che la parola di cui si desidera la traduzione sia la generica parola **W**, l'indirizzo generato sarà del tipo:

```
http://www.allwords.com/query.php?SearchType=3&Keyword="+W
+"&Language=ITA&v=17234684
```

Una volta ottenuto il codice HTML della pagina richiesta, si effettua lo scraping, andando a ricercare nel codice la stringa “word=” ed estraendo le parole che sono racchiuse tra essa e il carattere “?”, eliminando eventuali ripetizioni.

Queste parole, che rappresentano le traduzioni della parola originale, vengono quindi salvate in un apposito file di dizionario. In questo modo la creazione del dizionario avviene in maniera completamente automatica.

3.1.2 Traduzione

La fase di traduzione è la più semplice del processo. Questa fase si limita a ricercare nel file di dizionario le traduzioni dei sostantivi italiani inizialmente inseriti.

Inizialmente viene creato un vettore di dimensione pari alle parole inserite. Ciascun sostantivo, con le relative traduzioni, viene ricercato nel file di dizionario. Se il sostantivo è presente, viene identificato il numero di linea del file in cui esso compare e viene letto il numero di traduzioni che possiede. Le traduzioni del sostantivo vengono quindi salvate in una apposita struttura dati per potere essere reperite successivamente.

3.1.3 Scelta della migliore traduzione

Disponendo di tutte le possibili combinazioni di traduzioni inglesi dei sostantivi, è fondamentale che venga scelta la traduzione appropriata di ogni parola rispetto alle altre, cioè in funzione del contesto. Ad esempio, limitandoci al caso di due parole, se la frase composta da sostantivi italiani è “rete computer”, e le traduzioni possibili presenti nel file di dizionario sono {net, network} per “rete” e {computer} per “computer”, si desidera che la traduzione scelta sia quella inerente al contesto, e cioè “network computer”. Analogamente, se la coppia in esame è “rete pesca” e le traduzioni disponibili sono {net, network} per “rete” e {fishing, peach} per “pesca”, la traduzione più inerente al contesto sarà “net fishing”.

Per potere valutare quale sia la traduzione semanticamente più corretta tra tutte le possibili combinazioni delle traduzioni, è necessario introdurre una metrica; si è scelto di utilizzare a tale scopo la **distanza di Tanimoto**.

Su due insiemi A e B, tale similitudine è definita come:

$$T_{AB} = \frac{|A \cap B|}{|A \cup B|} \qquad (2.1)$$

Data una query Q, indicando con ò(Q) il numero di risultati ottenuti per essa da Google, si può estendere la distanza di Tanimoto al caso in cui i fattori della (2.1) siano delle query. Pertanto la (2.1) diviene:

$$T_{AB} = \frac{\partial(A \cap B)}{\partial(A \cup B) - \partial(A \cap B)} \tag{2.2}$$

Dati N sostantivi italiani S_1, S_2, …, S_N , con $N \leq 5$, siano S_1^1, S_1^2, …, S_1^n le traduzioni disponibili per S_1, siano S_2^1, S_2^2, …, S_2^m le traduzioni disponibili per S_2 e S_N^1, S_N^2, …, S_N^q le traduzioni disponibili per S_N rispettivamente.

Per valutare quale sia la migliore tra tutte le possibili traduzioni $S_1^i \times S_2^j \times \ldots \times S_N^k$, con i = 1, …, n, j = 1, …, m, e k = 1, …, q, per ciascuna combinazione di traduzioni si calcola la distanza di Tanimoto ottenuta estendendo la (2.1) a N insiemi, cioè prendendo le traduzioni $S_1^i \times S_2^j \times \ldots \times S_N^k$, con i = 1, …, n, j = 1, …, m, e k = 1, …, q, con $n, m, q \in \mathbb{N}$, dei rispettivi sostantivi S_1 S_2 … S_N.

Il numeratore e il denominatore della (2.2) sono il numero dei risultati ottenuti inviando al motore di ricerca utilizzato (Google) le query corrispondenti, pertanto la (2.2) si può riscrivere come:

$$T_{ij\ldots k} = \frac{\grave{o}\left(S_1^i \textbf{ AND } S_2^j \textbf{ AND } \cdots S_N^k\right)}{\grave{o}\left(S_1^i \textbf{ OR } S_2^j \textbf{ OR } \cdots S_N^k\right) - \grave{o}\left(S_1^i \textbf{ AND } S_2^j \textbf{ AND } \cdots S_N^k\right)} \tag{2.3}$$

per i = 1, …, n, j = 1, …, m e k = 1, …, q, con $n, m, q \in \mathbb{N}$.

Viene quindi selezionata la traduzione avente distanza di Tanimoto maggiore tra tutte.

3.1.4 Ricerca nella base di dati

Allo scopo di migliorare l'efficienza dell'algoritmo, prima della fase di disambiguazione è stata prevista la possibilità di potere verificare, tramite una base di dati, se esistano coppie di sostantivi già preventivamente disambiguate. In tal caso ciò permetterà, durante i vari passi dell'algoritmo, di disporre immediatamente dei risultati, evitando così i tempi di attesa dovuti alle numerose interrogazioni al motore di ricerca di Google.

La strategia di ricerca nella base di dati è incentrata sulla ricerca di coppie di sostantivi. Ciò non preclude l'utilizzo della base di dati per query formate da più di due parole. Infatti, dal momento che, come si vedrà nel paragrafo 3.1.5, l'algoritmo procede per coppie di parole, la sua interrogazione avviene per tutte le coppie che vengono formate durante la disambiguazione.

In altri termini, supponendo di introdurre una query di N parole W_1, W_2, …, W_N, per ogni coppia di parole W_i-W_j (i, j = 1, …, N, $i \neq j$) viene preventivamente interrogata la base di dati per verificarne la sua eventuale presenza. Se la coppia risulta essere presente, vengono restituiti i corrispondenti risultati, cioè il senso di W_i e la distanza di Tanimoto della coppia, altrimenti si procede mediante interrogazione al motore di ricerca.

3.1.5 Disambiguazione

Avendo classificato i sostantivi tradotti in funzione della loro distanza di Tanimoto, il passo successivo che viene effettuato è quello della disambiguazione dei sensi relativi a tali sostantivi.

La fase di disambiguazione prevede una serie di sotto-fasi. Per ogni sostantivo inglese, mediante l'utilizzo di WordNet, si generano una serie di query contenenti i sostantivi estratti dalle definizioni (gloss) e i sinonimi del sostantivo relativamente al senso in esame.

Queste query vengono utilizzate per potere sfruttare la metrica scelta, cioè la distanza di Tanimoto, allo scopo di classificare i sensi e scegliere quello più inerente al contesto.

3.1.5.1 Filtraggio delle descrizioni

La prima fase della disambiguazione consiste nel filtraggio delle descrizioni relative ai sostantivi inglesi ottenuti come miglior traduzione.

Come precedentemente descritto, ogni descrizione – o gloss – è costituito da una definizione, da commenti e da esempi, anche se in taluni casi può capitare che questi ultimi due non siano presenti e vi sia la sola definizione.

Per ciascun sostantivo si estrae il relativo gloss, quindi si provvede in prima istanza ad eliminare la punteggiatura in esso presente, poi si elimina tutto ciò che risulta racchiuso tra parentesi tonde (cioè un eventuale commento) e tra virgolette (eventuali esempi). Successivamente si analizzano le parole rimaste presenti, delle quali vengono prese in considerazione solo quelle aventi almeno un senso relativamente alla categoria "sostantivo".

Queste parole vengono confrontate con una lista di "stop words", cioè una lista di parole inglesi di uso molto comune, allo scopo di eliminare parole che potrebbero appesantire la ricerca in Internet a causa del loro significato troppo generico. Si noti, che trattando solo disambiguazione di sostantivi, si utilizza una lista di "stop words" relativa solamente a parole comuni presenti come sostantivi in WordNet. Vengono ad esempio eliminati i termini "something", "anything" etc.

Supponendo che la frase ottenuta dopo la scelta della migliore traduzione sia "W_1 W_2 … W_N", e supponendo che W_1 abbia m sensi W_1^1, W_1^2, …, W_1^m, $m \in \mathbb{N}$, per ciascun senso W_1^k di W_1 si ottengono q sostantivi N_{1k}^1, N_{1k}^2, …, N_{1k}^q, k = 1… m, $q \in \mathbb{N}$, estratti dal gloss relativo a tale senso.

In maniera analoga si estraggono i sostantivi dai gloss di tutte le altre W_i, i = 1, …, N. In questo modo si ottengono i seguenti insiemi:

$$\begin{aligned} &N(W_1): \{[N_{11}^1, N_{11}^2, \ldots, N_{11}^q], [N_{12}^1, N_{12}^2, \ldots, N_{12}^q], \ldots, [N_{1m}^1, N_{1m}^2, \ldots, N_{1m}^q]\} \\ &N(W_2): \{[N_{21}^1, N_{21}^2, \ldots, N_{21}^q], [N_{12}^1, N_{12}^2, \ldots, N_{22}^q], \ldots, [N_{2m}^1, N_{2m}^2, \ldots, N_{2m}^q]\} \\ &\ldots\ldots\ldots\ldots\ldots \\ &N(W_N): \{[N_{N1}^1, N_{N1}^2, \ldots, N_{N1}^q], [N_{N2}^1, N_{N2}^2, \ldots, N_{N2}^q], \ldots, [N_{Nm}^1, N_{Nm}^2, \ldots, N_{Nm}^q]\} \end{aligned} \quad (2.4)$$

3.1.5.2 Estrazione dei sinonimi

Sempre considerando che la migliore frase tradotta sia "W_1 W_2 … W_N", supponendo che W_1 abbia m sensi W_1^1, W_1^2, …, W_1^m, $m \in \mathbb{N}$, per ciascun senso W_1^k di W_1, oltre ai q sostantivi N_{1k}^1, N_{1k}^2, …, N_{1k}^q, k = 1, …, m, $q \in \mathbb{N}$, estratti dal gloss relativo a tale senso, si prendono anche i p sinonimi relativi a W_1^k, cioè s_{1k}^j, con j = 1, …, p, k = 1, …, m, $p \in \mathbb{N}$.

Analogamente si procede per tutti gli altri sostantivi W_i, ottenendo i seguenti insiemi:

$$\begin{aligned} &S(W_1): \{[s_{11}^1, s_{11}^2, \ldots, s_{11}^p], [s_{12}^1, s_{12}^2, \ldots, s_{12}^p], \ldots, [s_{1m}^1, s_{1m}^2, \ldots, s_{1m}^p]\} \\ &S(W_2): \{[s_{21}^1, s_{21}^2, \ldots, s_{21}^p], [s_{12}^1, s_{12}^2, \ldots, s_{22}^p], \ldots, [s_{2m}^1, s_{2m}^2, \ldots, s_{2m}^p]\} \\ &\ldots\ldots\ldots\ldots\ldots \\ &S(W_N): \{[s_{N1}^1, s_{N1}^2, \ldots, s_{N1}^p], [s_{N2}^1, s_{N2}^2, \ldots, s_{N2}^p], \ldots, [s_{Nm}^1, s_{Nm}^2, \ldots, s_{Nm}^p]\} \end{aligned} \quad (2.5)$$

3.1.5.3 Classifica dei sensi

Una volta effettuato il filtraggio delle descrizioni di una parola e l'estrazione dei sinonimi ad essa relativi, il passo successivo è quello di combinare i sostantivi ottenuti da tale filtraggio e i sinonimi con tutti gli altri sostantivi tradotti, allo scopo di inviare al motore di ricerca delle query e poterle classificare in base al numero di risultati ottenuti. Poiché, come precedentemente accennato nel paragrafo 3.1.4, si è scelto di procedere per coppie di parole, tale combinazione dovrà essere eseguita per tutte le possibili coppie di parole W_i-W_j, con i, j = 1… N, $i \neq j$.

Risulta pertanto chiaro che, essendo una query costruita in funzione di ciascun senso di ogni parola (cioè mediante le sue definizioni e i suoi sinonimi), ad ogni query è affidata la valutazione di un determinato senso.

Allo scopo di ottenere una classifica dei sensi, per ogni coppia di sostantivi inglesi si calcola la distanza di Tanimoto di una parola rispetto all'altra. Analogamente a quanto avviene per la scelta della migliore traduzione[3], dal momento che ogni query ottiene un determinato punteggio quando inviata al motore di ricerca, la distanza di Tanimoto si calcola prendendo il numero di risultati ottenuti inviando a Google delle query composte dal numeratore e dal denominatore che compaiono nella (2.2).

Nel caso della coppia di parole W_i-W_j, con i, j = 1, 2, …, N e $i \neq j$, riferendoci alla (2.2), l'insieme A sarà costituito dalla parola W_i tenuta fissa, mentre l'insieme B sarà l'OR logico dei sostantivi N_{jk}^q estratti dai gloss e dei sinonimi s_{jk}^p per ciascun senso W_j^k di W_j:

$$A = W_i \tag{2.6}$$

$$B = s_{jk}^1 \ \mathbf{OR}\ s_{jk}^2 \ \mathbf{OR} \cdots\ s_{jk}^p \ \mathbf{OR}\ N_{jk}^1 \ \mathbf{OR}\ N_{jk}^2 \ \mathbf{OR} \cdots N_{jk}^q \tag{2.7}$$

pertanto la (2.2) si può facilmente riscrivere come:

$$\begin{aligned} T_{ji}^k = {} & \grave{o}\left(W_i \ \mathbf{AND}\ (s_{jk}^1 \ \mathbf{OR}\ s_{jk}^2 \ \mathbf{OR} \cdots\ s_{jk}^p \ \mathbf{OR}\ N_{jk}^1 \ \mathbf{OR}\ N_{jk}^2 \ \mathbf{OR} \cdots N_{jk}^q)\right) \cdot \\ & \cdot \left[\grave{o}\left(W_i \ \mathbf{OR}\ s_{jk}^1 \ \mathbf{OR}\ s_{jk}^2 \ \mathbf{OR} \cdots\ s_{jk}^p \ \mathbf{OR}\ N_{jk}^1 \ \mathbf{OR}\ N_{jk}^2 \ \mathbf{OR} \cdots N_{jk}^q\right) - \right. \\ & \left. - \grave{o}\left(W_i \ \mathbf{AND}\ (s_{jk}^1 \ \mathbf{OR}\ s_{jk}^2 \ \mathbf{OR} \cdots\ s_{jk}^p \ \mathbf{OR}\ N_{jk}^1 \ \mathbf{OR}\ N_{jk}^2 \ \mathbf{OR} \cdots N_{jk}^q)\right)\right]^{-1} \end{aligned} \tag{2.8}$$

per k = 1, … m, i, j = 1, 2, …, N, $i \neq j$, p, q $\in$ □ .

E' importante considerare che risulta:

$$T_{ji}^k \neq T_{ij}^k \tag{2.9}$$

e pertanto la (2.8) non gode della proprietà di simmetria, poiché invertendo l'ordine delle parole, gli insiemi che compaiono come operandi nella applicazione della distanza di Tanimoto risultano differenti. Infatti, T_{ji}^k calcola la distanza di Tanimoto tra gli insiemi $A = W_i$ e $B = [s_{jk}^1 \ \mathbf{OR} \ldots s_{jk}^p \ \mathbf{OR}\ N_{jk}^1 \ \mathbf{OR} \ldots N_{jk}^q]$, mentre per T_{ij}^k si ha $A = W_j$ e $B = [s_{ik}^1 \ \mathbf{OR} \ldots s_{ik}^p \ \mathbf{OR}\ N_{ik}^1 \ \mathbf{OR} \ldots N_{ik}^q]$.

Avendo riscontrato, durante le prove effettuate, che query con molti termini legati tra loro da operatori booleani forniscono risultati differenti da query formate dagli

[3] Si veda a riguardo il paragrafo 3.1.3.

stessi termini considerati però uno alla volta, si è scelto di effettuare le query singolarmente, ad esempio inviando al motore di ricerca la generica query A **AND** (B **OR** C) come due query distinte A **AND** B e A **AND** C, e sommando poi i risultati ottenuti dalle singole query. La (2.8) si può quindi scrivere come:

$$T_{ji}^{k} = \frac{\sum_{r=1}^{p} \partial\left(W_i \ \mathbf{AND}\ s_{jk}^{r}\right) + \sum_{l=1}^{q} \partial\left(W_i \ \mathbf{AND}\ N_{jk}^{q}\right)}{\partial\left(W_i\right) + \sum_{r=1}^{p} \partial\left(s_{jk}^{r}\right) + \sum_{l=1}^{q} \partial\left(N_{jk}^{q}\right) - \sum_{l=1}^{q} \partial\left(W_i \ \mathbf{AND}\ N_{jk}^{q}\right)} \qquad (2.10)$$

per k = 1… m, i, j = 1, 2,…, N, $i \neq j$, $p, q \in$, dove le sommatorie si riferiscono al numero di risultati ottenuti dal motore di ricerca per quella data interrogazione.

Ovviamente il dovere generare numerose query singole comporta un notevole aumento dei tempi di calcolo, a discapito dell'efficienza.

3.1.5.4 Coefficiente di incertezza

Poiché si può presentare il caso in cui una parola ottenga dei punteggi molto vicini per alcuni dei suoi sensi, si è tenuto in considerazione il fatto che un senso potrebbe prevalere sugli altri con uno scarto talmente basso da non avere la certezza che esso sia quello effettivamente corretto.

Per avere un riscontro dello scarto ottenuto tra i primi due migliori sensi, si è quindi introdotto un **indice di incertezza I_u**, definito come il rapporto tra le distanze di Tanimoto di questi ultimi:

$$I_u = \frac{T_{ji}^{k_2}}{T_{ji}^{k_1}} \qquad (2.11)$$

Quanto più I_u risulta essere prossimo a 1, tanto più è possibile che vi sia una ambiguità tra il senso scelto come migliore e quello ad esso immediatamente successivo.

3.1.6 Salvataggio nella base di dati

Il salvataggio nella base di dati consente di memorizzare tutte le coppie di sostantivi presenti nella frase che hanno prodotto la migliore distanza di Tanimoto e il senso relativo alla prima delle due parole della coppia. Infatti, dal momento che la fase di disambiguazione viene compiuta per tutte le coppie di parole, nel salvataggio verranno memorizzate tutte le coppie W_i-W_j con la relativa distanza di Tanimoto T_{ji}, e col miglior senso W_i^* ottenuto per W_i.

3.1.7 Estensione della query

Una volta disambiguati tutti i termini della query e ottenuto il miglior senso per ciascuna di esse, il passo successivo prevede l'estensione della query.

Dati gli N sostantivi inglesi "W_1 W_2 … W_N" e chiamati W_1^*, W_2^*, …, W_N^* i migliori sensi ottenuti per W_1, W_2, …, W_N rispettivamente, siano S_1, S_2, …, S_N gli insiemi di sinonimi (synset) relativi a W_1^*, W_2^*, …, W_N^*. Ciascun synset S_i può contenere zero o più sinonimi $s_i^0, s_i^1, \ldots, s_i^{p_i}$ con i = 1, 2, …, N dove ovviamente $s_i^0 \equiv W_i$.

L'estensione della query prevede, per i termini disambiguati, la generazione di liste di similarità L (similarity lists) equivalenti alla prima, ottenute combinando tra loro tutti gli s_i^j, j = 0… p_i, dei termini. Pertanto si avrà:

$$L = \{s_1^0, \ldots, s_1^{p_1}\} \times \{s_2^0, \ldots, s_2^{p_2}\} \times \ldots \times \{s_N^0, \ldots, s_N^{p_N}\} \qquad (2.12)$$

Capitolo 4
Risultati sperimentali

Nel presente capitolo verranno riportate alcune prove di funzionamento del modello sviluppato ed esaminati i risultati sperimentali ottenuti.

4.1 Prove di funzionamento

Si riportano due prove di funzionamento eseguite su piattaforma Linux Red Hat 8.0 mediante i seguenti strumenti:

- J2SE™ SDK 1.4.1;
- WordNet 1.7.1;
- Jakarta Tomcat 4.1.18;
- PostgreSQL 7.3.2.

Verrà mostrata in principio una prova preliminare di funzionamento per la disambiguazione di due parole.

4.1.1 Prova di funzionamento per due parole

Si desidera determinare il senso corretto del sostantivo italiano “rete” nel caso in cui venga inserito in due contesti differenti, rispettivamente nell’ambito informatico e in quello marino. Sono state effettuate pertanto due query distinte, la prima accoppiando la parola “rete” con il sostantivo “computer”, e la seconda combinandola col sostantivo “pesca”.

Mediante il browser Mozilla, viene visualizzata la pagina Jsp che esegue il programma. Nell’apposita riga si inserisce la query della quale si desidera effettuare la disambiguazione e si clicca sul pulsante “search”.

Si introduca ad esempio per prima la query “rete computer”: il sostantivo “rete” presenta, nel file di dizionario utilizzato, due possibili traduzioni (net, network), mentre il sostantivo “computer”, nella traduzione, rimane invariato, disponendo di una sola traduzione coincidente con la parola stessa.

Vengono pertanto generate tutte le possibili coppie di traduzioni, valutando ciascuna di esse in base alla distanza di Tanimoto riportata, e scegliendo la coppia avente la distanza maggiore fra tutte.

In Figura 6 è riportata la visualizzazione del flusso di stampa prodotto, da cui si può notare che la traduzione “network computer” ottiene distanza di Tanimoto maggiore della traduzione “net computer”, come mostrato in Figura 7:

```
Dictionary initialized...

Number of words present into the query: 2
matching "rete" in line: 4551
Available translations for the word "rete":
net
network
matching "computer" in line: 1346
Available translations for the word "computer":
computer

Combining translations...
net computer
N results for: net computer: 5110000
N results for: net OR computer: 8520000
network computer
N results for: network computer: 5070000
N results for: network OR computer: 7780000
Ranking translations combinations:
"network computer" - Tdist = 1.8708487
"net computer" - Tdist = 1.4985337
```

Figura 6 - Visualizzazione del flusso di stampa prodotto

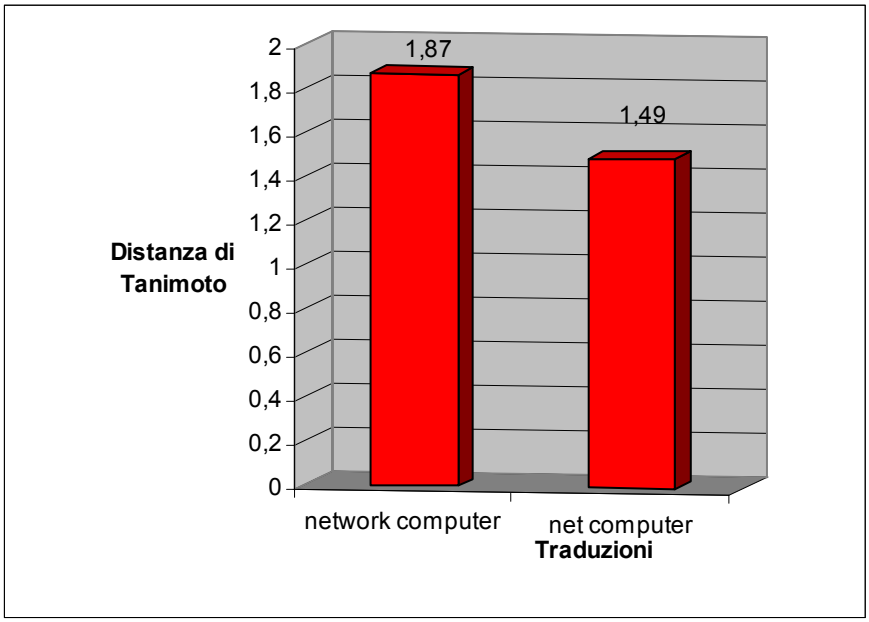

Figura 7 - Classifica delle traduzioni della query “rete computer”

Si esamini quindi la traduzione ottenuta: “network computer”. La parola “network” presenta in WordNet 3 sensi, come si può vedere dalla panoramica riportata in Figura 8:

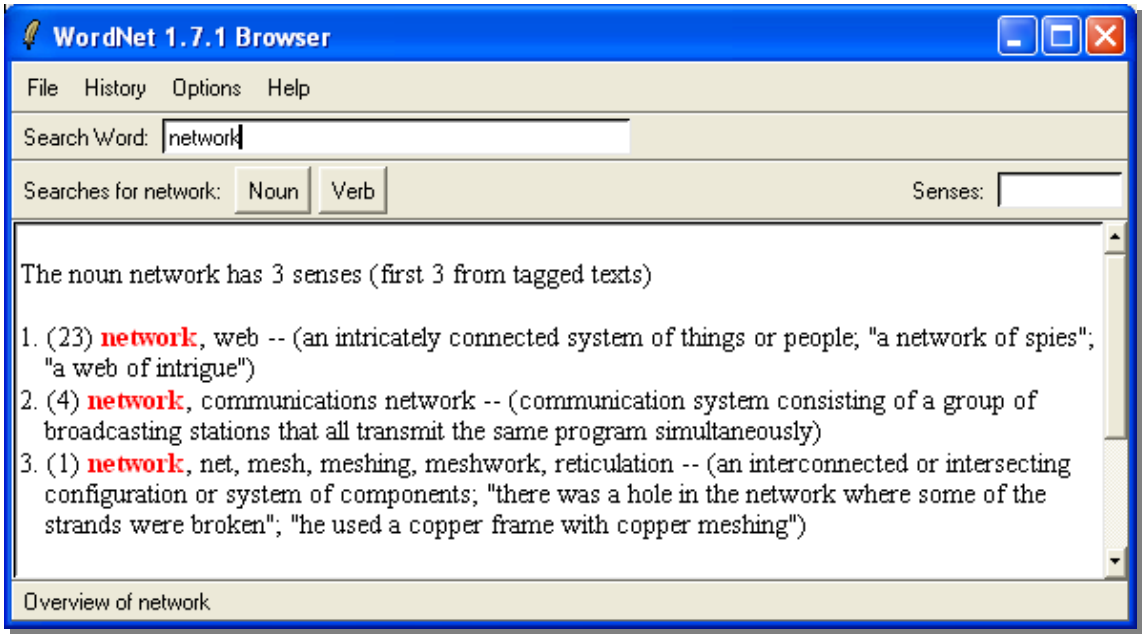

Figura 8 - Risultati ottenuti interrogando WordNet sulla parola "network"

La parola “computer”, in WordNet, presenta invece 2 sensi, come mostrato in Figura 9:

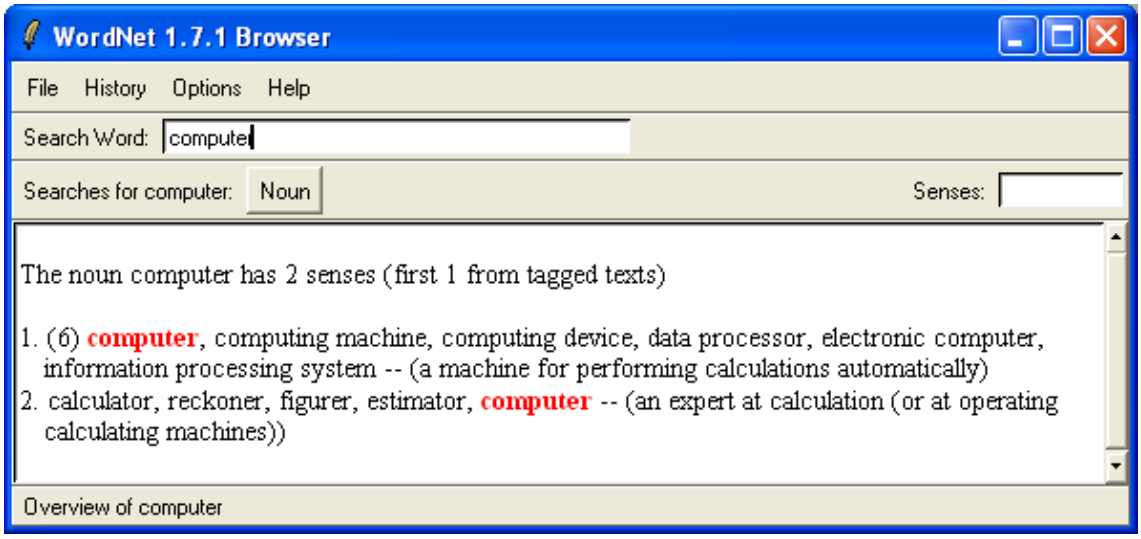

Figura 9 - Risultati ottenuti interrogando WordNet sulla parola "computer"

I sinonimi e i sostantivi estratti dai gloss per ciascun senso dei termini “network” e “computer” sono riportati in Tabella 1 e Tabella 2:

NETWORK		
Senso	**Sinonimi**	**Sostantivi estratti dal gloss**
1	web	system - things - people
2	communications network	communication - system - group - broadcasting - stations - program
3	net - mesh - meshing - meshwork - reticulation	configuration - system - components

Tabella 1 - Sinonimi e nomi estratti dai gloss della parola "network"

COMPUTER		
Senso	**Sinonimi**	**Sostantivi estratti dal gloss**
1	computing machine - computing device - data processor - elecronic computer - information processing system	machine - performing - calculations
2	calculator - reckoner - figurer - estimator	expert - calculation

Tabella 2 - Sinonimi e nomi estratti dai gloss della parola “computer”

L’algoritmo verifica inizialmente la presenza della coppia “network computer” nella base di dati:

```
Number of NOUNS found in WordNet dictionary: 2
"network" - "computer" not present into Database
```

Dal momento che la coppia non viene rilevata, si procede fissando per prima la parola “computer” e combinandola con i sinonimi e i sostantivi estratti dai gloss della parola “network” per ciascun senso di quest’ultima. In Figura 10 è mostrato il diagramma a barre contenente le query generate per il senso #1 della parola “network” ed i relativi risultati ottenuti da Google:

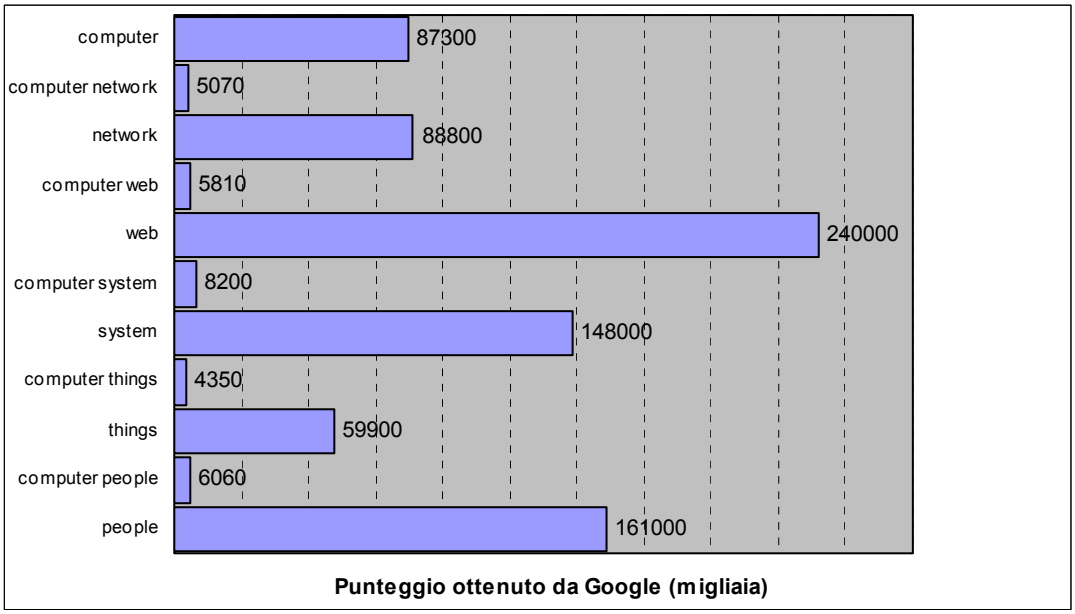

Figura 10 - Query generate per il senso #1 di "network" e numero di risultati ottenuti

In Figura 11 viene riportato il gloss relativo al senso #1 della parola "network" e la distanza di Tanimoto ottenuta:

```
GLOSS: an intricately connected system of things or people; "a
network of spies"; "a web of intrigue"

Tanimoto distance for sense [1] of "network" = 0.036536254
```

Figura 11 - Gloss relativo al senso #1 di "network" e distanza di Tanimoto ottenuta

In Figura 12 sono riportate le query generate per il senso #2 relativo alla parola "network" ed i punteggi ottenuti:

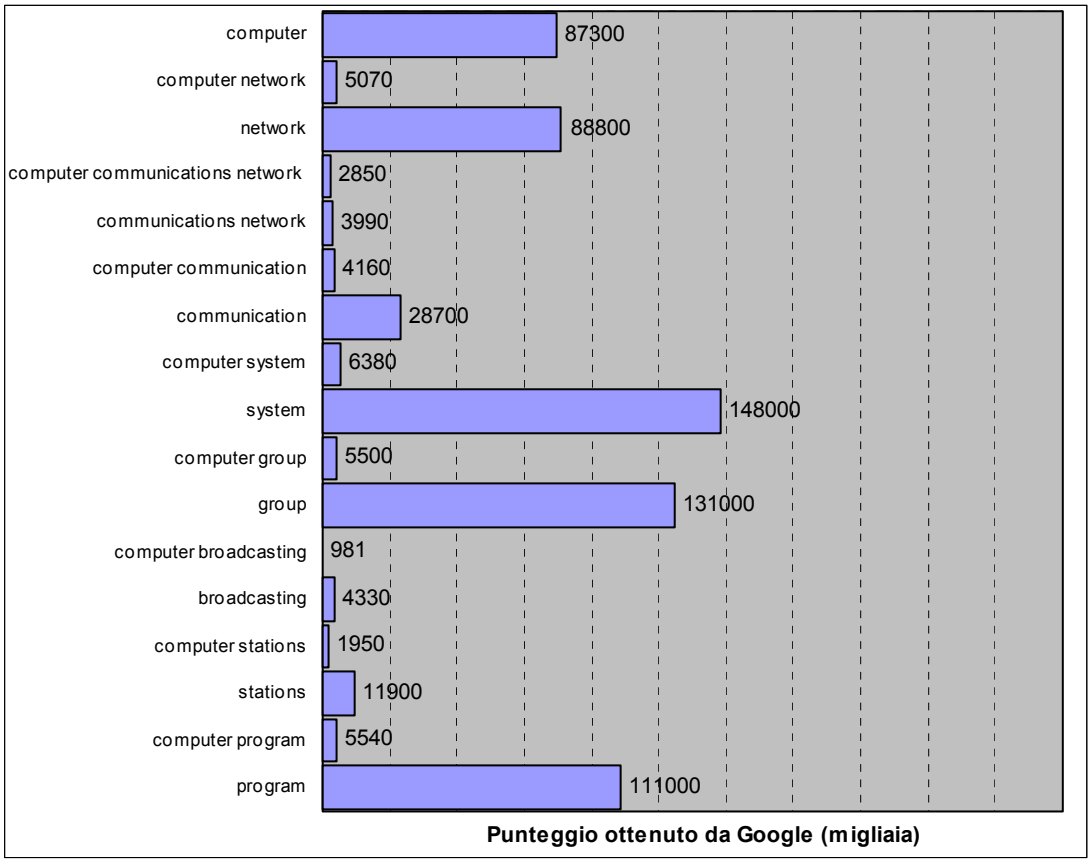

Figura 12 - Query generate per il senso #2 di "network" e numero di risultati ottenuti

Il gloss relativo al senso #2 della parola "network" e la distanza di Tanimoto ottenuta per tale senso sono riportati in Figura 13:

```
GLOSS: communication system consisting of a group of
broadcasting stations that all transmit the same program
simultaneously

Tanimoto distance for sense [2] of "network" = 0.055667024
```

Figura 13 - Gloss relativo al senso #2 di "network" e distanza di Tanimoto ottenuta

In Figura 14 vengono mostrate le query generate per il senso #3 di "network" ed i relativi risultati:

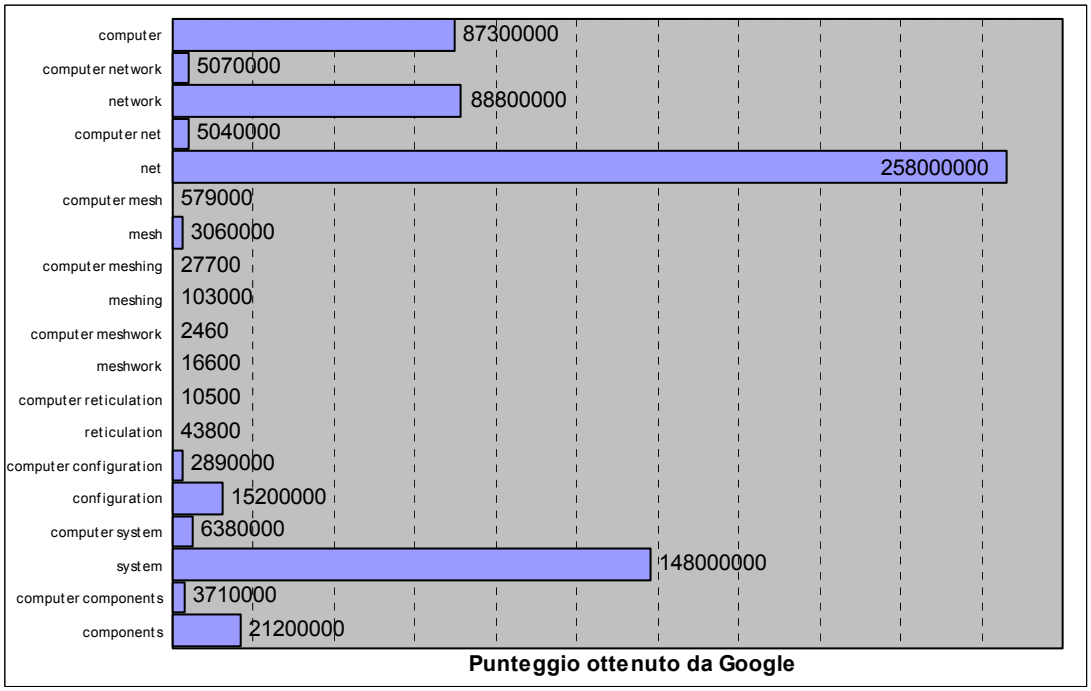

Figura 14 - Query generate per il senso #3 di "network" e numero di risultati ottenuti

Il gloss relativo al senso #3 della parola "network" e la distanza di Tanimoto ottenuta per tale senso sono riportati in Figura 15:

```
GLOSS: an interconnected or intersecting configuration or
system of components; "there was a hole in the network where
some of the strands were broken"; "he used a copper frame with
copper meshing"

Tanimoto distance for sense [3] of "network" = 0.03964735
```

Figura 15 - Gloss relativo al senso #3 di "network" e distanza di Tanimoto ottenuta

In seguito vengono ordinati i sensi della parola "network" in base alla distanza di Tanimoto ottenuta e viene calcolato l'indice di incertezza per il miglior senso, come mostrato in Figura 16:

```
"network" Sense ranking:
Sense: 2 - Tdist = 0.055667024
Sense: 3 - Tdist = 0.03964735
Sense: 1 - Tdist = 0.036536254
Current Uncertainty Index = 0.71222323
```

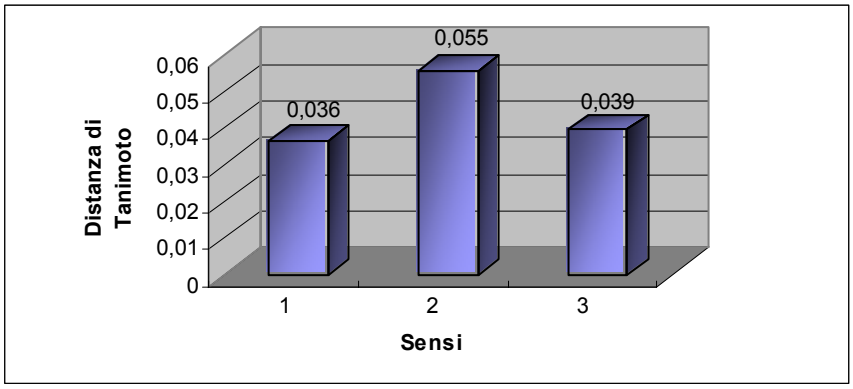

Figura 16 - Classifica dei sensi di "network"

Ottenuta la classifica dei sensi della parola "network", si procede alla disambiguazione della parola "computer", stavolta fissando "network" e combinandola con i sinonimi e i sostantivi estratti dai gloss di "computer". In Figura 17 si possono osservare le query generate per il senso #1 di "computer" ed i risultati riportati:

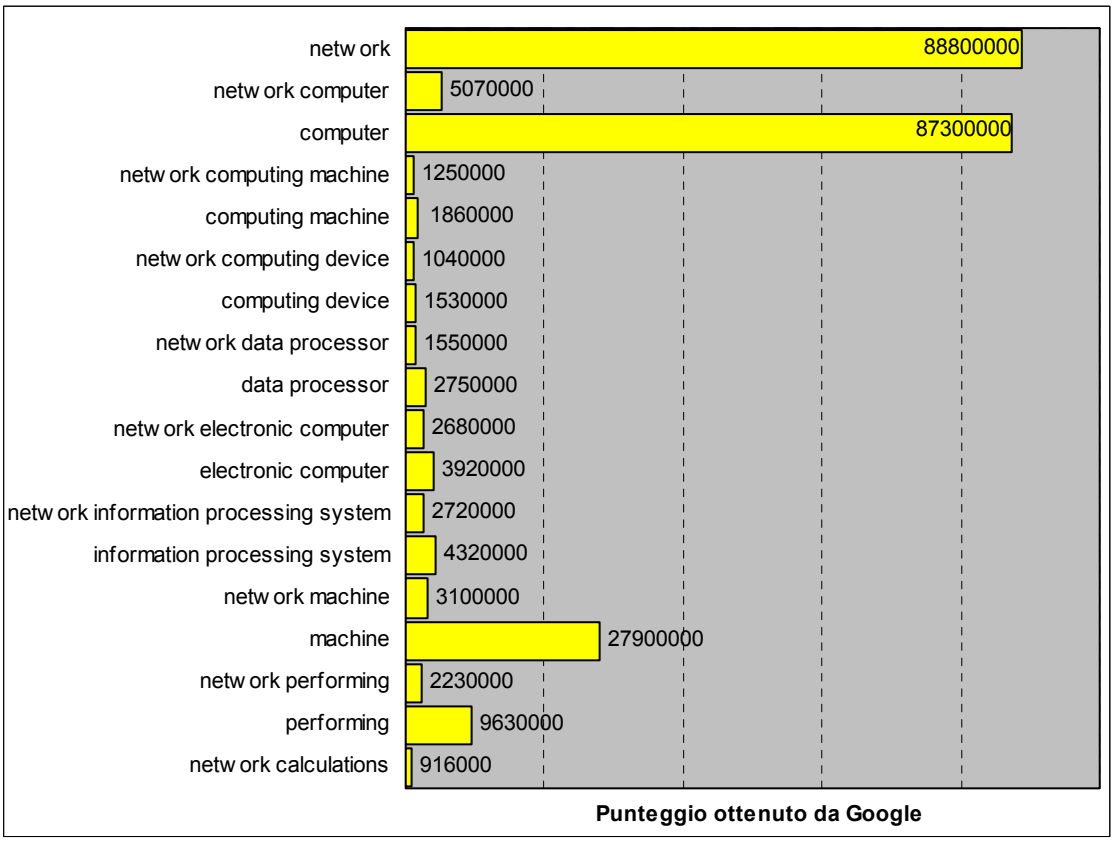

Figura 17 - Query generate per il senso #1 di "computer" e numero di risultati ottenuti

In Figura 18 viene riportato il gloss relativo al senso #1 della parola "computer" e la distanza di Tanimoto ottenuta:

```
GLOSS: a machine for performing calculations automatically

Tanimoto distance for sense [1] of "computer" = 0.09705199
```

Figura 18 - Gloss relativo al senso #1 di "computer" e distanza di Tanimoto ottenuta

In Figura 19 sono riportate le query generate per il senso #2 di "computer" ed i relativi punteggi ottenuti da Google:

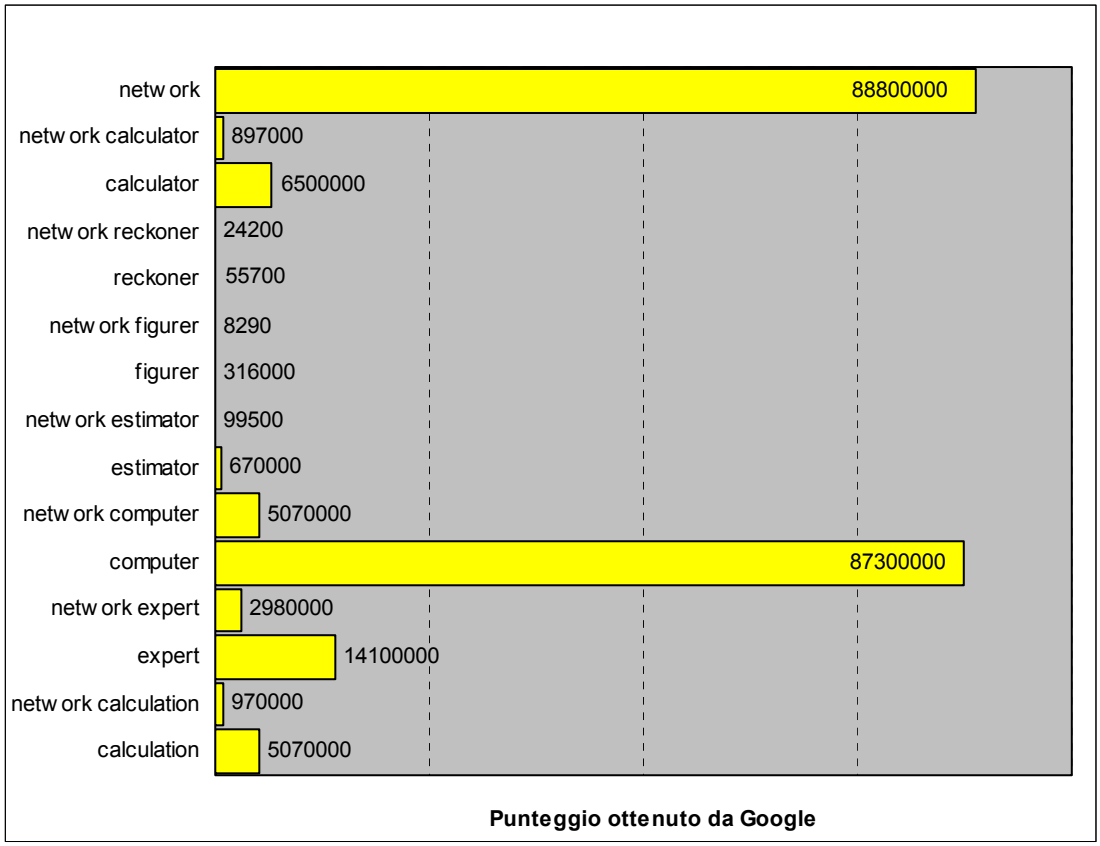

Figura 19 - Query generate per il senso #2 di "computer" e numero di risultati ottenuti

In Figura 20 si possono osservare il gloss relativo al senso #2 di "computer" e la distanza di Tanimoto ottenuta per tale senso:

```
GLOSS: an expert at calculation (or at operating calculating
machines)

Tanimoto distance for sense [2] of "computer" = 0.052131403
```

Figura 20 - Gloss relativo al senso #2 di "computer" e distanza di Tanimoto ottenuta

In Figura 21 è riportata la classifica ottenuta per i sensi della parola "computer" in funzione della distanza di Tanimoto:

```
"computer" Sense ranking:
Sense: 1 - Tdist = 0.09705199
Sense: 2 - Tdist = 0.052131403
Current Uncertainty Index = 0.53714925
```

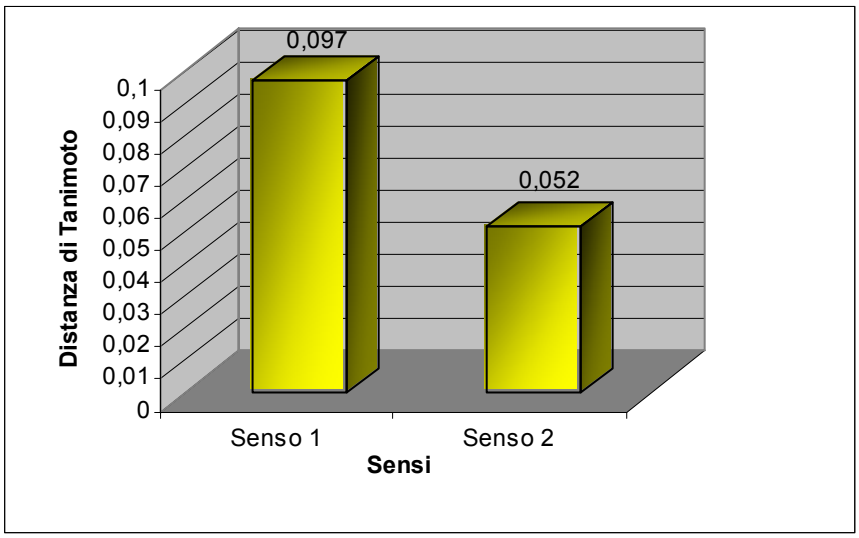

Figura 21 - Classifica dei sensi di "computer"

Avvenuta l'elaborazione dei dati, viene visualizzato nel browser ciascun termine con il relativo senso ottenuto, l'indice di incertezza relativo a tale senso, il gloss ad esso appartenente, l'estensione della query e l'interfaccia per il salvataggio dei risultati nella base di dati.

Se invece si accoppia la parola "rete" con il sostantivo "pesca", il quale a sua volta presenta due traduzioni, tra tutte le traduzioni possibili viene selezionata "net fishing", come mostrato in Figura 22:

```
matching "rete" in line: 4551
Available translations for the word "rete":
net
network
matching "pesca" in line: 3943
Available translations for the word "pesca":
fishing
peach

Combining translations...
net fishing
N results for: net fishing: 2060000
N results for: net OR fishing: 9730000
net peach
N results for: net peach: 353000
N results for: net OR peach: 3260000
network fishing
N results for: network fishing: 1260000
N results for: network OR fishing: 8700000
network peach
N results for: network peach: 154000
N results for: network OR peach: 3070000
```

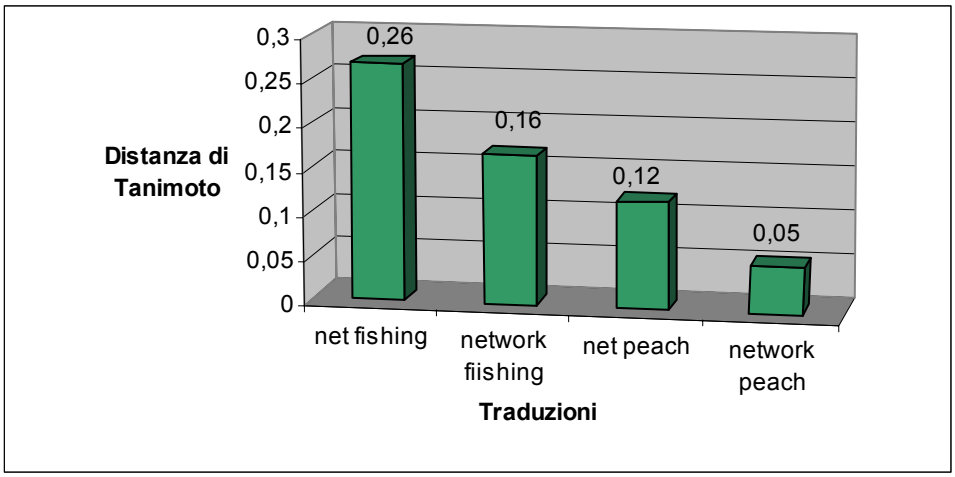

Figura 22 - Classifica delle traduzioni per la query "rete pesca"

La parola "net" presenta in WordNet 6 sensi, tra i quali figura quello inerente al contesto, cioè il #2 come mostrato in Figura 23:

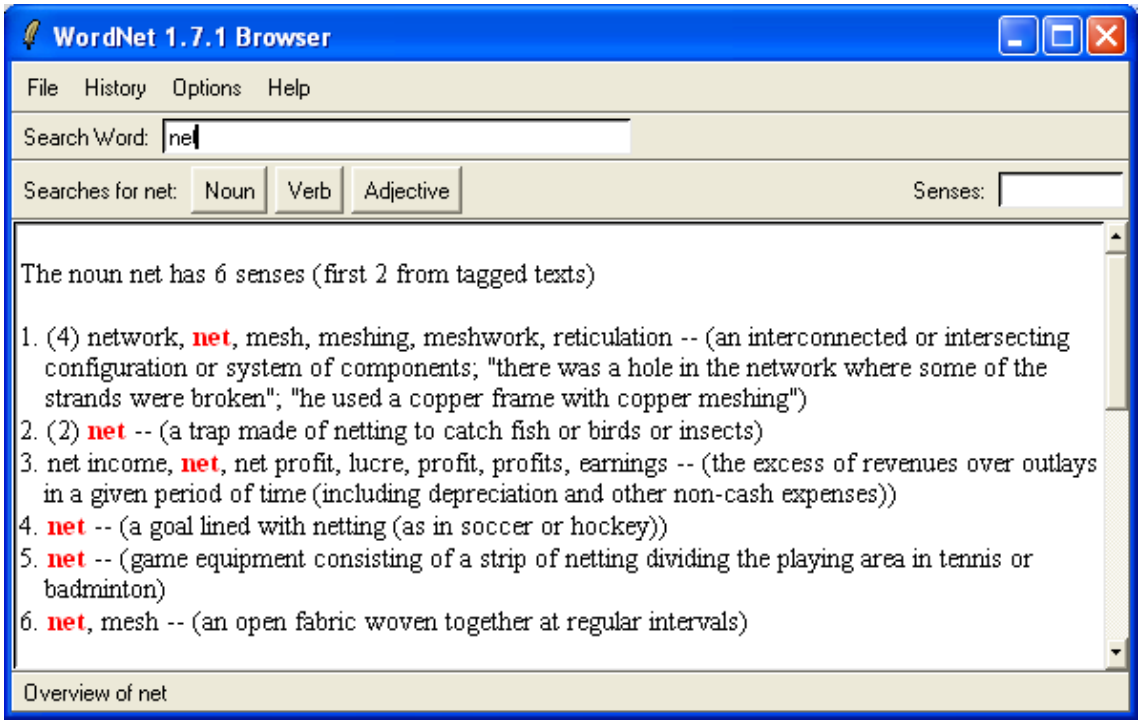

Figura 23 - Risultati ottenuti interrogando WordNet sulla parola "net"

La parola fishing invece presenta, come sostantivo, due sensi, come si può vedere in Figura 24:

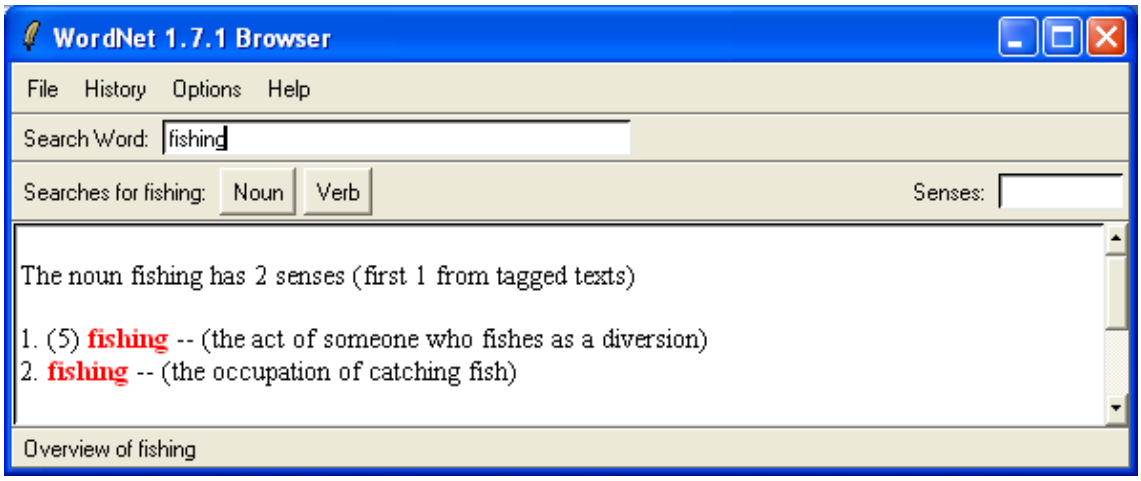

Figura 24 - Risultati ottenuti interrogando WordNet sulla parola “fishing”

Procedendo in maniera analoga a quanto visto per la query “rete computer”, per la query in questione si ottiene il senso inerente al contesto.

Capitolo 5
Conclusioni e prospettive future

A conclusione del lavoro svolto, si ritiene opportuno focalizzare l'attenzione sui punti nevralgici dello stesso, al fine di sottolinearne gli spunti di maggiore interesse e le possibili applicazioni future.

5.1 Considerazioni conclusive

Primo imprescindibile passo nello sviluppo dell'algoritmo è stata la creazione, di fatto realizzata in maniera totalmente automatica, di un dizionario di sostantivi. L'utilizzo di tale risorsa è da ritenersi necessario al fine di disporre delle traduzioni in inglese dei termini italiani costituenti le query da disambiguare. A partire da questa solida base, è stato possibile implementare un metodo atto all'analisi di tutte le plausibili traduzioni dei termini da disambiguare ed alla scelta di quella più opportuna. Tale scelta viene effettuata mediante l'ausilio del motore di ricerca Google che, corredato da un'utile collezione di API, consente la realizzazione di applicazioni in grado di interfacciarsi con il motore di ricerca.

La disambiguazione dei termini tradotti, da considerarsi l'operazione di maggiore interesse dell'intero algoritmo, consiste nella comprensione del significato di un sostantivo a partire dal contesto nel quale lo stesso è collocato. E' possibile effettuare tale disambiguazione mediante l'utilizzo di WordNet, base di dati lessicale in grado non solo di fornire le conoscenze proprie di un dizionario elettronico, ma anche di cogliere le sfumature che consentono di collocare un termine all'interno del contesto di appartenenza in maniera precisa.

L'ultimo rilevante aspetto dell'applicazione sviluppata consiste nella possibilità di estendere le query effettuate, sempre mediante l'utilizzo di WordNet, al fine di generarne altre equivalenti che tengano in considerazione relazioni semantiche, come la sinonimia, strettamente connesse al significato della parola. Un tale approfondimento nell'applicazione dell'algoritmo può essere quindi effettuato solo previa conoscenza del significato corretto del termine, fornito proprio dall'iniziale processo di disambiguazione.

In accordo con Resnik [31], a causa dei numerosi e differenziati approcci esistenti nell'ambito della WSD[4], risulta particolarmente oneroso un confronto tra gli stessi. Si intende comunque analizzare i risultati ottenuti in relazione a quelli propri del preesistente algoritmo di Moldovan [23], assumendo come termine di paragone la percentuale di disambiguazioni effettuate in maniera corretta.

---

[4] Si veda a tal proposito quanto riportato nel Capitolo 1.

Nonostante le sostanziali differenze fra le due soluzioni adottate – come l'utilizzo, da parte di Moldovan, di criteri di etichettatura dei vocaboli e di particolari operatori per la ricerca in Internet – è possibile effettuare un confronto sulla base dei punti di unione delle stesse, identificabili nell'applicazione della disambiguazione a coppie di vocaboli e nell'utilizzo di Internet come base statistica. Un'analisi comparata di questo tipo denota come, a fronte dell'86,5% di disambiguazioni effettuate in maniera corretta su sostantivi precedentemente etichettati dall'algoritmo di Moldovan, il presente lavoro produca un disambiguazione esatta nel 63% dei casi. Tale percentuale sale però al 70% se si prendono in considerazione esclusivamente le coppie di sostantivi correttamente tradotti.

5.2 Possibili applicazioni e prospettive future

L'ambito nel quale il presente progetto può rivelarsi particolarmente utile è, plausibilmente, quello concernente la ricerca in Internet. Infatti, proprio in tale ambito, l'applicazione delle tecniche di disambiguazione contribuisce in maniera determinante al miglioramento del recupero delle informazioni (Information Retrieval) ed al loro trattamento.

La prospettiva di utilizzo di maggiore interesse è costituita dall'opportunità di estendere il campo di applicazione della tecnica di disambiguazione a dizionari multilingua. Si intende con ciò la possibilità di reperire informazioni in lingua inglese senza alcuna imposizione di vincoli sulla lingua originaria di formulazione delle query.

L'utilizzo di WordNet fornisce, inoltre, le potenzialità per estendere le query disambiguate non solamente mediante le relazioni di sinonimia, ma anche mediante le numerose altre relazioni semantiche e lessicali precedentemente descritte.

Risulta infine di particolare interesse la prospettiva di effettuare la disambiguazione non soltanto nell'ambito di singole categorie linguistiche (sostantivi, verbi, etc.), ma anche di costrutti più complessi, quali ad esempio quelli tipici del linguaggio naturale.

Bibliografia

[1]. Y. A. Wilks and M. Stevenson, "*The grammar of sense: Is word sense tagging much more than part-of-speech tagging?*". Technical Report CS-96-05, University of Sheffield, Sheffield, United Kingdom, 1996.

[2]. AllWords.com – Dictionary, Guide, Community and More, http://www.allwords.com

[3]. I. Nancy and J. Véronis, "*Word sense disambiguation: The state of the art*". Computational Linguistics, 24:1, 1-40, 1998.

[4]. E. Agirre and D. Martinez, "*Knowledge Sources for Word Sense Disambiguation*". Lecture Notes in Computer Science, vol. 2166, 2001.

[5]. W. Weaver, "*Translation*". Mimeographed, 12 pp., July, 1949.

[6]. M. Masterman, "*Semantic message detection for machine translation, using an interlingua*". 1961 International Conference on Machine Translation of Languages and Applied Language Analysis, Her Majesty's Stationery Office, London, 1962, 437-475.

[7]. M. R. Quillian, "*Word concepts: A theory and simulation of some basic semantic capabilities*". Behavioral Science, 12, 410-30, 1967.

[8]. B. Boguraev, "*Automatic resolution of linguistic ambiguities*". Doctoral dissertation, Computer Laboratory, University of Cambridge, 1979 [available as technical report 11].

[9]. G. Adriaens, "*Word expert parsing: a natural language analysis program revised and applied to Dutch*". Proceedings of the 7th European Conference on Artificial Intelligence, ECAI'86, July 1986, Brighton, United Kingdom, 222-235.

[10]. M. Lesk, "*Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone*". Proceedings of the 1986 SIGDOC Conference, Toronto, Canada, June 1986, 24-26.

[11]. Y. A. Wilks and D. Fass, "*Preference semantics: A family history*". Report MCCS 90-194, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico, 1990.

[12]. D. Yarowsky, "*Word sense disambiguation using statistical models of Roget's categories trained on large corpora*". Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, 23-28 August, Nantes, France, 454-460, 1992.

[13]. P. Resnik, "*Disambiguating Noun Groupings with Respect to WordNet Senses*". Proceedings of the Third Workshop on Very Large Corpora, Cambridge, Massachusetts, 54-68, 1995.

[14]. P. Buitelaar, "*A lexicon for underspecified semantic tagging*". ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What and How?", April 4-5, 1997, Washington D.C., 25-33.

[15]. E. Viegas, K. Mahesh, and S. Nirenburg, "*Semantics in action*". In Saint-Dizier, Patrick (Ed.). Predicative Forms in Natural Language and Lexical Knowledge Bases. Text, Speech and Language Techonology Series. Kluwer Academic Publishers, Dordrecht.

[16]. E. Black, "*An Experiment in Computational Discrimination of English Word Senses*". IBM Journal of Research and Development, **32**(2), 185-194, 1988.

[17]. G. A. Miller, C. Leacock, R. Tengi and R. Bunker, "*A semantic concordance*". Proceedings of the 3rd DARPA Workshop on Human Language Technlogy, Plainsboro, New Jersey, March 1993, 303-308.

[18]. M. A. Hearst, "*Noun homograph disambiguation using local context in large corpora*". Proceedings of the 7th Annual Conf. of the University of Waterloo Centre for the New OED and Text Research, Oxford, United Kingdom, 1-19, 1991.

[19]. W. A. Gale, K. W. Church and D. Yarowsky, "*A method for disambiguating word senses in a large corpus*". Computers and the Humanities, 26, 415-439, 1993.

[20]. I. Dagan, A. Itai, and U. Schwall, "*Two languages are more informative than one*". Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 18-21 June 1991, Berkeley, California, 130-137.

[21]. B. Magnini, C. Strappavara, "*Experiments in Word Domain Disambiguation for Parallel Texts*". In proceedings of the ACL Workshop of Word Sense and Multilinguality, Hong Kong, China, 2000.

[22]. P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai and R. L. Mercer, "*Class-based n-gram models of natural language*". Computational Linguistics, Las Cruces, New Mexico, 139-145, 1992.

[23]. D. I. Moldovan and R. Mihalcea, "*Improving the search on the Internet by using WordNet and lexical operators*", 1999.

[24]. R. Mihalcea and D. I. Moldovan, "*An Iterative Approach to Word Sense Disambiguation*", 2000.

[25]. S. L. Lytinen, N. Tomuro and T. Repede, "*The Use of WordNet Sense Tagging in FAQ Finder*", 2000.

[26]. WordNet, a lexical database for English. Cognitive Science Laboratory, Princeton University, http://www.cogsci.princeton.edu/~wn/

[27]. G. A. Miller, R. Beckwith, C. Fellbaum, D.Gross and K. Miller, "*Five Papers on WordNet*". Technical report, Princeton University's Cognitive Science Laboratory, 1993.

[28]. G. A. Miller, "WordNet*: A Lexical Database for English*". In CACM 38, 1995.

[29]. Google, http://www.google.com

[30]. Google Web APIs (beta), http://www.google.com/apis

[31]. P. Resnik, "*Selectional preference and sense disambiguation*". In Proceedings of ACL Singlex Workshop on Tagging Text with Lexical Semantics, Why, What and How?, Washington DC, April 1997.

[32]. PostgreSQL, http://www.postgresql.org

[33]. M. Andreiana, "Web *Applications With Database Connectivity*". Linux Gazette (issue 50), 2000.

[34]. E. Brill, "*A Simple Rule-Based Part of Speech Tagger*". Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, 1992.

[35]. D. I. Moldovan and R. Mihalcea, "*A WordNet-based interface to Internet search engines*". In Proceedings of FLAIRS-98, Sanibel Island, FL, May 1998.

[36]. D.I Moldovan and R. Mihalcea, "*Using WordNet and Lexical Operators to Improve Internet Searches*". IEEE Internet Computing 4(1): 34-43, 2000.

[37]. R. Mihalcea and S. I. Mihalcea, "*Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web*". ICTAI, 280-287, 2001.

[38]. R. Mihalcea and D. I. Moldovan, "*A Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation*". International Journal on Artificial Intelligence Tools, vol.10 , 1-2, 2001.

[39]. CiteSeer, Scientific Literature Digital Library, http://citeseer.nj.nec.com/cs

[40]. Sourceforge, Open Source software development website, http://sourceforce.net

[41]. The Jakarta Site, Apache Tomcat, http://jakarta.apache.org