Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet

Bernardo Magnini
Carlo Strapparava
IRST, Istituto per la Ricerca Scientifica e Tecnologica
I - 38050 Povo TN, Italy
e-mail: {magnini | strappa}@irst.it

1. Motivazioni

L'idea di una base di conoscenza lessicale è stata recentemente proposta per indicare un insieme di conoscenze, principalmente di natura semantica, tra di loro strutturate in modo consistente e disponibili su supporto elettronico.

In questo lavoro descriviamo un progetto che ha come obiettivo la realizzazione di una base di conoscenza lessicale (LKB) di grandi dimensioni per la lingua italiana. Il progetto nasce dalla constatazione che attualmente non è disponibile una LKB per l'italiano, su supporto elettronico, tale che possa essere sfruttata per applicazioni sia nell'ambito dell'elaborazione del linguaggio naturale, sia della lessicografia. Per la realizzazione della LKB verranno inoltre usate tecniche sviluppate in Intelligenza Artificiale che permetteranno la acquisizione e la strutturazione automatica dell'informazione.

La principale fonte per l'acquisizione di informazione sul significato delle parole rimane tuttora un dizionario della lingua considerata. Tuttavia, l'utilizzo di dizionari che siano la semplice trasposizione elettronica di dizionari cartacei non è sufficiente. Infatti un semplice dizionario soffre spesso di mancanza di consistenza (e.g. circolarità delle definizioni) e di incompletezza delle informazioni semantiche relative ad un entrata lessicale, difetti che li rendono solo parzialmente sfruttabili per l'acquisizione della LKB. In questo progetto proponiamo una metodologia per la costruzione semiautomatica di una LKB per una lingua, in cui viene integrata la conoscenza descritta nel dizionario con la conoscenza, già strutturata, descritta in una LKB realizzata per la lingua inglese. La LKB di riferimento sarà WordNet [Miller 1990].

Il presente lavoro si situa all'interno di un più vasto programma di sviluppo di sistemi di dialogo in linguaggio naturale, che il gruppo linguaggio naturale dell'IRST sta portando avanti da alcuni anni [Stock et al. 1993]. Recentemente è emersa la necessità di potenziare la robustezza del sistema fornendo una interfaccia con grande copertura linguistica ed è in questa prospettiva che diviene necessaria la diponibilità di grosse porzioni di informazione sintattico/semantica, tramite le quali poter affrontare complessi problemi di interpretazione del testo.

2. WordNet

WordNet e' una base di conoscenza lessicale per l'inglese, disponibile gratuitamente, su supporto elettronico. In origine il progetto si e' ispirato alle correnti teorie psicolinguistiche sulla memoria lessicale umana. Nomi, verbi, aggettivi ed avverbi sono organizzati in insiemi di sinonimi, ciascuno dei quali rappresenta un concetto lessicale. Questi insiemi di sinonimi sono collegati tra di loro tramite un certo numero di relazioni ed organizzati in tassonomie. Nella attuale versione di WordNet sono presenti 95.600 forme lessicali organizzate in 70.100

significati (o synsets). Le corrispondenze tra forme lessicali e significati vengono mantenute tramite una matrice bidimensionale, nella quale ciascun synset è inteso essere un designatore non ambiguo del significato di una parola. Spesso (circa il 70%) ad un synset viene associata anche una breve definizione (gloss). WordNet distingue due tipi di relazioni: relazioni lessicali, quali la sinonimia, la antinomia e la polisemia, e relazioni concettuali, quali l'iponimia e la meronimia.

La relazione lessicale più importante per WordNet è la similarità di significato, dal momento che la capacità di riconoscere sinonimia tra parole è un prerequisito per la costruzione dei synsets e quindi per la rappresentazione dei significati nella matrice lessicale (vd. figura 1) .

Due espressioni sono sinonime se vale il principio di sostitutività (in altre parole se la sostituzione di una con l'altra non cambia il valore di verità di una frase). In realtà risulta più utile una definizione più debole, relativizzata ad un contesto. Due espressioni sono sinonime in un contesto linguistico C se la sostituzione di una con l'altra in C non cambia il valore di verità. È importante notare che la definizione di sinonimia in termini di sostitutività rende necessario partizionare WordNet in nomi, verbi, aggettivi e avverbi. Ovviamente l'appartenenza di una parola a più di un synset dà un'indicazione della sua polisemia.

La relazione di antinomia fornisce invece il principio organizzativo centrale per aggettivi ed avverbi.

WordNet non si limita a creare un insieme di significati relativi alle forme lessicali, ma indica anche le relazioni semantiche che sussistono tra di loro. L'ipo/iperonimia (o relazione ISA) mette in relazione significati subordinati e superordinati fornendo così una struttura gerarchica di concetti. La relazione meronimica (HAS-PART) induce invece sull'insieme dei significati una gerarchia delle parti.

3. La matrice lessicale multi linguale

Il punto di partenza del progetto di costruzione di una rete tipo WordNet multilingua si fonda sull'ipotesi che la rete dei significati (synsets) attualmente definita per la versione inglese possa essere in gran parte riutilizzata per altri linguaggi. Quest'ipotesi può essere considerata plausibile se ci limitiamo alle principali lingue indoeuropee tra le quali si può trovare una larga sovrapposizione culturale [Miller--comunicazione personale].

Il progetto prevede la realizzazione di una matrice lessicale multi linguale (MLLM) come estensione della matrice lessicale bidimensionale attualmente implementata in WordNet. Verrà aggiunta una terza dimensione alla matrice sulla quale sarà possibile considerare diverse lingue. L'estensione nella dimensione dei linguaggi verrà inizialmente considerata per l'italiano. La figura 1 visualizza le tre dimensioni della matrice (parole di una lingua, significati e linguaggi), insieme alle principali relazioni lessicali e semantiche. Per realizzare la matrice multilinguale in linea di principio occorre ri-mappare le forme lessicali italiane con i significati corrispondenti (M_i), costruendo l'insieme dei synsets per l'italiano (esplicitando gli E_{ij}^{I}). Il risultato sarà una completa ridefinizione delle relazioni lessicali, mentre per le relazioni semantiche verranno sfruttate, per quanto possibile, quelle già definite originariamente per l'inglese. Da questo punto di vista la dimensione dei significati viene considerata costante rispetto alle lingue e alle parole di ogni lingua. Se per un certo M_k si ottiene $E_{ik}^{L} = \{0,\ldots,0\}$ significa che per il linguaggio L non esiste nessuna parola che realizza lessicalmente quel significato.

4. Costruzione automatica della LKB

Le procedure di estrazione automatica dell'informazione su cui intendiamo basarci sono le seguenti:

a) estrazione dal dizionario macchina italiano

Per questo verrà usato un dizionario in formato elettronico, da cui estrarre informazione semantica (contenuta come testo). In prima approssimazione possiamo pensare alla estrazione del "genus" della definizione, tramite il quale verrà ristretto il campo di ricerca nella rete di significati di WordNet. Inoltre considereremo l'applicazione di tecniche di estrazione per ottenere informazione semantica strutturata (es. frames).

b) estrazione dalle "short gloss"

Un procedimento analogo a quello applicato alle definizioni del dizionario verrà usato sulle definizioni associate ai synset inglesi (short gloss). Anche in questo caso il risultato sarà una struttura a frames. Questo compito è semplificato dal fatto che le short gloss usano un vocabolario limitato (circa 7000 radici, esclusi i nomi propri) e costruzioni sintattiche semplificate.

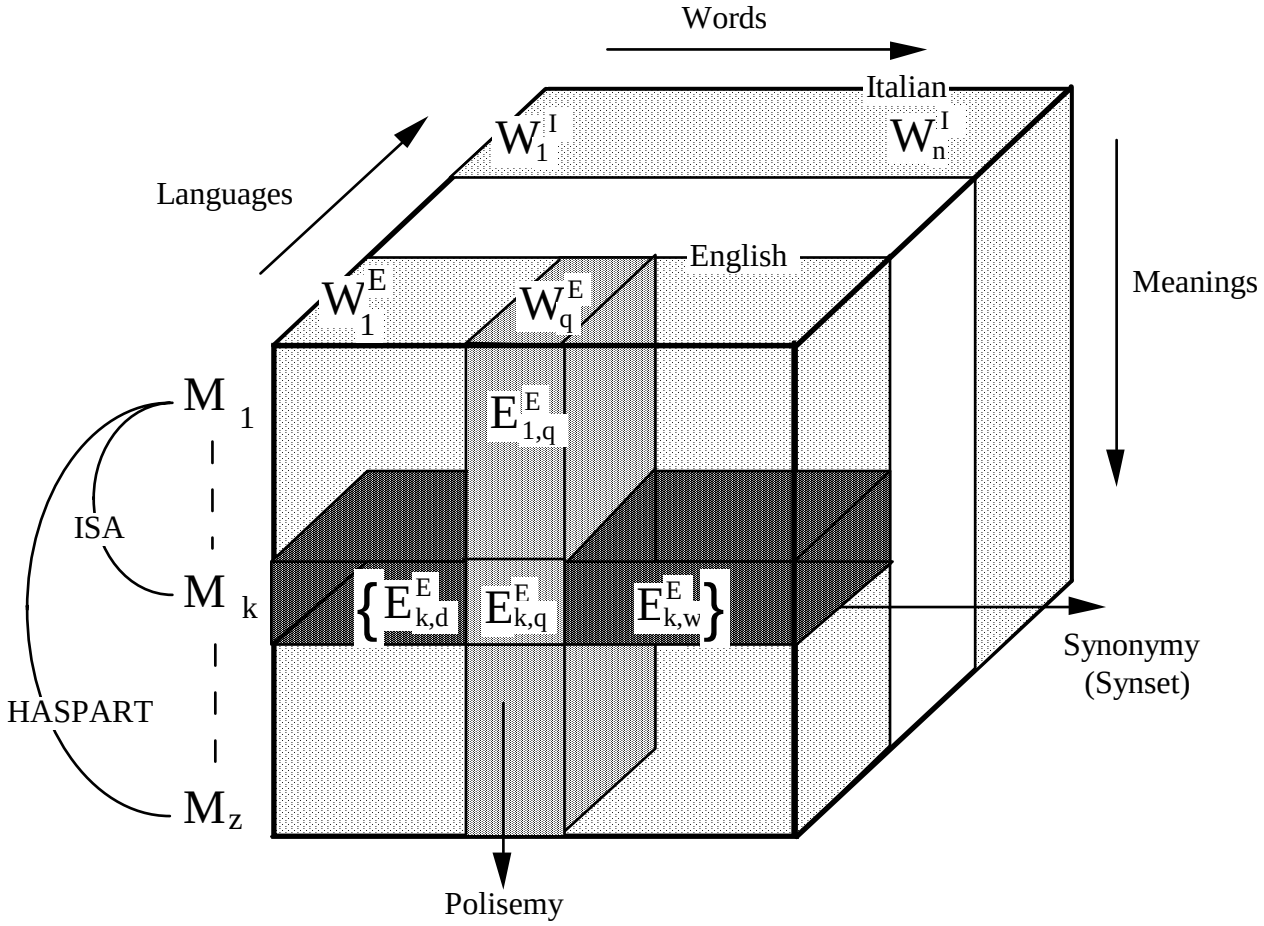

Figura 1: Multi Language Lexical Matrix

c) algoritmi di matching

La situazione che dovrà essere trattata a questo punto è quella di scegliere tra l'insieme di frames ricavati da WordNet quelli che soddisfano meglio un matching con il singolo frame ricavato dall'italiano. Questi algoritmi dovranno quindi confrontare e stabilire un grado di associazione tra due frames, rispettivamente contenenti informazioni ricavate dall'italiano e

dall'inglese. Il matching farà quindi uso di un dizionario bilingue, tramite il quale dovrà essere verificata la corrispondenza tra due termini nelle due lingue. In questo modo, all'interno di un singolo match, il problema della traduzione viene semplificato. La fase di automazione produrrà un insieme di possibili agganci tra una parola italiana e significati nella rete WordNet per i quali l'algoritmo ha superato una soglia prefissata. È importante comunque l'intervento di un lessicografo per validare le scelte proposte.

Bibliografia

[Miller 1990] Miller, G. A. (ed.), WordNet: "An on-line lexical database". *International Journal of Lexicography* (special issue), 3 (4), 235-312, 1990.

[Stock et.al. 1993] Stock, O. and the AlFresco Project Team, "AlFresco: enjoing the combination of NLP and hypermedia for information exploration", in M.T. Maybury (ed.), *Intelligent Multimedia Interfaces*, AAAI and MIT Press, 1993