Strumenti digitali e collaborativi per le Scienze dell'Antichità
a cura di Paolo Mastandrea

Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'Information Retrieval

Stefano Minozzi
(Università degli Studi di Verona, Italia)

Abstract This paper describes the creation of Latin WordNet, addressing the methods of construction, the data structure, possible applications and further developments. More notably, a possible expansion of Latin WordNet is the construction of an additional data structure in order to improve the chances of the network to represent the semantic shift taking diachrony into account. The original Princeton WordNet, on whose ground Latin WordNet is developed, does not take into account the semantic shift of words in time, considering the meanings identified by synsets as if existing in a model of vocabulary where a language has not an history: the purpose of Princeton WordNet was obviously the representation of in-use contemporary English, but this model is weak when we come to describe the semantic structures of a finished language such as Latin. In the second part, this paper shows the connections between semantic networks and information retrieval moving from the problems pertaining to the automatic disambiguation of contexts and the exploitation of ontologies in the process of sense retrieval on semantically annotated corpora.

Keywords WordNet. Ontologies. Lexicography. Dictionaries. Semantic network. Information retrieval.

1 La costruzione della base di conoscenza lessicale LatinWordNet

Il progetto di costruzione di una base di conoscenza lessicale per la lingua latina nacque, alla fine del 2004, con l'intento di fornire un modello di rappresentazione elettronica della conoscenza semantica utile alla applicazione di tecniche di *Natural Language Processing* per quella lingua.

Uno degli esempi più completi di rete semantica, all'epoca, era costituito da WordNet (Miller et al. 1990; Fellbaum 1998), un sistema disponibile pubblicamente, che contiene *frame* specificamente orientati alla rappresentazione delle parole: a partire dal riconoscimento della natura del tutto accidentale dell'ordinamento dei dizionari attraverso spelling, nel modello

Antichistica 14
DOI 10.14277/6969-182-9/ANT-14-10 | Submission 2017-08-14
ISBN [ebook] 978-88-6969-182-9 | ISBN [print] 978-88-6969-183-6

di WordNet le parole sono organizzate per blocchi di significato, denominati 'synsets', che raccolgono tutti i lemmi che lessicalizzano lo stesso concetto; i synsets sono collegati tra loro per mezzo di relazioni che includono, assieme alla sinonimia, anche l'iponimia, la meronimia e l'antinomia. L'ipo-iperonimia mette in relazione significati subordinati e superordinati fornendo così una struttura gerarchica di concetti. La relazione meronimica induce una gerarchia delle parti sull'insieme dei significati. In questo modo il livello lessicale è chiaramente separato da quello concettuale e questa distinzione è rappresentata dal medium semantico-concettuale e dalla relazione semantica che uniscono rispettivamente synsets e parole. Le relazioni presenti tra i verbi permettono di mettere in luce relazioni di implicazione (ingl. 'entailment') e di troponimia. Due verbi sono correlati dall'implicazione nel momento in cui il primo verbo implichi il secondo: per esempio la coppia comprare-pagare. La troponimia è la relazione presente nel momento in cui due attività collegate da implicazione avvengono allo stesso tempo: un esempio è la coppia zoppicare-camminare.

Il progetto di costruzione di una rete semantica per il latino ebbe come basi di partenza due modelli: WordNet e MultiWordNet (Bentivogli, Girardi, Pianta 2002), progetto sviluppato dall'allora Istituto Trentino di Cultura (oggi Fondazione Bruno Kessler) inteso a realizzare una rete semantica multilingue. Il modello adottato dal progetto MultiWordNet (MWN) consisteva nel costruire le reti semantiche specifiche per un linguaggio mantenendo il più possibile le relazioni semantiche disponibili nella WordNet di Princeton (PWN). Ciò è ottenibile costruendo i nuovi synsets in corrispondenza dei synsets della PWN, ogni volta che ciò sia fattibile, e importando le relazioni semantiche dai corrispondenti synsets inglesi; in questo modo si ipotizza che, se esistono due synsets nella PWN e una relazione che li collega, la stessa relazione leghi i corrispondenti synsets in una lingua diversa. Secondo Vossen (1996), il modello di MWN (o 'modello a espansione', *expand model*) garantisce un elevato grado di compatibilità tra differenti *wordnet*. Per constatare questo fatto basta considerare che la costruzione di qualsiasi rete semantica necessariamente implica un gran numero di decisioni soggettive (e discutibili). Così se due reti semantiche sono costruite indipendentemente per due diverse lingue, mostreranno differenze che dipendono solo parzialmente dalle differenze tra le due lingue: alcune non banali discrepanze strutturali dipenderanno, infatti, da scelte soggettive o da criteri di costruzione differenti. Il modello di MWN minimizza queste differenze aderendo strettamente ai modelli di costruzione di PWN.

Il modello MWN presentava anche degli inconvenienti potenziali: il rischio più serio era quello di forzare un'eccessiva dipendenza sulla struttura lessicale e concettuale di uno dei linguaggi coinvolti (Vossen 1996). Questo rischio però poteva essere scongiurato permettendo alla nuova rete semantica di divergere, quando necessario, dalla struttura di PWN.

Un altro importante vantaggio del modello MWN era la possibilità di utilizzare procedure automatiche per velocizzare la costruzione dei synsets corrispondenti e per l'individuazione delle divergenze tra PWN e la rete semantica che si stava costruendo. In tutte queste procedure la stessa PWN poteva essere usata utilmente come risorsa.

La costruzione di LWN (LatinWordNet) si è pertanto basata in un primo tempo su una procedura automatica per l'assegnazione. Seguendo il modello MWN, il nostro obiettivo era quello di costruire, ogniqualvolta fosse possibile, un synset latino sinonimo (semanticamente corrispondente) di un synset di PWN. Se ciò non era possibile, veniva individuata un'idiosincrasia Inglese-a-Latino o Latino-a-Inglese.[1]

I synsets sinonimi latini sono stati costruiti seguendo tre differenti strategie:

- La prima strategia era basata sui traducenti dall'inglese al latino. Per ciascun synset di PWN S, cerchiamo un gruppo di traducenti che siano i sinonimi delle parole inglesi di S. Se non è possibile costruire alcun synset sinonimo latino di S, si è trovata un'idiosincrasia lessicale inglese-a-latino.
- La seconda strategia era basata sui gruppi di traducenti Latino-a-Inglese. Per ciascun senso σ di una parola latina L, si cerca un synset di PWN che includa almeno un traducente inglese di L e si costituisce un legame tra L e S. Quando la procedura è stata applicata a tutti i significati della parola latina, possiamo costruire la classe di equivalenza di tutti i gruppi di parole latine che sono state collegate con lo stesso synset di PWN. Ciascun gruppo nella classe di equivalenza è il synset latino sinonimo con alcuni synsets di PWN. Se per un gruppo di sinonimi latini non c'è alcun synset sinonimo in PWN, si è trovata un'idiosincrasia lessicale Latino-a-Inglese.
- La terza strategia sfruttava la natura multilingue di MWN e i due dizionari di macchina latino-inglese e latino-italiano: attraverso di essa le parole latine che risultano avere come traduzione parole inglesi e parole italiane che sono contrassegnate dallo stesso identificativo di *synset* vengono attribuite al medesimo *synset* della rete LatinWordNet con lo stesso identificativo. Esse infatti presentano con certezza la lessicalizzazione latina del concetto espresso dai traducenti nelle due lingue moderne.[2]

1 Una trattazione estesa relativa al problema dei *lexical gap* si può trovare in Bentivogli et al. 2000.

2 In altre parole una parola, data una parola latina, sarà attribuita a quei *synset* che costituiscono l'intersezione dei gruppi di *synset* individuati dai rispettivi gruppi di traducenti: vale a dire quei *synset* ai quali rimandano sia i traducenti italiani, sia i traducenti latini.

Il miglior allineamento tra la WordNet di Princeton e quella latina è stato ottenuto utilizzando entrambe le strategie per cercare di validare i risultati incrociandoli.

Trovare collegamenti tra i significati delle parole latine e i synsets di PWN è un processo complesso e lungo, anche se è sempre molto più rapido rispetto alla costruzione da zero dei synsets latini, della loro organizzazione in una rete semantica e del metterli in corrispondenza con i synsets di PWN. Per ciascun significato latino, il lessicografo dovrebbe cercare i gruppi di traducenti equivalenti in un dizionario bilingue, trovare tutti i synsets che contengono questi traducenti equivalenti, valutare con attenzione il significato di questi synsets (sinonimi, glosse, relazioni semantiche) e, infine, decidere quale *synset* di PWN, se esiste, è sinonimo del significato latino della parola. Per alcuni significati di parola il lessicografo potrebbe dover valutare decine di synsets di PWN.

Per aiutare il lessicografo nel suo lavoro era stata realizzata una procedura che sceglieva, per ciascun significato di una parola latina, i synsets di PWN dal significato compatibile. Nella maggior parte dei casi la procedura trovava una rosa ristretta di candidati, aiutando il lessicografo a focalizzare quale fosse il synset di PWN più per l'assegnazione.

La procedura-assegnazione prendeva come input uno dei sensi della sezione Latino-a-Inglese del dizionario di macchina e forniva in output un gruppo di candidati, ciascuno dei quali era descritto da un *punteggio di certezza* e da un *synset* di PWN, dove il punteggio di certezza (PC) misurava il grado di certezza nel legame tra il significato della parola latina e il synset di PWN. Solo i candidati con un PC più alto di una certa soglia venivano proposti al lessicografo. Scegliere il livello di soglia è stata una questione di bilanciare precisione e richiamo. Maggiore era la soglia, minore era la probabilità che candidati erronei fossero proposti (alta precisione), ma era anche maggiore la possibilità che la scelta più idonea non fosse inclusa nel gruppo dei candidati (basso richiamo).

Per un determinato significato di parola listato nel dizionario latino-inglese, la procedura-assegnazione considerava il gruppo di parole inglesi che venivano proposte come traducenti equivalenti per quel significato e trovava tutti i synsets contenenti almeno un traducente equivalente. Questi *synsets* costituivano il gruppo di candidati (GCand) che doveva essere collegato con il significato di parola latina dell'input. Possiamo riassumere il primo passo dell'algoritmo dicendo che esso calcolava i GCand del significato di una determinata parola latina. Il resto dell'algoritmo consisteva nell'ordinare i GCand calcolando il PC di ciascuno dei synsets.

L'ordinamento dei GCand era basato su una serie di regole per stabilire i legami: ogni regola, se applicata con successo a un candidato, alzava il suo PC. Si deve notare che il PC parziale, contribuito da ciascuna regola, variava a seconda di fattori specifici alla regola. Accanto al dizionario di macchina, venivano utilizzate dalle regole anche altre risorse, come la se-

zione italiana di MultiWordNet e un dizionario italiano-latino, un dizionario dei sinonimi latini e la stessa PWN.

Le strategie individuate per la costruzione dei legami erano quattro:

- probabilità generica: la regola di probabilità generica si basa sulla supposizione che solo un elemento nel GCand è il corretto candidato per legare il senso di una parola latina. Di conseguenza si può supporre che maggiore è la cardinalità del GCand, minore è la probabilità che ciascun candidato sia quello esatto. La cardinalità del GCand dipende dal grado di ambiguità delle parole che sono proposte come traducenti equivalenti del significato della parola di input. Se c'è un solo synset nel GCand, ciò significa che tutti i traducenti equivalenti della parola di input sono monosemici: è quindi altamente probabile che l'unico synset nel GCand sia sinonimo del significato della parola di input.[3]
- traduzione incrociata: questa regola si basa sulla supposizione che se colleghiamo un significato di parola al corretto synset attraverso un traducente equivalente, è probabile che almeno alcuni dei sinonimi del traducente, presenti in PWN, abbiano la parola di input come traducente equivalente inglese-latino. Si prenda ad esempio il latino 'punctum': quando riferito a insetti, si traduce come 'sting'. 'Sting', però, appartiene a 4 synsets di PWN: 'sting', 'stinging'; 'pang', 'sting'; 'sting', 'bite', 'insect bite'; 'bunco', 'bunco game', 'sting'. Solo il terzo synset è sinonimo della parola latina. Se guardiamo ai sinonimi di 'sting' nel terzo synset possiamo trovare che la sezione inglese-latino dà 'punctum' come traduzione di 'bite'. Riassumendo, la regola della traduzione incrociata considera i sinonimi presenti in PWN di un traducente che crea il collegamento e calcola un PC parziale che è proporzionale al numero di sinonimi che hanno la parola italiana come traducente dall'inglese al latino.
- corrispondenza della glossa: un gruppo di regole di collegamento sfrutta le informazioni contenute nella glossa inglese che introduce la maggior parte del dizionario di macchina inglese-latino. La glossa può contenere un campo semantico specifico, un sinonimo, un iperonimo, o una specificazione di contesto d'uso. Queste informazioni possono essere utilizzate in vario modo.

L'informazione relativa al campo semantico è sfruttata grazie a una risorsa sviluppata parallelamente a MWN, cioè la marcatura di tutti i synsets di PWN con una etichetta relativa al campo semantico (Magnini, Cavaglià 2000). La glossa del dizionario infatti contiene una etichetta relativa al

3 Cf. il criterio monosemico usato in Atserias et al. 1997.

campo semantico e se questa etichetta corrisponde a un synset individuato come candidato, allora il candidato ottiene un maggiore PC. Le varianti nelle etichette dei campi semantici sono gestite attraverso una tabella di corrispondenze.

Quando le glosse contengono parole o frasi, si cerca un corrispondente tra di esse e le parole contenute nelle glosse di PWN. Per fare ciò, si estraggono i lemmi delle parole inglesi delle glosse, e si controlla la loro presenza nelle glosse del traducente equivalente in PWN. La forza della corrispondenza dipende dal grado di ambiguità del traducente. Maggiore è la polisemia, minore è il peso attribuito alla corrispondenza.

Il meccanismo ha due estensioni basate sul fatto che le glosse spesso specificano il genere della parola che stanno definendo al posto di un sinonimo. La prima estensione cerca una corrispondenza tra una parola latina e un iperonimo del suo traducente equivalente. Il secondo meccanismo cerca una corrispondenza tra una parola latina e una parola inglese contenuta nella glossa di un iperonimo del synset candidato. Se la corrispondenza tra la parola latina e la parola inglese viene ottenuta attraverso uno dei meccanismi indiretti il PC parziale sarà più basso rispetto all'individuazione diretta.

La regola dell'intersezione di synsets sfrutta il fatto che i gruppi di traduzione possono includere più traducenti equivalenti, che sono ovviamente sinonimi. Se uno dei traducenti equivalenti è ambiguo, possiamo usare gli altri traducenti equivalenti per disambiguare. In pratica, la regola prende i differenti gruppi di candidati che sono accessibili attraverso diversi traducenti equivalenti e li interseca. I synsets che sono nell'intersezione ottengono un PC. Per esempio la parola latina 'pila' è tradotta nel suo senso metaforico come 'pillar', 'mainstay'. La parola 'pillar' appartiene a 5 synsets di PWN, mentre 'mainstay' appartiene a tre synsets. C'è però un solo synset che li contiene entrambi.

Una volta terminata l'assegnazione si è ottenuta una struttura reticolare che forniva una modelizzazione dei rapporti semantici tra le parole: il modello di stoccaggio dei dati in *MultiWordNet* rifletteva i principali elementi teorici della rete semantica multilingue. Il database era costruito sull'idea che esista un gruppo di dati comuni a tutte le lingue e altri specifici di ciascuna lingua. Nell'implementazione le relazioni semantiche di PWN sono contenute in un modulo chiamato COMMON-DB, mentre le relazioni lessicali per il latino e per l'inglese sono conservate in altri due moduli LATIN-DB e ENGLISH-DB. In altre parole l'informazione relativa a quali lemmi appartengano ai synsets si trova nei database delle lingue, mentre l'informazione relativa alle relazioni tra i *synsets*, che rimangono costanti tra le lingue, è immagazzinata nel COMMON-DB. La corrispondenza tra i synsets realizzati nelle diverse lingue si ottiene utilizzando sempre lo stesso codice identificatore: i *synsets* di lingue diverse che hanno lo stesso codice di identificazione appartengono al medesimo multisynset.

Il COMMON-DB descrive le relazioni tra i *multisynsets* di MWN. Quindi, tutte le informazioni semantiche che sono indipendenti dalla lingua possono essere aggiunte al COMMON-DB.[4]

Il modello di dati di MWN rappresenta le costanti concettuali presenti in lingue differenti. Tale modello di dati, inoltre, evidenzia anche le divergenze semantiche tra le lingue.[5] Inoltre, anche se si mantengono le relazioni semantiche evidenziate da PWN come base del COMMON-DB, è possibile aggiungere nuove relazioni o modificare quelle esistenti. La possibilità di modificare le relazioni semantiche di PWN e di rappresentare le idiosincrasie concettuali nei linguaggi specifici è stata implementata attraverso dei moduli aggiuntivi che sovrascrivono, senza modificarli fisicamente, i dati originali di PWN. Il COMMON-DB infatti contiene tutte le relazioni semantiche originali di PWN e una risorsa chiamata COMMON-ADD-ON che ne riscrive una parte. Ciascuna lingua contiene un language-ADD-ON che specifica le relazioni semantiche che sono proprie di quella lingua.

Le peculiarità lessicali vengono codificate all'interno delle aggiunte specifiche di ciascuna lingua. Se c'è prova che la lessicalizzazione di un determinato concetto manchi in una lingua, nella sezione lessicale del database di quella lingua viene inserita un'etichetta vuota per quel nodo.[6] Per la rappresentazione delle differenze denotative e dei gap lessicali, vengono seguite due diverse strategie: se il nodo vuoto corrisponde a una differenza denotativa, una o più relazioni vicine vengono usate per collegare il nodo ad un synset più generico o a molti *synsets* più specifici. Se il nodo vuoto corrisponde a un gap lessicale, viene riportata nella glossa del nodo vuoto una parafrasi di traduzione appropriata, preceduta dalla parola chiave TE (Translating Equivalent). Le relazioni più vicine vengono inserite nella risorsa linguistica aggiuntiva specifica della lingua in questione.

Ciascun database linguistico contiene anche un modulo con informazioni lessicografiche relative ai collegamenti tra i sensi delle parole e i synsets.

Per quel che riguarda le relazioni, tutte quelle semantiche erano state importate da PWN e sono disponibili assieme alle relazioni più vicine, cioè le nuove relazioni specifiche di ciascuna lingua che sono state aggiunte nella MWN per rappresentare le differenze denotative.

L'attuale implementazione della parte latina di MWN si basa sull'aggiunta di un modulo (non disponibile online) in grado di rendere indipendente il livello grafico/ortografico dall'individuazione dei lemmi. In pratica per ciascun lemma di dizionario è stata introdotta una grafia normalizzata,

4 In particolare le relazioni relative ai campi semantici.

5 Nella fattispecie i gap lessicali.

6 Il termine 'nodo' è usato in quanto MWN si compone di una struttura reticolare, dove i lemmi sono inseriti come nodi e le relazioni semantiche costituiscono i collegamenti tra i nodi.

associata a un numero espandibile di grafie alternative. Nei synsets della parte latina non vengono registrati direttamente i lemmi, ma dei codici identificativi: in questo modo possono essere utilizzate diverse grafie per la rappresentazione dello stesso lemma, e sono inoltre collegate all'interno del synset anche tutte le realizzazioni morfologiche della flessione dei lemmi. L'implementazione della base dati è stata effettuata attraverso un database relazionale, in modo da permettere l'interfacciamento con il sistema di IR in maniera versatile, sfruttando le possibilità di consultazione anche attraverso un ambiente distribuito.

La consistenza della base dati è di 9.378 lemmi collocati in 8.973 synsets con 143.701 archi di relazione: la copertura lessicale e i risultati dell'assegnazione automatica necessiterebbero di una ulteriore fase di valutazione e di controllo. Lo strumento è stato reso consultabile attraverso il sito della Fondazione Bruno Kessler[7] che ha sviluppato e messo a disposizione l'interfaccia per effettuare il browsing della rete semantica latina contestualmente a quelle realizzate per altre lingue.

Uno dei punti deboli che si ravvisano nella trattazione di una lingua conclusa con il sistema di una rete lessicale è la mancanza di una collocazione diacronica dei significati all'interno delle stadiazioni della lingua in esame: un'interessante espansione per LWN potrebbe essere quella di un'ulteriore marcatura delle parole afferenti ai gruppi sinonimici che possa far tenere traccia dello spostamento semantico in una prospettiva temporale di storia della lingua. Tale marcatura permetterebbe di identificare l'ingresso di una parola all'interno di un synset (o la sua uscita) nel tempo. Questo lavoro di marcatura storica non può che avvenire, al momento, che attraverso la revisione delle singole parole e dei singoli nodi da parte del lessicografo.

2 Ipotesi di utilizzo di una rete basata su WordNet per l'*Information Retrieval*

Uno dei problemi delle tecniche di reperimento dell'informazione basate sulla corrispondenza di parole è determinato dal fatto che la corrispondenza tra parole e concetti non è una funzione in senso matematico. Nel caso di omografi le parole che sembrano uguali rappresentano concetti diversi, nel caso dei sinonimi, invece, due parole distinte rappresentano lo stesso concetto. Gli omografi rappresentano un ostacolo in quanto diminuiscono la precisione creando falsi positivi e i sinonimi diminuiscono i valori di richiamo in quanto si presentano come falsi negativi. Nel costru-

7 URL http://multiwordnet.itc.it/online/multiwordnet.php (2017-10-19).

ire un sistema di reperimento dell'informazione si parte dall'assunto che l'efficacia degli algoritmi di ricerca dovrebbe migliorare se il confronto non viene operato direttamente sulle parole ma sui concetti che le parole rappresentano.

Le basi di conoscenza lessicale costruite sul modello di WordNet definiscono un concetto attraverso il synset. Pertanto lo sfruttamento dell'informazione semantica contenuta in WordNet è stato investigato in vari modi, soprattutto per quel che riguarda l'interazione tra synsets e concetti nelle operazioni di ricerca testuale. Sussna (1993) e Richardson (1994) trattano la struttura creata dai puntatori relazionali di WordNet come una rete semantica e definiscono alcuni modelli di misurazione per calcolare la distanza tra synsets. La somiglianza tra una *query* e un documento viene poi calcolata dalla similarità tra il gruppo di synsets della *query* e i synsets presenti nei documenti. Quest'ultimo tipo di operazione è particolarmente esosa in termini di risorse di calcolo a causa della sua natura combinatoria: vengono infatti individuate tante coppie di synsets quanti sono gli elementi della *query* e quanti sono i documenti e devono essere valutati tutti i percorsi possibili tra ciascuna coppia.

Per contenere e ottimizzare il numero di confronti, negli esperimenti di Chakravarthy (1994) e di Chakravarthy e Haase (1995), WordNet è stata utilizzata per trovare corrispondenze quando sia le *query* sia i documenti sono brevi e strutturalmente prevedibili. Questo interessante approccio è stato applicato all'indicizzazione delle didascalie in collezioni di immagini. La rappresentazione delle *query* e delle didascalie viene realizzata automaticamente e identifica il ruolo delle parole nel testo: il confronto individua corrispondenze se le parole della *query* e quelle della didascalia sono collegate semanticamente nel *database* lessicale e rivestono lo stesso ruolo nei rispettivi testi.

Un terzo tentativo di confronto tra concetti e *synsets* è motivato dall'obiettivo di migliorare l'efficienza del sistema di ricerca, mantenendo la robustezza e l'affidabilità del modello vettoriale (Sussna 1993). Il focus di questo approccio è dato da una procedura di indicizzazione completamente automatica progettata per scegliere un solo synset per ciascuna parola nel testo. Il risultato di questa procedura di indicizzazione è un vettore nel quale alcuni dei termini rappresentano i synsets anziché le parole. Una volta creato un vettore basato sui synsets, esso viene gestito esattamente come uno basato sulle parole (vale a dire utilizzando l'analisi del coseno come parametro su cui calcolare la similarità fra richiesta e risultato restituito).

Un'alternativa di sfruttamento della rete semantica atto soltanto ad aumentare il richiamo dei risultati all'interno del sistema di ricerca può essere basato sull'utilizzo della struttura dei synsets per espansione della *query* utente. Le strategie di espansione possono essere molteplici: espansione solo attraverso i sinonimi, espansione attraverso i sinonimi e i legami di parentela nella gerarchia, espansione attraverso i sinonimi e

tutti i synsets direttamente collegati (aventi distanza 1 all'interno della catena dei collegamenti).

Gli esperimenti generalmente condotti hanno portato a rilevare come il livello di precisione delle *query* espanse sia direttamente proporzionale al numero di parole specifiche presenti nella *query* originale: l'espansione attraverso i synsets dovrebbe essere combinata in modo da operare sull'intersezione degli insiemi costituiti dai risultati delle *query* individuali. L'intersezione in genere porta a un aumento del richiamo nella maggior parte dei casi ma con una certa degradazione della precisione.

Un sondaggio più approfondito di queste ipotesi implicherebbe in ogni caso il completamento della rete semantica LatinWordNet, onde coprire con continuità tutta la catena ipo-iperonimica e la realizzazione di un vasto corpus di testi semanticamente annotati. Campo di fertile ricerca sarebbe senz'altro l'investigazione di modalità di attribuzione automatica e disambiguazione che superassero quanto già operato da Ellen M. Voorhees (1998) nei suoi ormai classici scritti: ampi orizzonti di ricerca per cui si auspica che questo convegno possa contribuire ad aprire sentieri.

Bibliografia

Atserias, Jordi et al. (1997). «Combining Multiple Methods for the Automatic Construction of Multilingual Wordnets». Mitkov, Ruslan; Nicolov, Nicolas (eds.), *Recent Advances in Natural Language Processing II. Selected Papers from Ranlp '97*. Amsterdam; Philadelphia: John Benjamins, 327-40.

Bentivogli, Luisa; Pianta, Emanuele (2000). «Looking for Lexical Gaps». Heid, Ulrich (ed.), *Proceedings of the Ninth EURALEX International Congress* (Stuttgart, 8th-12th August 2000). Universität Stuttgart: Institut für Maschinelle Sprachverarbeitung, 663-9.

Bentivogli, Luisa; Pianta, Emanuele; Pianesi, Fabio (2000). «Coping with Lexical Gaps When Building Aligned Multilingual Wordnets». Gavrilidou, Maria et al. (eds.), *2nd International Conference on Language Resources and Evaluation = Proceedings of LREC-2000* (Athens, 31st May-2nd June 2000). Paris: European Language Resources Association, 993-7.

Bentivogli, Luisa; Girardi, Christian; Pianta, Emanuele (2002). «MultiWordNet. Developing and Aligned Multilingual Database». Central Institute of Indian Languages (ed.), *Proceedings of the First International Conference on Global WordNet* (Mysore, India, 21st-25th January 21-5). Mysore: Central Institute of Indian Languages, 293-302.

Chakravarthy, Anil S. (1994). «Toward Semantic Retrieval of Pictures and Video». Centre de hautes études internationales d'informatique documentaire (ed.), *Intelligent Multimedia Information Retrieval Systems*

and Management = RIAO '94 Conference Proceedings with Presentation of Prototypes and Operational Systems. Paris: CID-CASIS, 1: 676-86.

Chakravarthy, Anil S.; Haase, Ken B. (1995). «Net.Serf. Using Semantic Knowledge to Find Internet Information Archives». *SIGIR '95 = Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, 9th-13th July 1995). New York: ACM Press, 4-11.

Fellbaum, Christiane (1998). *WordNet. An Electronic Lexical Database (Language, Speech, and Communication)*. Cambridge: MIT Press.

Gavrilidou, Maria et al. (eds.) (2000). *2nd International Conference on Language Resources and Evaluation = Proceedings of LREC-2000* (Athens, 31st May-2nd June 2000). Paris: European Language Resources Association.

Magnini, Bernardo; Cavaglià, Gabriele (2000). «Integrating Subject Field Codes into Wordnet». Gavrilidou et al. 2000, 1413-18.

Miller, George A. et al. (1990). «Introduction to WordNet. An on-Line Lexical Database». *International Journal of Lexicography,* 3(4), 235-44.

Richardson, Ray (1994). *A Semantic-Based Approach to Information Processing* [PhD Dissertation]. Doublin: Doublin City University.

Sussna, Michael (1993). «Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network». *CIKM* '93 = *Proceedings of the Second International Conference on Information and Knowledge Management.* New York: ACM Press, 67-74.

Voorhees, Ellen M. (1993). «On Expanding Query Vectors with Lexically Related Words». Harmna, Donna (ed.), *Proceedings of the First Text REtrieval Conference (TREC-1)* (Gaithersburg, Maryland, November 1-3 1995). Gaithersburg: National Institute of Standards and Technology, 223-32.

Voorhees, Ellen M. (1998). «Using WordNet for Text Retrieval». Fellbaum 1998, 285-303.

Vossen, Piek (1996). «Right or Wrong. Combining Lexical Resources in the Eurowordnet Project». Gellerstam, Martin et al. (eds.) (1996), *Proceedings of Euralex-'96*. Goetheborg University: Department of Swedish, 2: 715-28.