Information retrieval lessicale e semantico sui testi mediolatini.

Stefano Minozzi
Dipartimento di Linguistica Letteratura Scienze della Comunicazione,
Università degli Studi di Verona,
Verona,
Italy
stefano.minozzi@lettere.univr.it

Lezione seminariale per il dottorato in Filologia e Letteratura

Sommario

Si affronta il problema dell'*information retrieval* semantico sui *corpora* testuali di documenti in latino. A partire da una analisi dei metodi di orientamento nel testo e delle implicazioni semiotiche delle attività finalizzate al rinvenimento dei significati e alla creazione di strumenti per l'orientamento testuale, si propone un modello linguisticamente orientato per l'indicizzazione lessicale dei documenti e un modello multilinguale di mappatura semantica degli elementi lessicali.

1 Creare mappe: indicizzazione e orientamento nel testo

1.1 Testualità, lettura e scrittura

Lettura e sapere venatorio

Intendo iniziare questa comunicazione con la definizione di *sapere venatorio* data da Carlo Ginzburg (1979, pp. 66 e 67). Egli definisce la nostra attitudine a leggere e a decifrare come frutto di un sapere venatorio, cioè della

> [...] capacità di risalire da dati sperimentali apparentemente trascurabili ad una realtà complessa e non sperimentabile direttamente

per questo motivo

> [...] il cacciatore sarebbe stato il primo a «raccontare una storia» perché era il solo in grado di leggere, nelle tracce mute (se non impercettibili) lasciate dalla preda, una serie coerente di eventi.

Il cacciatore che decifra tracce di animali compie un processo del tutto simile a quello del lettore che, attraverso la materialità di simboli geometrici, ricostruisce

la rete di significati espressa dal testo: il percorso va da un oggetto materiale ad un oggetto immateriale, il significato, che nulla ha a che vedere con la propria rappresentazione fisica.

Scrittura: risultato di transcodifiche

La scrittura è un atto grafico, l'atto finale di un processo che, a partire dal senso, codifica quest'ultimo attraverso la sintassi e il lessico di un linguaggio naturale e ne dà una convenzionale rappresentazione nell'ordinamento lineare di una sequenza di *glifi*.

Allo stesso modo il processo di localizzazione dell'informazione nel testo è ancora un processo di ricerca simile alla caccia: da un insieme di orme e rami spezzati all'individuazione della preda, dalle tracce del significante al rinvenimento del significato, del quale il testo codifica graficamente la realizzazione linguistica.

Testo: luogo da mappare

Il testo è un luogo in cui ci si può orientare attraverso strumenti fisici e gli indici e le tabelle stanno allo spazio testuale così come le mappe e le bussole stanno all'orientamento spaziale.

L'ente unificatore che permette di stabilire l'equivalenza semantica di ciascun codice è il soggetto-lettore, interprete in grado di superare attraverso la bontà delle sue inferenze l'ambiguità delle codifiche linguistiche e grafiche.

1.2 Il testo elettronico

Venendo a discutere della versione elettronica del testo ci accorgiamo come quanto detto fino ad ora si riproponga: il testo elettronico presenta una sua materialità nella codifica in impulsi che permette di renderlo memorizzabile nell'elaboratore e, allo stesso tempo, può essere rappresentato graficamente sullo schermo del computer perché continui ad essere leggibile al lettore umano, al quale la codifica binaria rimane nascosta.

La memorizzazione di un testo nel computer non è altro che un processo di transcodifica che sostituisce i significanti con altri convenzionalmente equivalenti per una necessità causata da un cambiamento del mezzo. L'aver memorizzato un testo in un elaboratore elettronico non fa dell'elaboratore il destinatario del messaggio, né più né meno di quanto lo sia il foglio su cui scriviamo con penna ed inchiostro.

Inferenza e significazione sono processi umani che nella *semiotica cognitiva* di Peirce, come essa viene definita da Massimo Bonfantini, vengono accomunati:

> Un segno, o *representamen*, è qualcosa che sta a qualcuno per qualcosa sotto qualche rispetto o capacità [...] Definisco un *Segno* come qualcosa che da un lato è determinato da un *oggetto* e dall'altro determina un'idea nella mente di una persona, in modo tale che quest'ultima determinazione, che io chiamo *interpretante* del segno, è con ciò stesso mediatamente determinata da quell'*oggetto* (Peirce, 1980, p.132 e 194)

Interpretazione vs elaborazione

La funzione *interpretante* è un fatto puramente umano ed è bene per questo distinguere tra interpretazione di un testo e processi di elaborazione elettronica di un testo: il testo rappresentato a video, infatti, continua ad essere oggetto di inferenza e di interpretazione per il lettore umano, né più né meno del testo cartaceo, ma il testo come insieme di dati è un oggetto elaborabile da parte del *computer* soltanto nella dimensione del significante.

In altre parole il formato dati *testo* non fornisce una rappresentazione del messaggio adeguata all'elaborazione automatica del contenuto, né il *computer* ha possibilità di realizzare una semiotica intesa con Hjelmslev come individuazione dei rapporti che intercorrono tra espressione e contenuto.
Dunque per poter rendere elaborabili alcune delle caratteristiche del testo che, pur essendo evidenti al lettore umano, non possono essere immediatamente oggetto di elaborazione elettronica è necessario operare una ulteriore transcodifica che attraverso l'aggiunta di segni possa consentire una diversa rappresentazione del contenuto la cui espressione sia sottoponibile al processo di elaborazione, per mezzo di formalismi logici e matematici che modellizzino una teoria del significato. Si tratta in altre parole di realizzare, all'interno dello spazio testuale, una sorta di segnaletica che permetta la creazione di una mappa per l'elaborazione di elementi e di dimensioni del testo che vanno oltre la reticente sequenza simbolica e di individuare algoritmi per il trattamento di questi segni secondo un ben preciso modello semiotico.
Il testo elettronico, quindi, per poter essere in qualche modo *elaborato dal calcolatore nella dimensione del significato* deve essere mappato o come nel caso del *markup* di tipo *strongly embedded* deve contenere segni che ne conducano il tracciamento di una mappa. La logica e la semantica modale[1] elaborano una teoria dei *mondi possibili* che mostra la possibilità del testo di poter essere rappresentato in un altro testo, sulla base di atteggiamenti proposizionali di un terzo interpretante.
Mi sia ora permessa una breve citazione da Borges, *Storia Universale dell'infamia*:

> ...In quell'Impero, l'Arte della Cartografia giunse a una tal Perfezione che la Mappa di una sola Provincia occupava tutta una Città, e la Mappa dell'Impero tutta una Provincia. Col tempo, queste Mappe smisurate non bastarono più. I Collegi dei Cartografi fecero una Mappa dell'Impero che aveva l'Immensità dell'Impero e coincideva perfettamente con esso. Ma le Generazioni Seguenti, meno portate allo Studio della Cartografia, pensarono che questa Mappa enorme era inutile e non senza Empietà la abbandonarono alle Inclemenze del Sole e degli Inverni.

Necessità di una scelta di scala

L'opera di chi si appresta ad editare il testo per renderlo elaborabile dalla macchina è accostabile a quella paradossale dei cartografi imperiali [ma speriamo che non sia simile anche negli esiti finali]: alla base della rappresentazione digitale del testo deve esserci una scelta, per così dire, di scala che ne selezioni le caratteristiche che si desidera rendere elaborabili, dato che la marcatura del testo, come operazione ermeneutica di esplicitazione di strutture di significato, può facilmente produrre una mappa più grande dell'impero.

La scala scelta

Per abbandonare quindi la vertigine della ricorsività e la logica di una semiosi illimitata, la proposta di un sistema di mappatura semantica dei testi latini, che qui si formula, opera nell'ambito di una *scala* che va dalla ricognizione della dimensione lessicale del testo alla sua categorizzazione semantica, nel tentativo di fornirne una sistemazione computazionale a partire dal riconoscimento dell'insufficienza degli strumenti di orientamento basati sull'ordinamento alfabetico.

[1] A tale proposito Kripke (1980)

La consapevolezza dell'utilità degli indici alfabetici per il reperimento delle informazioni nel testo è stata bene illustrata in Papia, che, nella prefazione del suo *Elementarium doctrinae rudimentum*, si prefiggeva di fissare *regulas certas*, perché il lettore potesse risalire velocemente ai contenuti, e sceglieva di disporre i lemmi in rigorosa successione alfabetica. Allo stesso tempo, però, Papia si rendeva conto della difficoltà che una tale disposizione comportava a causa della varietà delle grafie: egli ricorda come

> Hyaena a quibusdam per i, ab aliis per y vel per aspirationem cum diphthongo in penultima scribitur

La difformità colpiva anche la pronuncia dei vocaboli e lo stesso Papia segnalava che

> quam verbenam quidam, alii berbenam vocant herbam

Folia librorum querere

Un altro esempio di reperimento dell'informazione del testo attraverso i suoi aspetti materiali è la mnemotecnica che Ugo di San Vittore proponeva ai suoi allievi nel *De tribus maximis circumstantiis gestorum*:

> Multum ergo valet ad memoriam confirmandam ut, cum libros legimus, non solum numerum et ordinem versuum vel sententiarum, sed etiam ipsum colorem et formam simul et situm positionemque litterarum per imaginationem memoriae imprimere studeamus, ubi illud et ubi illud scriptum vidimus, qua parte, quo loco (suppremo, medio, vel imo) constitutum aspeximus, quo colore tractum litterae vel faciem membranae ornatem intuiti sumus.

Egli consigliava ai propri allievi di mandare a mente le caratteristiche grafiche della pagina, segnando nella memoria visiva la posizione delle *sententiae* nel testo. Si trattava di una forma di orientamento senza indice che si serviva della memoria posizionale e degli aspetti materiali del testo come elemento di collegamento tra il senso e la sua realizzazione in una posizione della pagina.

Precognizione dei risultati

Il reperimento dell'informazione da *corpora* testuali elettronici presenta elementi comuni al primo e al secondo metodo: è infatti quasi sempre un orientamento attraverso un indice (poco importa se questo sia un indice statico pre-generato o oppure venga creato al volo dalla scansione lineare dei documenti) e la consultazione di una base dati testuale prevede la formulazione di ipotesi su come i concetti siano stati lessicalizzati nel linguaggio (o nei linguaggi) dei documenti archiviati e allo stesso tempo deve ricostruirne la forma di rappresentazione grafica; in altre parole l'accesso al testo ideale è possibile soltanto formulando ipotesi sul testo materiale: riprendendo la metafora venatoria potremmo affermare che è come se il cacciatore per risalire alla preda dovesse ipotizzarne la forma delle tracce [e certamente, della caccia, questo tipo di ricerca condivide tutti gli aspetti aleatori]. È chiaro il contrasto con l'esempio medievale: laddove gli allievi di Ugo di San Vittore ritrovavano la posizione dei segni nella pagina, avendone memorizzato gli aspetti materiali dopo una accurata lettura, il ricercatore, che si serve dello strumento digitale, formula una ipotesi su un aspetto materiale del testo (la sua sequenza di glifi) per ritrovare un significato che non è stato, per così dire, visitato in precedenza.

2 Un modello di Information retrieval per i testi latini

IR su testi latini: un modello

Parlando di testi latini, la realizzazione di un modello specifico per l'*information retrieval* semantico si scontra con tre ordini di problemi:

- l'aspetto grafico-ortografico delle parole contenute nel testo;
- la flessione;
- l'imprevedibilità delle realizzazioni lessicali della dimensione semantica.

La proposta che qui si formula si basa sull'impiego di uno strumento di indicizzazione che permetta la compressione delle forme *allografe* o *alloglife*, la loro riconduzione ad un lemma e il collegamento dei lemmi ad una struttura, un *thesaurus semantico*, che organizzi i significati a partire dalla loro realizzazione lessicale, permettendo di simulare la competenza semantica nello strumento di ricerca.
I primi due momenti di indicizzazione operano una progressiva *reductio ad unum* degli elementi lessicali, che va dal catalogo delle forme ad un lemmario, organizzato in modo da permettere l'individuazione della posizione dei lemmi nel testo. Il terzo momento collega i lemmi individuati ad una struttura più ampia (quindi si può considerare come un processo di espansione) che secondo quella che con Eco (1981) chiameremmo una *semantica a dizionario* costituisce ed esplicita una rete di significazione.

2.1 Ricerche lessicali

2.1.1 Compressione delle allografie

Apparentemente l'identificazione degli elementi dell'indice duplicati (o moltiplicati) a causa di grafie alternative potrebbe presentarsi come un tipico problema decisionale: se Σ è l'alfabeto finito utilizzato per la codifica delle parole presenti nei documenti e Σ^* l'insieme delle stringhe finite di simboli di Σ, dato il problema decisionale Π, si individua come D_Π l'insieme di stringhe che codificano un'istanza di Π, i cui elementi sono le forme presenti nell'indice in esame, e come $Y_\Pi \subseteq D_\Pi$ l'insieme di stringhe che codificano *istanze positive* di Π. Quest'ultimo insieme sarà costituito dalle parole presenti nell'indice in forma allografa. In altre parole, si intenderebbe individuare il linguaggio $L(\Pi)$ costituito dalle stringhe di Σ^* che appartengono a Y_Π:

$$L(\Pi) = \left\{ s \in \Sigma^* \mid s \in Y_\Pi \right\}$$

ciò sarebbe possibile utilizzando una funzione di confronto fra gli elementi appartenenti a D_Π che isolasse le forme che possiedono un allografo.
Ad una più attenta analisi risulta evidente che la creazione di tale funzione attraverso un approccio booleano è impossibile: in primo luogo, le grafie alternative di una stringa, soprattutto per quel che riguarda la situazione dei testi mediolatini,

Fuzzy problem

sono frequentemente imprevedibili, pur essendo possibile individuare alcune tendenze[2]. Risulta chiaro, inoltre, che decidere se una stringa dell'insieme *è allografa* di un'altra significa sindacare sulla loro similarità. Non necessita di dimostrazione l'affermazione che similarità e differenza tra stringhe si snodano in un continuo, tanto che ogni stringa di un insieme è simile ad un'altra, purché esse abbiano almeno un simbolo (carattere) in comune. Appare evidente che si sta operando nell'ambito di un problema di tipo *fuzzy*[3].
Sia quindi $\chi_{X,Y}$ il confronto di due stringhe appartenenti all'insieme D_Π ed N il numero di stringhe presenti nell'insieme: si avranno $\frac{1}{2}N(N-1)$ operazioni di confronto, per le quali è necessario un algoritmo in grado di restituire un valore indicativo delle differenze tra le due stringhe, per poter successivamente operare raggruppamenti in base ad una soglia di similarità (a) opportunamente scelta. In altre parole, per ciascuna stringa dell'insieme verrà individuato un linguaggio di D_Π che raggruppa tutte le stringhe che rientrano nella soglia di similarità:

$$\forall\ s \in D_\Pi\ \exists\ A(X) = \left\{ X \in D_\Pi \mid dist(X,Y) \leq a \right\}$$

occorre sottolineare che gli insiemi A(X) individuati possono presentare punti di intersezione, problema di cui si discuterà più avanti.
Prendendo in esame i numerosi algoritmi di confronto esistenti[4] è parso opportuno utilizzare l'algoritmo di Levenshtein (1966) conosciuto come *edit distance* ($dist_{Lev}(X,Y)$): esso permette di calcolare il numero minimo di operazioni di cancellazione, inserimento e inversione di simboli necessarie per trasformare una stringa ($X = x_1 \dots x_m$) nella stringa su cui si opera il confronto ($Y = y_1 \dots y_n$), fornendo una valutazione quantitativa della differenza tra due stringhe. La distanza di Levenshtein può essere definita come $\delta(m,n)$:

Edit distance

$$\delta(0,0) := 0$$

$$\delta(i,j) := min \begin{cases} \delta(i-1,j)+1 \\ \delta(i,j-1)+1 \\ \delta(i-1,j-1)+costo(x_i,y_j) \end{cases}$$

dove

$$costo(x_i,y_j) := \begin{cases} 0, & \text{se } x_i = y_i \\ 1, & \text{se } x_i \neq y_i \end{cases}$$

La complessità dell'algoritmo è $O(N)$, dovuta alla necessità di operare una ricerca lineare.
Il sistema di confronto, come fino ad ora illustrato, non tiene conto del modello linguistico, pertanto devono essere introdotti alcuni passi correttivi che permettano di tenere conto della flessione: le forme flesse appartenenti al medesimo lemma,

[2]La monottongazione dei dittonghi, l'alternanza delle lettere y e i, l'inserimento occasionale e arbitrario di h, la geminazione o lo scempiamento delle consonanti, gli scambi vocalici e consonantici ecc. Per una descrizione estesa dei fenomeni: Cremaschi (1959); De Prisco (1991); Stotz (1996)
[3]Per quel che riguarda la *fuzzy logic*: Novak (1992); Zadeh (1992)
[4]Per una trattazione esaustiva: Knuth (1998); Cormen *et al.* (2001)

senza una correzione di questo tipo, sarebbero considerate tutte allografe tra loro. Inoltre, deve essere individuato un sistema che permetta di abbattere il numero di confronti completi tra le stringhe dell'insieme.

Soundex

La soluzione qui proposta si serve di una raggruppamento iniziale operato da un versione opportunamente modificata dell'algoritmo *Soundex*, originalmente sviluppato da Odell e Russell (1918/1922)[5]. L'insieme D_Π viene preliminarmente suddiviso in partizioni che raggruppano le stringhe aventi il medesimo codice *Soundex* così ricavato:

1. si isola il primo carattere della stringa, sopprimendo tutte le vocali e la lettera h
2. vengono assegnati i seguenti valori alle lettere rimanenti:
   - b, f, p, v $\rightarrow$ 1
   - c, g, k, q, s, x, z $\rightarrow$ 2
   - d, t $\rightarrow$ 3
   - l $\rightarrow$ 4
   - m, n $\rightarrow$ 5
   - r $\rightarrow$ 6
3. se due lettere con lo stesso codice sono adiacenti, se ne mantiene solo una
4. il codice risultante nella forma *carattere,numero,numero,numero* è dato dalla prima lettera seguita dai primi tre valori numerici individuati; nel caso in cui restino meno di tre cifre, si completa con degli 0 (es: imperator→i516; lupi→l100).

I confronti fra stringhe attraverso l'algoritmo di Levenshtein saranno operati solo internamente alle partizioni così costituite e il numero di confronti necessari sarà sempre minore di $\frac{1}{2}N(N-1)$, tranne nel caso in cui tutte le stringhe abbiano il medesimo codice.

Allografie preconosciute

Il confronto puramente computazionale viene affinato grazie ad una tabella di accelerazione, contenente numerose allografie conosciute, che permette di identificare come equivalenti alcune sequenze di caratteri[6], prima dell'applicazione del confronto algoritmico, che restituirà le probabili allografie non normate. Inoltre, il numero dei falsi positivi viene ridotto attraverso un ulteriore controllo sulla base della parte terminale di ciascuna stringa, identificando le forme che probabilmente appartengono alla medesima flessione: di conseguenza, la forma *rose* sarà correttamente identificata come allografa di *rosae* ma non di *rosis*.

2.1.2 Risultati dell'implementazione e discussione

Specimen

Il metodo illustrato è stato implementato in linguaggio di programmazione Visual Basic e applicato alla collezione dei testi retorici attualmente presente nel database

[5] Descritto in Knuth (1998, p.30)

[6] es: i fenomeni che investono i dittonghi, y:i, v:u, cia:tia, cio:tio, cie:tie, x:s, æ:ae, ph:f, y:i, v:u, cia:tia, cio:tio, cie:tie ecc.

ALIM.

Precisione

L'analisi quantitativa dei risultati ha evidenziato come su un indice di 22028 forme siano stati individuati 6681 gruppi di possibili allografi, con una percentuale di falsi positivi del 14,9%: ciò permette di valutare l'efficacia dell'algoritmo in termini di precisione[7], data dal rapporto tra il numero di gruppi individuati e il numero di gruppi rilevanti. Il valore di precisione ottenuto (0,851) può essere considerato un ottimo risultato, tenendo conto che si opera prescindendo dal contesto e da informazioni di tipo semantico.

2.1.3 Lemmatizzazione

E pluribus unum

Il secondo momento di compressione dell'indice, vede la realizzazione di un meta-indice che raggruppi le forme per lemmi apponendo ad esse un codice di descrizione. Il tipo di implementazione realizzata nella presente ricerca si serve di un approccio semi-automatico e di una base di conoscenza lessicale di circa quarantamila lemmi. La costruzione dell'indice dei lemmi avviene risolvendo le situazioni di omografia esolemmatica (tra forme di due o più lemmi)[8] in maniera assistita attraverso un algoritmo che su base statistica individua i candidati più probabili per la disambiguazione. Questo tipo di organizzazione, permette ricerche lessicali molto più efficienti rispetto a quelle possibili sulle collezioni elettroniche che operano solo sulle forme flesse.

2.1.4 Trattamento dell'omografia esolemmatica

Omografia esolemmatica

Uno dei problemi relativi alla fattibilità di un algoritmo per la lemmatizzazione è quello legato alla possibilità di disambiguare automaticamente le situazioni di omografia esolemmatica, quei casi, cioè, dove una parola presenti una forma ascrivibile a lemmi di categorie grammaticali differenti a causa dell'omografia.

Hidden Markov Model

Nel corso della ricerca è emerso che l'attribuzione della parte del discorso corretta può essere ottenuta con un margine di errore sufficientemente basso attraverso l'applicazione di un processo probabilistico basato sul modello statistico chiamato *Modello nascosto di Markov*. Questo modello presuppone che il sistema modellizzato sia un processo markoviano con parametri sconosciuti: l'obiettivo è riuscire a inferire dai parametri osservabili il parametro nascosto. Questo modello è applicato frequentemente in processi di *pattern recognition* e può essere adattato all'analisi automatizzata del discorso.

Modellizzazione

Nel problema esposto i dati che devono essere analizzati sono le categorie grammaticali delle parole che possono essere ricondotte a più di una parte del discorso. Sia quindi (Ω, A, P) uno spazio probabilizzato dove Ω rappresenta l'insieme delle categorie grammaticali della lingua latina, A la tribù delle parti di Ω e P una misura di probabilità su Ω che verrà specificata in seguito.
In un tale spazio si organizza un blocco di variabili aleatorie

$$[C_i]_{1 \leq i \leq k}$$

[7] Cfr. Grossman e Frieder (2004, p.5)

[8] Per una classificazione tassonomica dell'omografia latina si rimanda a: Passarotti e Ruffolo (2004)

che, per ogni i, rappresenta la categoria grammaticale della i-esima parola analizzata, essendo k il numero totale delle parole presenti nella frase analizzata (C è la categoria grammaticale della parola i, con i compreso largamente tra uno e il numero totale delle parole della frase). Usando l'ipotesi di Markov si assume che

$$P\{C_i \mid C_{i-i}, C_{i-2}, \ldots C_1\} = P\{C_i \mid C_{i-i}, C_{i-2}\}$$

cioè la probabilità che la i-esima parola sia ascrivibile a una determinata categoria grammaticale (C_i) dipende dalle categorie grammaticali delle due parole che la precedono (o, eventualmente, in una modellizzazione più complessa, che viene discussa più ampiamente nella tesi, dalle categorie delle due parole adiacenti, precedenti o conseguenti). Data la formula della probabilità condizionata:

$$P\{C_i \mid C_{i-1}, C_{i-2}\} = \frac{P\{C_i \cap C_{i-1} \cap C_{i-2}\}}{P\{C_{i-1} \cap C_{i-2}\}}$$

nell'impossibilità di ricavare direttamente la legge di P, deve essere scelto un valore di C_i che massimizzi tale formula.

Corpus

Perché sia possibile individuare tale valore si ricorre all'analisi automatizzata di un congruo numero di testi raccolti in un *corpus* di natura omogenea, estraendo, attraverso un algoritmo di confronto sulla base di un dizionario completo, l'insieme di tutte le terne che presentano sequenze di parti del discorso attribuibili in maniera univoca e conservando una stima della frequenza. Si ottengono così insiemi di terne accompagnate dalle rispettive occorrenze nel corpus in esame (per esempio Verbo-Preposizione-Sostantivo-80; Aggettivo-Sostantivo-Verbo-56; ecc.). Nel caso di una situazione di ambiguità vengono presi in esame tutti i gruppi che presentino sequenze di categorie grammaticali identiche a quelle che precedono la parola che si deve disambiguare. Alla parola da disambiguare viene assegnata la categoria grammaticale della parola che si trova nella terna più frequente tra quelle considerate.

Il procedimento descritto può essere applicato in tutti quei casi in cui sia possibile individuare una sequenza di tre termini tra i quali sia ambigua la categoria grammaticale di uno: il problema viene risolto su base statistica e i risultati sono direttamente influenzati dalla omogeneità delle frasi analizzate con quelle delle strutture presenti nel *corpus* utilizzato come strumento per l'estrazione delle terne.

2.2 Ricerche semantiche

Semantica e NLP

La possibilità di operare ricerche attraverso l'astrazione dei contenuti è uno dei principali interessi di studio della linguistica computazionale: di particolare interesse può risultare l'applicazione di queste tecniche alle lingue concluse, come il greco antico e il latino. Qui si propone di applicare alla struttura della frase latina una indicizzazione semantica, prendendo in parte spunto dal modello di Roussey *et al.* (1999). La base per l'implementazione del modello è la realizzazione di un *thesaurus* semantico che permetta di unire più lemmi all'interno di una definizione di dominio e che permetta di chiarire le relazioni tra i lemmi, definendo due livelli di conoscenza:

1. un livello concettuale che dia un modello del campo di studio formato dai concetti e dalle relazioni che intercorrono tra di essi;

2. un campo terminologico che rappresenti l'insieme delle manifestazioni linguistiche di un concetto nel testo.

2.2.1 *Thesaurus* semantico

WordNet

Tra i tipi di modellizzazione della conoscenza lessicale è parsa particolarmente interessante la rappresentazione dei rapporti semantici nei dizionari definita dagli studi di Miller *et al.* (1990) e Fellbaum (1998): a partire dal riconoscimento della natura del tutto accidentale dello *spelling* delle parole, nel modello di *WordNet* le parole sono organizzate per blocchi di significato, denominati *synset*, che raccolgono tutti i lemmi che lessicalizzano lo stesso concetto; i *synset* 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 *synset* e parole.
L'applicazione di tale modello ad un *thesaurus* semantico per la lingua latina rappresenta l'orientamento dei prossimi studi di chi vi parla: si ritiene utile riportare in questa sede una schematizzazione del metodo di costruzione di tale rete semantica.
Risultando impraticabile costruire da zero l'insieme di relazioni di una rete di questo tipo, il punto di inizio del progetto di costruzione si basa sull'ipotesi che la rete dei significati, definita per la versione inglese, possa essere in gran parte portata verso altri linguaggi. Una ipotesi plausibile, se ci si limita alle principali lingue indoeuropee che presentano una vasta sovrapposizione culturale.

Mapping

2.2.2 Matrice lessicale multilingue

Multilingual lexical matrix

La matrice lessicale della *Wordnet* inglese è bidimensionale (si estende nelle due dimensioni dei lemmi e dei significati); sulla scorta di quanto effettuato in progetti analoghi[9] per le lingue moderne, aggiungendo una terza dimensione alla matrice (la dimensione delle lingue) diventa possibile considerare la lingua latina. Per realizzare la matrice multilingue, si rende necessario mappare i lemmi latini sui significati corrispondenti, andando a costruire l'insieme dei *synset* per il latino. Il risultato è una completa ridefinizione delle relazioni lessicali, mentre per la creazione della rete di relazioni semantiche vengono impiegate, per quanto possibile, quelle già definite per l'inglese. La dimensione dei significati, pertanto, viene considerata costante rispetto alle possibili lessicalizzazioni a livello linguistico. In un primo tempo, attraverso l'utilizzo di algoritmi di *matching* e di un dizionario bilingue, viene verificata la corrispondenza tra i lemmi delle lingue che devono essere aggiunte alla rete. Questa prima fase di automazione produce un insieme di possibili

[9] Alcuni tra i più significativi: Artale *et al.* (1997); Chen *et al.* (2002); Lee *et al.* (2004)

connessioni tra una parola latina e i significati nella rete *WordNet*. A questa fase segue l'intervento umano per validare le scelte proposte.
Il campo di applicazione di una rete così costituita va dal disambiguamento dei contesti, all'IR semantico. *Non solo IR*
Il collegamento tra lemmi presenti nel testo e rete semantica, costituisce una mappa della catena di significati realizzati, che prescinde della loro lessicalizzazione: l'insieme dei riferimenti ai *synset* contenuti nella struttura di una frase diventa, così, il descrittore del contenuto di quella frase e l'elemento su cui deve operare il processo di ricerca.

2.2.3 Misura della similarità semantica

Per l'utilizzo nell'ambito di un sistema di *information retrieval* della rete semantica è stato applicato il modello vettoriale di Salton *et al.* (1975). In questo modello i documenti e le *query* sono rappresentati da vettori in uno spazio N-dimensionale dove N è il numero di lemmi distinti presenti nella collezione di documenti. Il processo automatico di lemmatizzazione contribuisce ad una compressione dell'indice e a una attribuzione di pesi alle parole indicizzate in maniera direttamente proporzionale al numero di occorrenze della parola nel testo e inversamente proporzionale al numero di documenti in cui la parola è presente. Data una *query* il sistema produce una lista ordinata di documenti in base alla similarità rispetto alla *query*. La misura della similarità è data dal coseno dell'angolo formato tra il vettore della *query* e il vettore che rappresenta il segmento di testo che chiamiamo documento. Se d_i è il peso del termine i nel vettore D e q_i è il peso del corrispondente termine nel vettore della query Q la misura della similarità tra il documento e la query è:

$$sim(Q, D) = \cos(Q, D) = \frac{\sum_{i=1}^{N} q_i d_i}{\sqrt{\sum_{i=1}^{N} q_i^2 \sum_{i=1}^{N} d_i^2}}$$

Il contesto minimo su cui opera l'indicizzazione e quindi la costruzione dei vettori è la frase, intesa come sequenza di simboli tra due segni di interpunzione forte. I vettori, vengono costruiti in base all'assegnazione delle parole contenute nella frase ai *synset* della rete semantica: in questo modo la frase viene considerata come una sequenza di *synset*, vale a dire marche che puntano ad un'area di significati. Ciò consente migliorare l'efficienza del sistema di ricerca.

3 Conclusione

Alla fine della descrizione del percorso di ricerca, intendo sottolineare che ci si rende conto che la capacità di rappresentazione e, di *produzione* di senso propria dell'oggetto testo non è limitata, e non si esaurisce, nella sua dimensione lessicale, ma si estende anche nella dimensione sintattica e nelle capacità attributive del soggetto, a tal proposito, rimando a quanto detto precedentemente sulla necessità di operare una *scelta di scala* nella costruzione del modello.
Il percorso fin qui descritto ha portato alla realizzazione di una serie di strumenti di analisi automatica del testo, come aspetto applicativo (o, se vogliamo, come

sottoprodotto) della ricerca: il già citato sistema di identificazione delle allografie e il meccanismo di indicizzazione e lemmatizzazione. Si segnala inoltre che recentemente è stato rilasciato al pubblico dominio il un dizionario latino compatibile con il *software* WinEdt per il sistema di editoria elettronica LaTeX: il dizionario composto da una lista di 1243950 forme latine, permette il controllo ortografico dei testi ed è utilizzabile, con lievi adattamenti, in applicazioni di riconoscimento dei caratteri, per migliorare la qualità dei risultati di acquisizione.
Inoltre il progetto del *thesaurus* semantico descritto nella presente tesi di dottorato ha incontrato l'interesse dell'Istituto Trentino di cultura e si è concretizzato in una convenzione interdipartimentale tra il Dipartimento di Linguistica Letteratura e Scienze della Comunicazione dell'Università di Verona e la Divisione di ricerca per le Tecnologie Cognitive e della Comunicazione del Centro per la Ricerca Scientifica e Tecnologica di tale Istituto (IRST-ITC).
Nel chiudere questa comunicazione, che ha voluto proporre un metodo per il rinvenimento automatico del significato, mi siano concessi un'ultima breve citazione e un augurio. Prima la citazione:

> So d'una regione barbarica i cui bibliotecari ripudiano la superstiziosa e vana abitudine di cercare un senso nei libri, e la paragonano a quella di cercare un senso nei sogni o nelle linee caotiche della mano... Ammettono che gli inventori della scrittura imitarono i venticinque simboli naturali, ma sostengono che questa applicazione è casuale, e che i libri non significano nulla *di per sé*.

L'augurio è che i libri possano sempre, e comunque, significare *per noi*.

Riferimenti bibliografici

Artale, A., Magnini, B. e Strapparava, C. Proceedings of ACL/EACL-97 Workshop Lexical discrimination with the Italian version of WordNet. In P. V. et al., cur., *Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications (Madrid, Spain, July 1997)*. Association for Computational Linguistics, ACL 1997

Chen, H.-H., Lin, C.-C. e Lin, W.-C. Building a Chinese-English wordnet for translingual applications. In *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1(2) [2002]:pp. 103–122

Cormen, T. H., Leiserson, C. E., Rivest, R. L. e Stein, C. *Introduction to Algorithms, Second Edition*. The MIT Press 2001

Cremaschi, G. *Guida allo studio del latino medievale*. Liviana editrice, Padova 1959

De Prisco, A. *Il latino tardoantico e altomedievale*. Jouvance 1991

Eco, U. *Enciclopedia*, vol. XII, cap. Significato. Einaudi 1981, pp. 830–876

Fellbaum, C., cur. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press 1998

Ginzburg, C. *Crisi della ragione*, cap. Spie. Radici di un paradigma indiziario. Einaudi 1979

Grossman, D. A. e Frieder, O. *Information Retrieval: algorithms and heuristics*. Springer, Dordrecht 2004, seconda ed.

Knuth, D. E. *Art of Computer Programming, Volume 3: Sorting and Searching*. Addison-Wesley Professional 1998, seconda ed.

Kripke, S. *Naming and Necessity*. Basil Blackwell, Oxford 1980

Lee, C., Lee, G. G. e Seo, J. Multiple Heuristics and Their Combination for Automatic WordNet Mapping. In *Computers and the Humanities*, vol. 38(4) [2004]:pp. 437–455

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, vol. 10(8) [1966]:pp. 707–710

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. e Miller, K. J. Introduction to WordNet: an on-line lexical database. In *International Journal of Lexicography*, vol. 3(4) [1990]:pp. 235 – 244

Novak, V. Fuzzy Sets in Natural Language Processing. In R. R. Yager e L. A. Zadeh, cur., *An Introduction to Fuzzy Logic Applications in Intelligent Systems*. Kluwer, Boston 1992, pp. 185–200

Odell, M. K. e Russell, R. C. U.S. Patents 1261167 (1918), 1435663 (1922)† [1918/1922]. Brevetto non pubblicato. Citato in Knuth (1998)

Passarotti, M. e Ruffolo, P. L'utilizzo del lemmatizzatore LEMLAT per una sistemazione dell'omografia in latino. In *Euphrosyne*, vol. 32 Nova Série [2004]:pp. 99–110

Peirce, C. S. *Semiotica*. Einaudi, Torino 1980

Roussey, C., Calabretto, S. e Pinon, J.-M. Etat de l'art en indexation et recherche d'information. In *Revue Document Numérique*, vol. 3(3-4) [1999]:pp. 121–150

Salton, G., Wong, A. e Yang, C. S. A Vector Space Model for Automatic Indexing. In *Commun. ACM*, vol. 18(11) [1975]:pp. 613–620

Stotz, P. *Handbuch zur lateinische Sprache des Mittelalters*. Beck 1996

Zadeh, L. A. Knowledge Representation in Fuzzy Logic. In R. R. Yager e L. A. Zadeh, cur., *An Introduction to Fuzzy Logic Applications in Intelligent Systems*. Kluwer, Boston 1992, pp. 1–25