

METODOLOGIE INFORMATICHE E DISCIPLINE UMANISTICHE (MODULO B)

7. I SISTEMI DI ANALISI DEL TESTO E LA LINGUISTICA COMPUTAZIONALE

La disponibilità di corpora di dati strutturati o semi-strutturati in ambienti documentali distribuiti ha di recente accresciuto le possibilità di lavorare con *sistemi automatici di analisi del testo* nel campo dello studio del linguaggio naturale.

Gli ipertesti e gli archivi digitali di testi sono due esempi di collezioni che possono essere sottoposto all'indagine testuale.

Un *corpus* si deve basare su una serie di regole di aggregazione, selezione e organizzazione precise, necessarie affinché la collezione possa essere sottoposta a un'analisi linguistica funzionale a ottenere risultati significativi.

Tre aspetti importanti:

1. *text retrieval* → ricerca per stringa di caratteri;
2. i diversi livelli della manipolazione del testo;
3. le tipologie di interrogazione del testo così manipolato, finalizzate all'*information retrieval* e *information extraction*.

La linguistica computazionale è la disciplina che per prima ha affrontato il problema del trattamento automatico del linguaggio naturale, con l'obiettivo di acquisire conoscenza da documenti testuali. Superato il livello di semplice riconoscimento di stringhe di simboli è possibile approdare a un secondo momento di trattamento della lingua: sistemi di manipolazione di stringhe, analisi automatica del testo ed estrazione di informazione dai dati.

Nel campo degli studi letterari, i sistemi di analisi del testo consentono lo studio del vocabolario d'autore e la verifica dell'uso di certi costrutti sintattici e finalizzati all'*attribuzione di paternità*.

1. FORME DI TEXT RETRIEVAL

Verifica della presenza delle parole o diverse forme testuali all'interno del testo.

La *tokenizzazione* è l'individuazione delle unità minime del testo; estrapolare tutte le forme presenti all'interno del testo per verificare il numero di volte in cui compaiono: definire le occorrenze. Il fine è la creazione dell'indice (composto da *types*: i token lemmizzati)

Il passo successivo è la collocazione delle parole rispetto al contesto testuale d'uso (concordanze) e la verifica di quante volte ciascuna forma grafica occorre (frequenze).

Esistono molti programmi in grado di fare operazioni di *text retrieval*, cioè che sono in grado di estrarre tutte le sequenze di caratteri che stanno tra due spazi o caratteri speciali (se istruiti dall'utente).

L'operazione preliminare è la scelta oculata del testo base. Questa scelta si mostra indispensabile nel caso di un testo che riporta differenti edizioni; soprattutto occorreranno, una volta caricato il testo in formato elettronico, riferimenti espliciti al testo impiegato come modello.

- a. Concordanze → Un programma di concordanze è un applicativo che consente di enucleare tutte le parole presenti nel testo, presentandole in ordine alfabetico, accompagnate da un contesto e da una serie di indicazioni che permettono il reperimento e la localizzazione del passo all'interno della struttura del testo.

Non sempre nel preparare le concordanze di un'opera vanno elaborate tutte le parti del testo; generalmente si tende a togliere le "parole vuote", ma, vi sono casi dove proprio queste parole hanno significato → es. quando si vuole fare uno studio sui *legami sintattici*.

L'analisi delle concordanze permette di disambiguare impieghi diversi del lessico, agevolando il confronto fra significati diffusi della stessa forma grafica.

- b. Indici → Un indice può essere considerato come un caso particolare di concordanza prima di contest; è una lista di vocaboli contenuti all'interno di un testo – o *corpus* – dove ogni parola è accompagnata dal riferimento al luogo in cui è possibile tracciare l'occorrenza. Talvolta l'indice può riportare alcune statistiche relative alla frequenza relativa o assoluta dei vocaboli all'interno del testo.
- c. Frequenze → la lista di frequenze di un testo mostra le parole che lo compongono accompagnate dal numero di volte in cui occorrono e a volte la percentuale rispetto al numero totale di parole. La posizione che ogni vocabolo occupa all'interno della lista di frequenze è il "rango". Bisogna stabilire una soglia al di sopra/sotto della quale le parole vengono considerate rare o frequenti; *hapax*: parole che compaiono una sola volta.

L'informatica, in senso strumentale, non produce in realtà risultati differenti da quelli ottenibili tradizionalmente ma permette una maggiore rapidità con l'utilizzo di più materiali.

2. DAL TEXT RETRIEVAL ALLA TEXT ANALYSIS

Dalla fase di recupero stringhe è possibile passare all'analisi del testo. Normalizzazione, lemmatizzazione, *part of speech*, *tagging*, *parsing*, riduzione della sinonimia ecc. sono alcuni dei processi che caratterizzano l'attività del *natural language processing*.

Innanzitutto, è necessario operare sul testo → al fine di ottenere risultati attendibili nella fase della *text analysis* è necessario intervenire con l'annotazione del testo.

Annotare il testo significa arricchirlo di informazioni sui differenti aspetti di interesse ai fini dell'analisi. La più comune forma di annotazione di un testo è l'assegnazione di marcatori o etichette. Tali indicazioni sono generalmente standardizzate; possono essere aggiunte direttamente nel corpo del testo o definite in un file separato ed essere richiamate tramite collegamento.

Tre livelli di analisi del testo: *morfo-lessicale*, *sintattico*, *semantico*.

Per lavorare a questi livelli, prima di tutto, bisogna disporre di risorse linguistiche, cioè i repertori linguistici e lessicali, che possano coadiuvare nell'operazione.

Il dizionario macchina è la versione elettronica di un dizionario tradizionale che elenca tutti i lessemi e associa a ciascuno le informazioni tipiche di un dizionario tradizionale. Può essere usato nella fase di analisi morfo-lessicale, oggetto principale di studio della lessicografia.

Un lessico di frequenza è un elenco di forme e di lemmi con indicazioni della frequenza d'uso rispetto a un corpus definito. Generalmente è usato per l'estrazione di parole chiave del testo.

Una rete semantico-concettuale vuole associare ogni lessema a un concetto e quindi a una classe semantica di riferimento all'intero della struttura gerarchica, ma anche individuare le relazioni che i concetti e le classi intrattengono.

È necessario che la fase di tokenizzazione sia accompagnata da un procedimento di identificazione dei *tokens* significativi ai fini della successiva analisi → operare a diversi livelli:

- normalizzazione delle varianti ortografiche delle parole che possono presentarsi sotto diverse forme;
- separazione di parole che sono costituite da più *token*;
- unione di elementi differenti in un unico *token*.

Effettuata la normalizzazione, ogni token potrebbe essere associato alla parte del discorso cui ogni forma può essere ricondotta.

Tramite l'ausilio di un analizzatore morfologico, ogni componente lessicale può essere descritta in termini di categoria grammaticale di appartenenza.

Il limite della tokenizzazione è che consente di estrarre il solo elenco delle forme; bisognerà allora adottare procedure che permettano di trasformare i token in lemmi → *stemming*.

Lemmatizzare un testo significa individuare un unico lemma → un'unica forma grammaticale. La lemmatizzazione si pone come obiettivo di ricondurre a unità queste forme raccogliendole sotto un'unica forma base. Esistono programmi specifici per la lemmatizzazione o per l'analisi morfologica dei testi.

Strumenti informatici come i dizionari macchina possono agevolare l'operazione di lemmatizzazione. Il compito della lemmatizzazione è anche risolvere problemi di ambiguità fra diverse forme base cui può corrispondere una stessa forma flessa.

L'analisi sintattica è rappresentata dal processo di *parsing* → gli strumenti consentono di associare agli elementi della frase un determinato valore sintattico. Il primo passo è costituito dal *tagging*, si tratta dell'annotazione: associare un'etichetta descrittiva a ogni costituente grammaticale.

Sulla base del contesto d'uso della parola gli strumenti di tagging consentono di disambiguare la *part-of-speech* di ciascun componente; si occupa di associare una categoria sintattica con cui ogni forma occorre in un dato contesto linguistico.

Chunking → processo di segmentazione del testo analizzato morfologicamente in gruppi sintattici.

Le parole che compongono una frase possono essere ricondotte a un gruppo funzionale di riferimento e i sintagmi essere inseriti in uno schema di relazioni di dipendenza grammaticale. L'uso di sistemi formali per l'analisi del linguaggio si colloca a seguito della fondazione della grammatica generativa di Chomsky.

L'analizzatore sintattico identificherà le dipendenze grammaticali principali; sarà così possibile realizzare una rappresentazione della struttura morfo-sintattica del testo necessaria a processi di estrazione della conoscenza linguistica dai testi annotati.

Un corpus annotato dal punto di vista della struttura sintattica è detto *treebank*.

L'oggetto dell'analisi semantica è costituito dalla ricostruzione del significato del vocabolario presente del testo.

Categorizzazione semantica → associazione di una comune categoria a diverse parole che rientrano nel medesimo raggruppamento concettuale. L'annotazione semantica si distingue fra:

- specificazione del significato di un elemento desunto da una risorsa lessicale, che associa la parola ad una descrizione semantica;
- marcatura dei ruoli semantici.

Lo strumento usato ai fini dell'associazione semantica è la rete semantico-concettuale: *iponimi*, *iperonimi* e *olonimi*.

La disambiguazione del significato di una parola è una delle attività più difficilmente risolvibili in modo automatico e l'annotazione deve essere svolta dallo studioso sulla base del contesto d'uso del vocabolario.

Da un lato per risolvere il problema della specificazione dei collegamenti fra le risorse, vale a dire della possibilità di distinguere le diverse tipologie di relazioni che sussistono fra i vocabolari correlati; dall'altro per lo sviluppo di ontologie, per stabilire i concetti e creare le relazioni sussistenti fra i concetti relativi a un dominio.

3. TIPOLOGIE DI RICERCA SUL TESTO E INTERROGAZIONE SIGNIFICATIVA

La verifica della frequenza è funzionale al procedimento di estrazione delle parole chiave: tanto più un termine è presente, tanto più è significativo ma solo se non è frequente in altri testi.

Disponendo di lessici di frequenza specialistici questa operazione può raggiungere livelli di precisione significativi.

La verifica delle concordanze permette di ragionare sull'impiego del vocabolario nei diversi contesti d'uso e disambiguare facilmente gli omografi.

Oltre all'interrogazione delle singole parole è possibile ricercare forme di co-occorrenze.

Gli operatori logici sono impiegati anche dalla ricerca effettuata usando le espressioni regolari.

Alla ricerca di co-occorrenze si affianca quella di *collocations*, cioè la verifica della presenza di parole che co-occorrono fornendo un senso specifico, cioè parole o lemmi che, quando compaiono in abbinamento, esprimono un preciso concetto.

Le nuove frontiere dell'analisi del testo si collocano nel *text mining*, cioè nell'estrazione di informazione significativa dal testo non strutturato, con l'obiettivo di ottenere una nuova conoscenza: sistemi di *clustering* dei documenti sulla base dei contenuti trasmessi e di classificazione in categorie, permettono di trasformare i documenti in informazione e quindi in conoscenza.