Linguistic Miner

Eugenio Picchi, Maria Luigia Ceccotti, Laura Cignoni, Nella Cucurullo, Giovanni Fiorentini, Manuela Sassi, Eva Sassolini, Giovanna Turrini - Istituto di Linguistica Computazionale - Area della Ricerca di Pisa - CNR
[eugenio.picchi@ilc.cnr.it]

Abstract: In this paper we present a project titled "Linguistic Miner", designed and coordinated by Eugenio Picchi. The project arises from the availability of the PiSystem tools and the familiarity with the automatic treatment of human language. The project goal is the extraction of linguistic information from the texts and the validation of linguistic patterns. We show the objectives and the results of the project as achieved in the first months of work.

Introduzione

Alla fine degli anni ’90 la rivoluzione Internet ha aperto nuove e multidisciplinari prospettive di ricerca. La più ricca biblioteca del mondo, un insieme eterogeneo di dati in gran parte testuali, che viene incrementato senza soluzione di continuità, senza regole, senza piani di acquisto, senza autorizzazioni, può essere considerata sia la torre di babele dei nostri giorni sia lo strumento più adeguato per la diffusione della conoscenza. Il recupero dei dati dal web, l’estrazione dai dati delle informazioni, l’archiviazione delle informazioni sono le fasi delicate di un percorso che da pochi anni è oggetto di importanti progetti di ricerca impegnati nella soluzione di problemi quali la definizione di metodologie adeguate ad integrare dati strutturati e non, la creazione di architetture per *data warehouse*, la definizione di tecniche che permettano di estrarre conoscenza dai dati, il *data mining*.
La *Linguistica Computazionale*, caratterizzata sin dai primi passi da due filoni di ricerca, quello ‘lessicografico’, impegnato a perfezionare, ottimizzare, velocizzare tecniche di lavoro collaudate da secoli, quello ‘cognitivo’ alla ricerca di modelli, di formalizzazioni dei vari aspetti dell’attività linguistica, si colloca in questo importante settore con il vantaggio di una notevole esperienza maturata nel campo del TAL.
Linguistic Miner è un progetto nato nell’ILC, progetto che ha come obiettivo la costruzione di un sistema che estragga automaticamente da testi in lingua italiana, di varia provenienza e formato, conoscenza linguistica utilizzabile per scopi molteplici: didattici, editoriali, culturali, etc. Il progetto nasce da una considerazione preliminare a tutti i sistemi di analisi linguistica *corpus based*: una lingua, rappresentata da un insieme, il più grande possibile, di testi delle più varie tipologie è la miglior fonte di informazione linguistica, a qualunque livello di analisi la si consideri. Quanto più grandi sono i corpora disponibili e quanto più rappresentano in maniera eterogenea i vari ambiti linguistici (differenziati secondo le tipologie comunicative) tanto maggiore è la loro rappresentatività della realtà linguistica di una lingua. Importante sarà in questo scenario elaborare tecniche che permettano di monitorare continuamente il bilanciamento tra i vari settori della LM. Quindi la capacità di costruire grandi corpora di riferimento di una lingua è il primo e fondamentale obiettivo; ma immediata necessità successiva è quella di poter creare e disporre di efficaci strumenti per la gestione di tali corpora, per la loro analisi e per la realizzazione automatica di sintesi linguistiche. La conoscenza prodotta sarà di tipo e associativo e verificativo. La letteratura in proposito già da alcuni anni sottolinea la differenza sostanziale tra l’informazione ottenuta tramite operazioni di recupero (*data retrieval*) o di statistica e la conoscenza (*data mining*) che sfrutta relazioni, esplicite ed implicite, tra i dati testuali.

Basi di Partenza del Progetto

La prima fase del progetto è stata dedicata:
- al disegno di un'architettura integrata del sistema LM per la gestione, l'archiviazione e l'utilizzo di tutte le risorse e gli strumenti che entreranno a far parte del sistema;
- al censimento del corpus di testi, in lingua italiana, dell'ordine di grandezza di centinaia di milioni di parole, utile per la definizione di una prima classificazione dei testi, classificazione di riferimento iniziale per la fase di acquisizione dati dal web; attualmente sono disponbili circa duecento milioni di parole;
- al controllo dei moduli che fanno parte del PiSystem (Picchi 94), un sistema integrato per il trattamento di materiali testuali e lessicali, la cui componente principale è il DBT (Data Base Testuale), una procedura per l’analisi testuale, attuale versione della procedura di spoglio elettronico intorno alla quale l’ILC è sorto e si è sviluppato. Lo scopo è stato quello di individuare la possibilità di integrare, adattare le diverse componenti del sistema per costruire procedure di elaborazione finalizzate ai nuovi obiettivi.

Figura 1: Schema del Linguistic Miner con le sue fasi di lavoro

Fase di Acquisizione

Il progetto prevede l'acquisizione di testi provenienti dalle più svariate fonti (dati web e non web). Se per questi ultimi (ad esempio CD contenenti gli articoli di quotidiani, riviste) esistono già modalità consolidate, per il recupero di dati web sono state realizzate procedure di acquisizione automatica (*spider*) di siti ad aggiornamento periodico ed altre per lo scaricamento ragionato e guidato da scelte dei ricercatori. Il sistema prevede l'analisi del materiale già nella prima fase di recupero e l'inserimento nella grande banca-dati dei singoli elementi linguistici che costituiscono la specifica ricchezza di ogni testo, inibendone al tempo stesso la capacità di riproduzione e la lettura. Tale procedimento rende il materiale non più strettamente dipendente dalla sorgente originale e, al tempo stesso, nulla toglie alle potenzialità ed agli obiettivi del progetto che rimangono puramente linguistici e non di analisi dei contenuti o di ricerca documentale.

Fase di Analisi/Codifica

Le procedure di acquisizione hanno il compito di individuare e scaricare pagine testo in formato HTML. La necessità di identificare in tali pagine la parte testuale e di classificarla ha richiesto la realizzazione di procedure per l'interpretazione e la corretta codifica dei materiali. Tale strumento di codifica è stato realizzato, con diverse percentuali di rendimento, anche per altre tipologie di materiale (Word, RTF, PDF). La qualità dei risultati ottenuti da procedure di analisi linguistica automatica dipende dalla quantità di elementi del testo correttamente etichettati e per questo è molto importante predisporre il pre-editing automatico dei materiali per ottimizzare le successive elaborazioni. A questo scopo sono state inserite e raffinate procedure, mutuate in parte dal progetto PiSystem, per l'individuazione ed il trattamento di vari fenomeni quali: struttura del testo, sigle, numeri, abbreviazioni, nomi propri (parole isolate o espressioni), collegamenti ipertestuali, indirizzi di posta elettronica, etc. Una volta individuati e marcati gli elementi, il testo viene immagazzinato nella banca-dati secondo il formato interno del DBT per successive analisi.

Monitor di Sistema

La fase di classificazione di primo livello prevede la categorizzazione di ogni testo immesso: l'informazione necessaria per tale operazione sarà rilevata dalla provenienza del materiale e da una analisi sommaria, che ne fornisce un indice provvisorio di classificazione per una utilizzazione ragionata e comparativa di settore. Nella Fig.2 viene proposto un esempio di schermata del monitor di controllo:

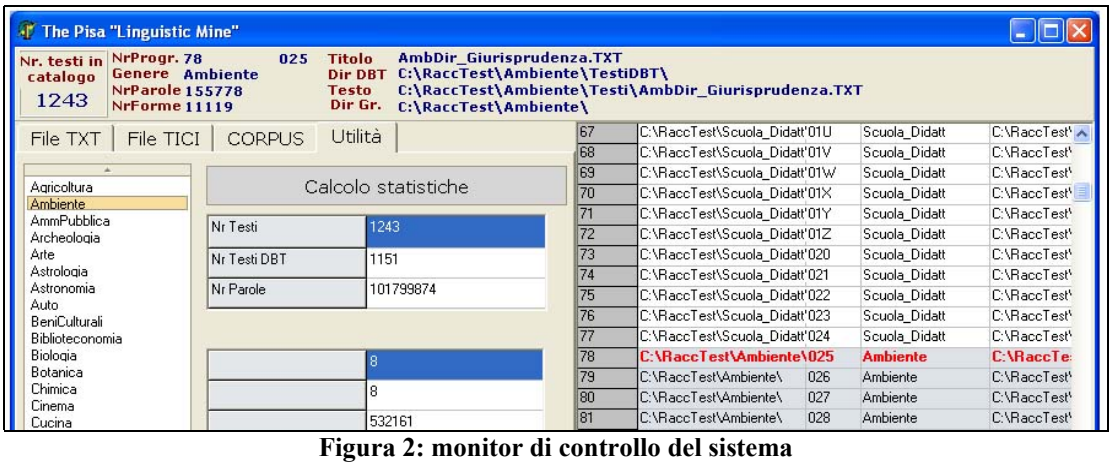


Figura 2: monitor di controllo del sistema

Fase di Estrazione dalla *Linguistic Mine*

La fase di sfruttamento dei dati testuali che verranno stratificandosi all'interno della miniera costituisce il momento più importante di tutto il progetto. Gli strumenti già disponibili, l'ambiente PiSystem con la sua procedura di base DBT, opportunamente integrati e riadattati, costituiscono il nucleo centrale del processo di estrazione e di sintesi delle informazioni linguistiche. Già in questa prima fase possono essere ottenuti importanti risultati: per esempio, a fini lessicografici, si possono estrarre le concordanze di singole parole, di singoli lemmi, di specifiche locuzioni, le co-occorrenze strutturate, in ordine sinistro o in ordine destro, operando incroci secondo la tipologia dei testi analizzati.

Alle funzioni già disponibili per l'analisi e la navigazione nel testo, è stato aggiunto un nuovo strumento le cui specifiche sono già definite e che è in fase di sperimentazione e perfezionamento: ossia un ambiente di lavoro per la definizione di *pattern linguistici* e successiva applicazione alla miniera e/o ad un suo sottoinsieme. La Fig. 3 propone il menu di selezione, a partire da un elenco di regole precedentemente definite, e il sottoinsieme del corpus a cui applicare la regola scelta. All'interno di tale schermata sono presenti anche i menu di estrazione dati, di definizione delle regole, di fusione di più output e di stampa dei risultati.

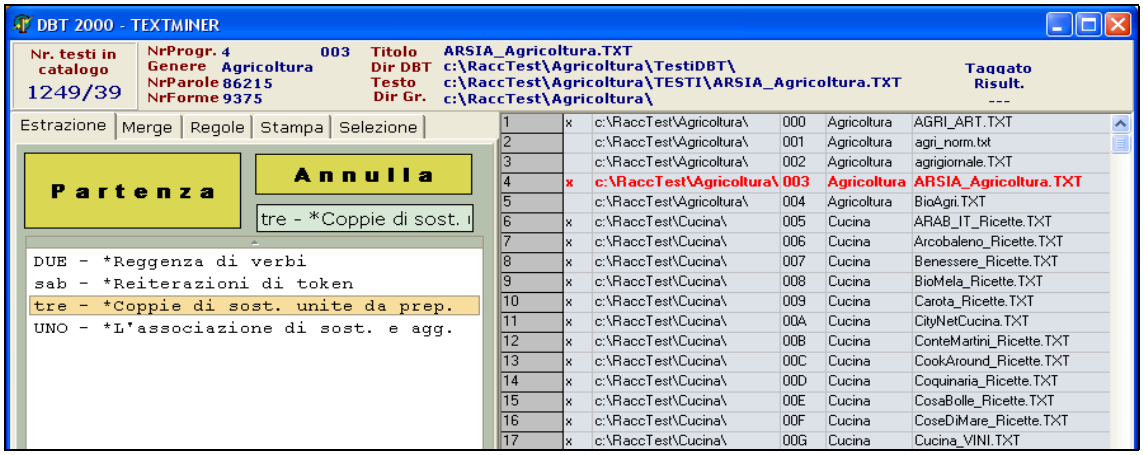


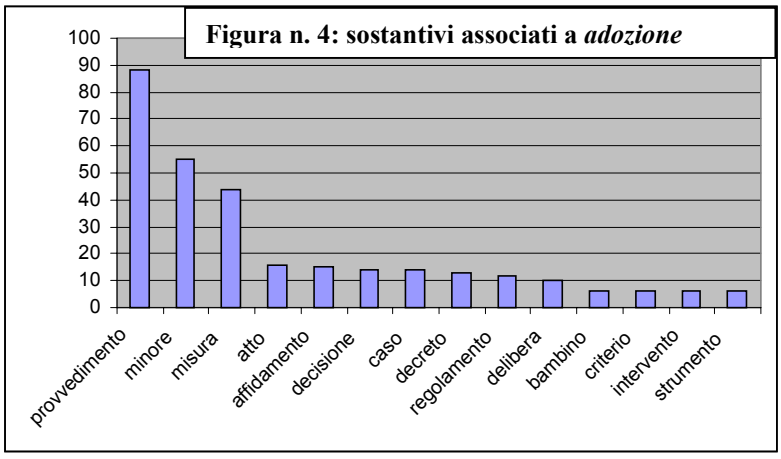
Figura 3: monitor di gestione delle regole di estrazione e di stampa dei risultati

Altre Procedure di Analisi Linguistica

Per poter analizzare in maniera più efficace il corpus, il sistema provvede alla esecuzione, in maniera automatica, della fase di *tagging*, altrimenti detta *lemmatizzazione* (Picchi 94). Tale fase viene eseguita per ciascun testo, nel momento in cui venga utilizzato per la ricerca dei pattern. Il risultato della lemmatizzazione viene poi memorizzato in modo da essere disponibile per una successiva ricerca che comprenda quel testo.

Risultati: alcuni esempi

L'estrazione di *pattern* linguistici dal corpus di riferimento costituisce un importante strumento per la analisi e la sintesi della lingua, permettendo non soltanto la ricerca di informazioni e la verifica di ipotesi linguistiche ma anche la costruzione di banche-dati di quanto ricercato, analizzato, estratto e sintetizzato. Alla base di queste considerazioni sta proprio un altro dei punti fondamentali del progetto LM che è quello di poter costruire un sistema di *repositories* di dati linguistici integrati con il compito di contenere la memoria e la sintesi di tutti i dati e le nozioni linguistiche raccolte precedentemente.



Il sistema, a regime, elabora la grande mole di testi inseriti e memorizza le varie fasi di analisi eseguite. Si avrà quindi non soltanto una grande miniera di testi ma anche una grande banca-dati di testi marcati/lemmatizzati e disponibili per altre elaborazioni che potranno essere maggiormente efficaci e produttive grazie a questo valore aggiunto associato.

Una massa di informazioni linguistiche utili, sia in fase di analisi che di sintesi, sarà

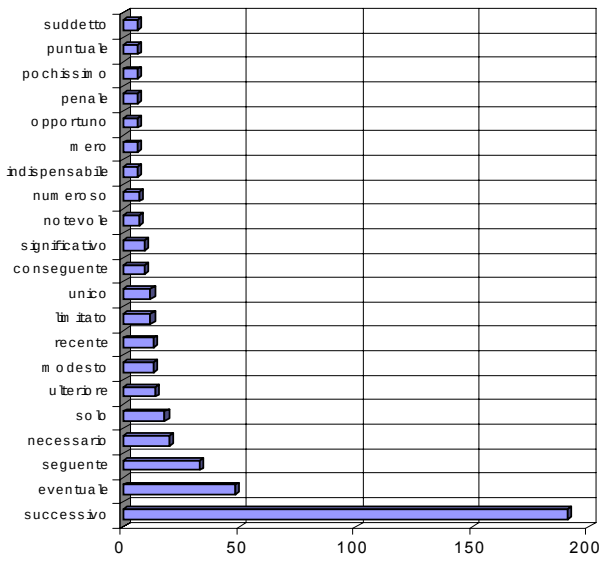


Figura n. 5: aggettivi più frequenti associati a *modifica*

disponibile per diverse tipologie di studio e di sviluppo di sistemi applicativi, il cui fondamento sarà una conoscenza linguistica della lingua italiana incredibilmente ricca e documentata dall'insieme dei testi raccolti.

Come esemplificazione mostriamo due tipi di risultati di ricerca di *pattern* linguistici in un campione di testi giuridici di circa 8 milioni di parole.

Il primo esempio è relativo alla ricerca di tutte le coppie di sostantivi collegate da preposizioni.

Nella Fig. 4 si mostrano i sostantivi più frequenti associati con una preposizione alla parola *adozione*.

Il secondo esempio è estratto da una ricerca sull'aggettivazione: per tutti i sostantivi ritrovati nei testi, vengono individuati gli aggettivi associati.

Nella Fig. 5 si propone una sintesi dell'aggettivazione della parola *modifica* nelle sue combinazioni più frequenti.

Conclusioni

Agli strumenti e alle risorse finora descritti possono essere interessati varie tipologie di utenti, quali i linguisti, per lo studio di vari aspetti della lingua, i lessicografi, strutturalmente interessati a tutte le fasi di sintesi lessicale e linguistica, i docenti per proporre nuovi percorsi didattici e più in generale tutti gli operatori della società dell'informazione. L'analisi e la disambiguazione di procedure automatiche, il miglioramento del processo di traduzione automatica possono essere, ad esempio, facilitati dalla disponibilità nella banca-dati dell'aggettivazione di ciascun termine, articolata anche in vari ambiti linguistici, che potrà permettere ad un sistema automatico o traduttore di comporre in maniera guidata, maggiormente rispondente alla consuetudine della lingua settoriale, e di trovare le migliori combinazioni di termini. Un altro risultato importante del progetto sarà la diffusione delle risorse linguistiche, grazie alla realizzazione della procedura di conversione degli archivi DBT in formato XML, universalmente riconosciuto per l'interscambio dei dati in Internet.

Bibliografia

Picchi, E. (1994). *Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian*, in Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Euralex '94 Proceedings, Papers submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, Free University of Amsterdam, The Netherlands.

Biagini, L., & Picchi, E. (1996). *INTERNET and DBT*. Gellerstam, Martin, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström and Catarina Röjder Papmehl (Eds). 47-53.

Picchi, E. (1997). *DBT 3 - Data Base Testuale: Guida all'uso*, versione 3.1. Lexis Ricerche s.r.l. su licenza del C.N.R., Roma.

Picchi, E. (1999) *Informatica e scienze umane: procedure di analisi testuale,* in MARIA ASSUNTA ZANETTI (a cura di) Parola e immagine, Facoltà di Lettere e Filosofia dell'Università di Pavia, 88, Istituto di Psicologia, La Nuova Italia Editrice, Firenze. 181-190.

Sassi, M., Ceccotti, M. L. (2000) *L' utilizzo didattico di corpora: proposte metodologiche*, Didamatica 2001, a cura di A. Andronico, A.M. Fanelli, G. Piscitelli, T. Roselli, Edizioni Laterza. 350-358.