da "Statistica & Società - SIS", anno IV numero 2 (2006)

Una strategia per l'analisi statistica del linguaggio giornalistico: sei annate di *Le Monde Diplomatique*

Domenico Attardi, Michelangelo Misuraca
Dipartimento di Matematica e Statistica, Università di Napoli "Federico II"

1. Introduzione

L'età dell'informazione, locuzione che ricorre come uno dei principali attributi dell'era moderna, è caratterizzata in modo significativo da un costante e diffuso processo di produzione di fonti testuali.

Per far fronte alle problematiche connesse all'utilizzo di questa consistente mole di informazioni, sempre più comunemente disponibile in formato elettronico, sono necessarie delle strategie di ricerca, analisi e organizzazione della conoscenza che consentano di soddisfare i bisogni informativi degli utenti finali.

Le aziende e le istituzioni hanno mostrato una crescente attenzione verso tali problematiche, data la sempre maggiore necessità di selezionare nell'enorme mole di materiale testuale disponibile i dati più interessanti e capaci di produrre valore.

2. Lo studio del *linguaggio naturale*

Il *linguaggio naturale* è un fenomeno complesso e in continua evoluzione, difficile da analizzare con procedure di tipo automatico. Lo sviluppo di metodologie per il trattamento di dati qualitativi secondo una logica di confronto e non di misura, insieme alle enormi possibilità offerte dall'informatica, ha determinato il potenziamento delle tecniche di analisi dei testi, diffondendone l'uso in contesti disciplinari molto differenziati rispetto a quelli originari.

Inizialmente, soprattutto linguisti, sociologi e psicologi erano interessati allo studio del linguaggio. Negli ultimi anni anche informatici e statistici hanno mostrato interesse per tale ambito di ricerca, ponendo una maggiore attenzione agli aspetti quantitativi.

Il progredire delle capacità informatiche ha permesso non solo di codificare e riconoscere i caratteri dell'alfabeto di una qualsiasi lingua, ma anche di trattare in maniera più semplice e veloce collezioni documentali di notevole dimensione. Sono nati software per il trattamento dei testi e per calcolare indici di vario tipo, come la ricchezza del vocabolario utilizzato o le misure lessicografiche.

Possono essere considerate come oggetto di studio le più diverse fonti testuali: i testi di natura letteraria, gli articoli di stampa periodica, le indagini qualitative con domande aperte, i messaggi pubblicitari, le trascrizioni di messaggi non testuali come i confronti televisivi (ad esempio i faccia a faccia tra uomini politici sui programmi elettorali), e così via.

Bisogna però dire che il linguaggio naturale è difficilmente definibile da un punto di vista statistico. La *parola* come unità elementare del linguaggio non è di per sé univocamente definibile: può infatti denotare un oggetto (sostantivo), un'azione o uno stato (verbo), una qualità (aggettivo, avverbio), una relazione (proposizione). Nonostante la contrarietà di alcuni linguisti, il linguaggio può essere considerato in una dimensione quantitativa, ma è indispensabile adottare delle convenzioni precise ed il più possibile coerenti da un punto di vista linguistico (De Mauro, 1995).

Gli anni Settanta segnano il punto di svolta per gli studi quantitativi della lingua. È in questo periodo che si passa da una logica di tipo linguistico ad una di tipo lessicale. Successivamente negli anni Ottanta, grazie ai contributi di Ludovic Lebart e André Salem, si inizia a porre una maggiore attenzione alla *testualità* della base dei dati analizzata.

L'interesse attuale sembra rivolto ad un approccio integrato di tipo *lessico/testuale*, in cui l'esame di una raccolta di testi d'interesse viene supportata da meta-informazioni di carattere linguistico (lessici di frequenza, grammatiche locali) e da interventi di natura differente sul testo stesso (Bolasco, 1999).

3. L'analisi automatica dei testi

L'interdisciplinarietà dello studio dei testi ha ingenerato nel corso degli anni alcune ambiguità nella definizione degli ambiti di ricerca.

Il settore cui fa riferimento l'analisi automatica dei testi viene indicato con il termine *Text Analysis*, ma sempre più spesso nel trattamento di grandi basi di dati documentali si sente parlare di *Text Mining*. In tale contenitore possono essere compresi tanto i metodi per il trattamento del linguaggio

naturale (*Natural Language Processing*) quanto i metodi statistici.

Relativamente a questi ultimi, è possibile passare da un livello di studio "unidimensionale", quale ad esempio l'*analisi delle concordanze*, a uno "multidimensionale", nel quale principalmente si considera l'*Analisi dei Dati Testuali*.

In tale ambito si utilizzano approcci fattoriali tipici per lo studio di dati non strutturati, allo scopo di identificare regolarità nei comportamenti di tipo linguistico, esplorare e visualizzare l'informazione contenuta nei *corpora*.

La metodologia maggiormente utilizzata è l'*Analisi delle Corrispondenze* (AC), proposta negli anni Settanta da J.P. Benzécri per l'analisi di tabelle di contingenza.

L'AC è una tecnica attraverso la quale è possibile descrivere da un punto di vista sia geometrico che algebrico le relazioni tra le distribuzioni, espresse in forma matriciale, delle modalità di due o più caratteri in un set di unità statistiche. Questa metodologia si è rivelata molto utile per ricavare induttivamente alcune regolarità linguistiche. Infatti, non ci si limita a mettere in evidenza le parole più ricorrenti all'interno di un *corpus*, ma si analizza l'associazione presente all'interno della tabella, arrivando a proiettare le parole su un piano fattoriale per determinare dei profili lessicali specifici e studiare le similarità tra di essi (Lebart et al., 1998).

Anche se si utilizzano tecniche d'analisi di tipo quantitativo su dati tipicamente qualitativi (quali quelli testuali) risulta comunque opportuno il ricorso a strumenti interpretativi derivati da altre discipline che consentono una più ampia prospettiva del fenomeno linguistico indagato.

4. Concetti e definizioni preliminari

Nell'effettuare un'analisi automatica dei testi bisogna considerare alcune operazioni per trasformare l'informazione testuale in *dato*: si parla in generale di *pre-trattamento* del testo[1].

La *numerizzazione* del *corpus* consiste in una lettura automatica del testo per associare alle diverse parole il numero di volte che si presentano, ottenendo così il "vocabolario" delle *forme grafiche* (ossia le parole così come sono scritte nel testo).

Con la *normalizzazione* si rendono omogenee le grafie utilizzate, per evitare sdoppiamenti nel dato, eliminando le differenze tra caratteri minuscoli e maiuscoli, uniformando nomi propri, sigle, date e così via.

La *lessicalizzazione* unisce due o più forme in una sola, identificando i segmenti ripetuti più significativi (es. primo_ministro) e al contempo eliminando i casi più banali di ambiguità lessicale.

Il *tagging grammaticale* identifica la "parte del discorso" associata alla forma, in termini di categoria grammaticale (es. partito_V, casa_N), e quindi prepara il *corpus* alla *lemmatizzazione*, che riconduce le parole al lemma presente nel dizionario della lingua (sostantivi al singolare, aggettivi al maschile singolare, verbi all'infinito). Alla fine di queste operazioni si ottiene un vocabolario modificato, contenente voci meno ambigue rispetto alle forme iniziali, costituito da unità minimali di senso non ulteriormente decomponibili (Reinert, 1988).

Una delle criticità dell'analisi statistica dei testi riguarda la necessità di selezionare le parole presenti nel *corpus* a maggior contenuto informativo, rispetto al tipo di conoscenza che si vuole evidenziare.

La selezione delle unità tipiche o caratteristiche (*parole chiave*) non può essere basata solo sul criterio della loro frequenza, in quanto può accadere che forme che appaiono una sola volta possono risultare significative. Si può in alcuni casi ricorrere al calcolo di indici mutuati da altre discipline, quale il *Term Frequency/Inverse Document Frequency* (TFIDF) che considera non solo l'importanza relativa di una parola in un testo, ma anche il suo "potere di discriminazione" rispetto all'intero *corpus* (Balbi e Misuraca, 2005).

5. Il linguaggio giornalistico

I diversi tipi di linguaggio possono essere ricondotti ad un sistema di codici e regole utilizzato da specifiche classi di utenti.

Focalizzando l'attenzione sulle cosiddette lingue speciali è possibile riconoscere principalmente due sottoinsiemi: le *lingue specialistiche* delle discipline a specializzazione avanzata (es. le scienze, l'informatica, la politologia, e così via) e le *lingue settoriali* di ambiti meno specialistici o comunque dirette ad un pubblico più eterogeneo (il linguaggio

[1] Tra i diversi *software* utilizzati si ricorda TALTAC, una libreria di metodi che consente il trattamento e l'analisi lessicografica di un'insieme di dati testuali.

della moda, della pubblicità, e soprattutto quello burocratico).

Il linguaggio utilizzato dalla stampa, per la funzione sociale e civile di quest'ultima, è stato nel corso degli ultimi anni più volte investigato, data anche la sempre maggiore disponibilità di fonti e di strumenti avanzati.

Nel linguaggio *giornalistico* è possibile individuare tanto codici diversi provenienti da settori specifici (politico, tecnico-scientifico, economico-finanziario, e così via) quanto vari registri riconducibili ai diversi livelli della lingua parlata (aulico, colto, ufficiale, colloquiale).

È quindi difficile identificare i differenti modi di creazione di un linguaggio così articolato, infarcito di neologismi, di stereotipi e vocaboli mutuati da altri linguaggi.

La diffusione di Internet ha portato ad un'ulteriore evoluzione del linguaggio dell'*Informazione* rispetto al tradizionale supporto cartaceo.

A differenza dei cosiddetti *media generalisti*, è possibile diversificare il contenuto informativo, in relazione ai diversi bisogni di conoscenza degli utenti. Il vantaggio è quello di poter confrontare rapidamente come una stessa notizia è stata riportata da fonti differenti e facilitare al contempo il lavoro di rassegna per la presenza di archivi (più o meno completi) delle annate precedenti. Il risultato è la possibilità pressoché illimitata di accedere in tempo reale ad una *emeroteca* virtuale completa e personalizzabile.

6. *Le Monde Diplomatique* in Italia

L'applicazione presentata di seguito si riferisce ad uno studio degli articoli pubblicati dal 1998 al 2003 nell'edizione italiana del periodico francese *Le Monde Diplomatique* (LMD)[2] fondato nel 1954 e pubblicato in 20 lingue e in 30 paesi differenti. Lo scopo prefissato è quello di esaminare il linguaggio utilizzato nel descrivere gli avvenimenti di attualità degli anni a cavallo tra la fine del XX secolo e l'inizio del XXI. Il giornale si è contraddistinto negli anni per il suo punto di vista critico e di respiro internazionale, rispetto ad argomenti politici, economici, sociali e culturali. In Italia, a partire dal 1995, è regolarmente pubblicato come supplemento mensile al quotidiano *Il manifesto*.

Il linguaggio utilizzato è sufficientemente omogeneo, anche se spesso le stesse parole sono tradotte in maniera differente dai diversi traduttori presenti nella redazione. Dalle circa 2000 pagine iniziali sono stati selezionati manualmente e convertiti in formato testo, 1914 articoli. Per ciascuno di essi si è considerato nell'analisi solo il corpo, escludendo titolo e occhiello.

Sulla base di conoscenza esperta ogni articolo è stato classificato in 32 categorie, considerando l'argomento principale trattato.

La base di dati così ottenuta contiene più di 3.000.000 di parole ed è stata attentamente pre-trattata secondo le procedure già viste in precedenza. Dopo aver effettuato una lessicalizzazione sufficientemente approfondita è stato ottenuto un vocabolario di circa 77.000 forme testuali, da cui sono state poi eliminate le *forme strumentali*, utili a discernere il senso generale del fenomeno analizzato ma non interessanti ai fini dell'analisi.

I dati così ottenuti sono stati strutturati in una particolare tabella di contingenza, detta *lessicale*, che ha in riga le circa 5.000 forme selezionate ed in colonna i 32 argomenti. Applicando una *Analisi delle Corrispondenze Lessicali* è stato possibile ottenere una rappresentazione grafica delle relazioni tra le forme e le categorie di argomenti.

L'interpretazione di tale grafico è effettuata seguendo alcune regole di lettura:

- la dispersione delle forme e delle categorie intorno all'origine degli assi mostra la forza dell'associazione nella tabella;
- se due forme sono vicine allora sono utilizzate in maniera simile. Allo stesso modo, se due argomenti sono vicini utilizzano un vocabolario simile;
- la prossimità di una singola forma ad un argomento (o viceversa) non è letta in modo diretto ma valutata in riferimento all'intera nube degli argomenti (o delle forme);
- ogni forma e ogni categoria contribuisce alla determinazione dell'asse fattoriale. Le forme e gli argomenti che maggiormente contribuiscono alla formazione di un'asse consentono di spiegarlo ed eventualmente etichettarlo.

Di seguito sono riportate la rappresentazione degli argomenti (Fig. 1) e la rappresentazione delle forme (Fig. 2).

---

[2] Le sei annate complete del giornale sono tratte dal sito web <*http://www.ilmanifesto.it/MondeDiplo*>.

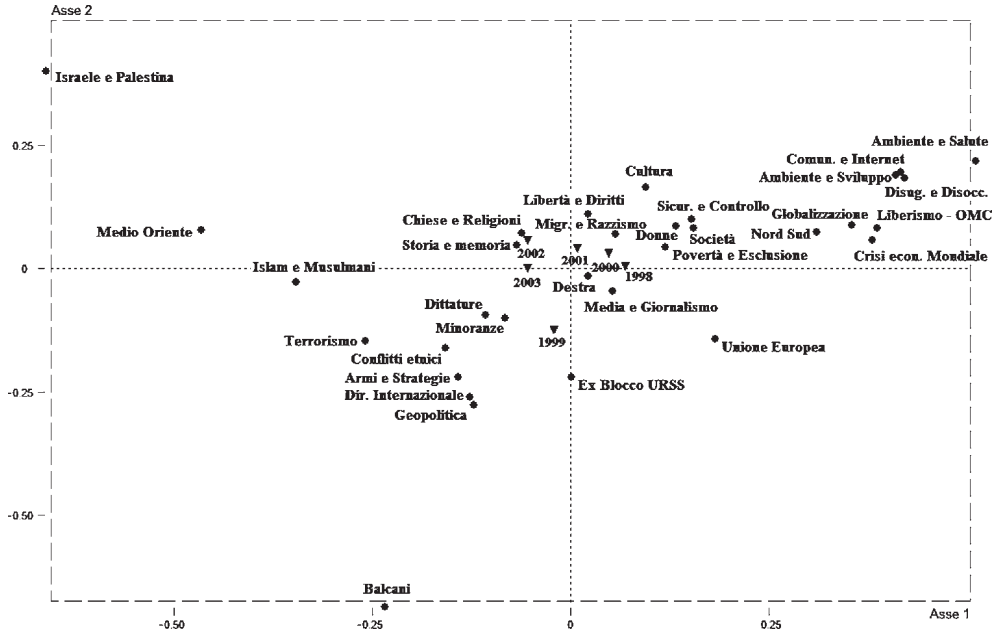

Fig. 1: Rappresentazione fattoriale degli argomenti (1º e 2º asse).

Nella figura 1 si evidenzia una contrapposizione tra le tematiche relative alla tensione internazionale e alla guerra al terrorismo (in basso a sinistra) e quelle relative a problematiche economiche e sociali (in alto a destra). Le annate di pubblicazione sono state proiettate come punti supplementari.

Nella figura 2 sono state riportate le forme che maggiormente caratterizzano gli assi fattoriali. Nella sezione inferiore del grafico è possibile individuare le forme legate ai conflitti che hanno interessato l'area balcanica sul finire degli anni Novanta (*Nato*, *Rambuillet*, *Uck*, *Bosnia*, *pulizia_etnica*).

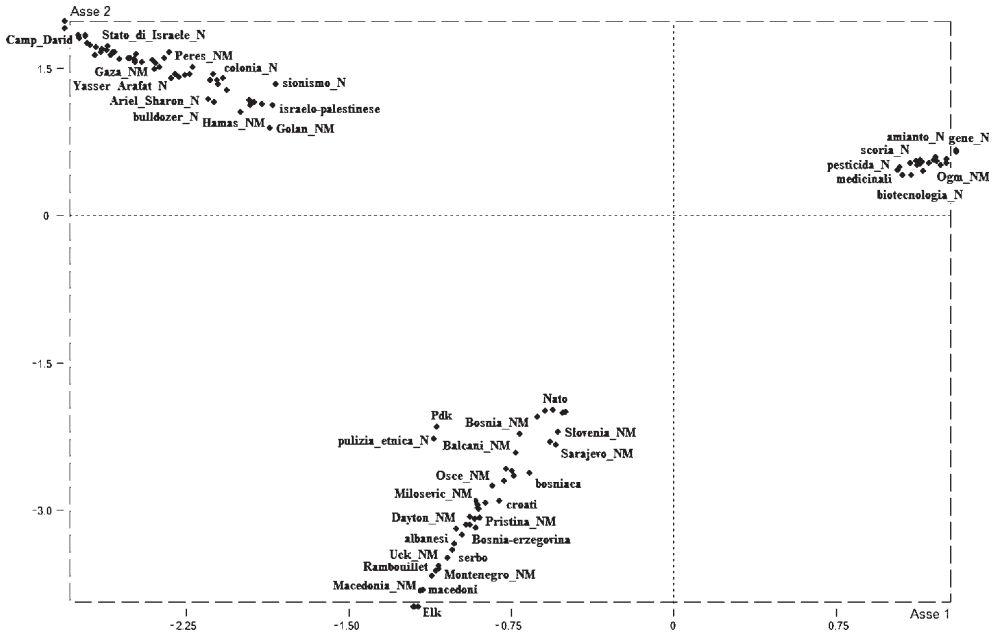

Fig. 2: Rappresentazione fattoriale delle forme testuali (1º e 2º asse).

In alto a sinistra si trovano le forme relative alle tensioni in Medio Oriente, ed in modo particolare del contrasto tra israeliani e palestinesi (*coloni*, *Hamas*, *bulldozer*, *Golan*). In alto a destra vi sono invece le forme caratteristiche delle problematiche connesse all'ambiente e alla salute (*amianto*, *Ogm*, *biotecnologia*).

Le forme più prossime all'origine degli assi non sono state visualizzate perché più diffusamente utilizzate negli articoli.

Tale scelta è motivata dalla particolare metrica euclidea ponderata utilizzata nell'AC (*Chi-quadro*), che assegna alle forme rare, ossia quelle con un minor numero di occorrenze nel *corpus*, una "importanza" eccessiva. Ciò se da un lato mette in evidenza le forme con un uso peculiare o che caratterizzano in modo particolare un documento o un gruppo di documenti, rischia al contempo d'indurre una distorsione nella fase interpretazione dei risultati ottenuti.

In taluni casi, a seconda dell'obiettivo dell'analisi e del bisogno informativo connesso, può risultare utile il passaggio da un approccio simmetrico, come nello schema classico dell'AC ad un approccio non simmetrico (Balbi, 1995).

7. Conclusioni

L'analisi presentata ha evidenziato come sia possibile studiare l'evoluzione e l'utilizzo di un certo linguaggio specialistico, nell'ottica di un approccio esplorativo quale quello della Statistica Testuale, basato su tecniche fattoriali proprie dell'Analisi dei Dati.

Tale strategia può essere vista come una fase di un processo integrato di Text Mining, nel quale si voglia non solo estrarre conoscenza ma anche procedere ad una sua "sistematizzazione". I termini che caratterizzano maggiormente certi argomenti possono infatti essere utilizzati come *parole chiave* per migliorare il processo di classificazione degli articoli e definire così procedure automatiche che limitino la necessità di dover ricorrere a conoscenza esperta.

Bibliografia

Balbi S., (1995), "Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms", in S. Bolasco et al. (Eds.), *Actes des 3es Journées internationales d'Analyse statistique des Données Textuelles*. CISU, Roma, 2, 5-12.

Balbi S., Misuraca M., (2005), "Visualization techniques for non symmetrical relationships". In: S. Sirmakessis (Ed.) *Knowledge Mining (Studies in Fuzziness and Soft Computing)*. Springer, Heidelberg. (in corso di stampa)

Bolasco S., (1999), *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*. Carocci Editore, Roma.

De Mauro T., (1995), "Quantità – qualità: un binomio indispensabile per comprendere il linguaggio". In: R. Cipriani, S. Bolasco (Eds.) *Ricerca qualitativa e computer – Teorie, metodi e applicazioni*. Franco Angeli, Milano, 21-30.

Lebart L., Salem A., Berry L., (1998), *Exploring textual data*, Kluwer Academic Publisher, Dordrecht.

Reinert M., (1988), "Un logiciel d'analyse des données textuelles: Alceste". In: E. Diday (Ed.) *Data Analysis and Informatics*. NH, Amsterdam, 5, 469-476.