Aaron Visaggio

La linguistica computazionale nell'analisi automatica dei contenuti nei social media

Summary: The computational linguistic applied to the automatic analysis of social media's contents

Computational linguistics can be defined as the meeting point between the theoretical linguistics and the information technologies. This paper aims at providing an in-depth analysis of the tools of the computational linguistics and discusses how they can support and automatate the analysis of the discourse.
A specific attention will be paid to those features that allow language specialists to have access to those characteristics of the language which cannot be reached otherwise. In order to pursue this goal, the computational linguistics has developed tools of automatic and semi-automatic analysis of natural language which avoid a manual analysis of the text to the language specialist.
In this work the main software applications of computational linguistics will be discussed as well as: text mining, information retrieval, text indexing, sentiment analysis, and opinion mining.

Keywords: *Computational Linguistics, Discourse Analysis, Text Mining, Sentiment Analysis.*

1. Introduzione

Fin dalla formulazione del test di Turing, è stato chiaro che i rapporti tra la linguistica e l'informatica dovevano essere estremamente stretti. Ed infatti nella seconda metà del Novecento è stato evidente, oltre che fruttuoso, l'interesse che tutti coloro che analizzavano il linguaggio hanno rivolto all'informatica, vedendola come promettente strumento di analisi e sperimentazione. L'informatica ha incontrato gli strumenti di analisi del linguaggio solo dagli anni 50 in poi, essendo prima stata considerata come la disciplina che si occupa della raccolta e del trattamento delle informazioni o, più specificatamente, dell'elaborazione di dati per mezzo di calcolatori elettronici. Negli ultimi anni l'integrazione degli strumenti informatici con quelli linguistici e letterari, è stata così intensa da favorire una fortissima specializzazione e settorializzazione di questa branca dell'informatica, definendola come una disciplina a sé stante che prende il nome di "linguistica computazionale". La linguistica computazionale, può essere definita come il luogo di incontro tra linguistica teorica e tecnologie informatiche. L'informatica, la statistica, la matematica, e l'intelligenza artificiale forniscono strumenti e metodi per le analisi linguistiche e per le loro applicazioni. Le applicazioni computazionali prevedono due tipologie di obiettivi:

- Lo sviluppo di strumenti informatici per lo studio e la ricerca specialistica delle lingue;
- Lo sviluppo di applicazioni informatiche, ossia software che sfruttano le competenze linguistiche per produrre programmi di utilità generale.

Questo lavoro si pone l'obiettivo di introdurre gli strumenti più noti della linguistica computazionale ed esaminare in che modo questi possono contribuire ad automatizzare l'analisi del discorso. Si presterà particolarmente attenzione alle applicazioni informatiche destinate a specialisti del linguaggio, che attraverso tali strumenti, tentano di portare alla luce caratteristiche delle lingue altrimenti non rilevabili. Di tali strumenti si serve la *linguistica dei corpora*, che esamina grandi quantità di produzioni linguistiche, osservandone le caratteristiche quali: il lessico, la sintassi, la struttura morfologica. La linguistica computazionale, per favorire tale studio, ha sviluppato strumenti informatici di analisi automatica o semi-automatica dei testi scritti in linguaggio naturale che evitano al linguista di analizzare e cercare i dati linguistici manualmente. In questo lavoro saranno esposte le principali applicazioni informatiche legate a questo ramo di studi come il Text Mining, Information Retrieval, Text Indexing, Sentiment Analysis e Opinion Mining.

2. NLP Natural Language Processing

Il Natural Language Processing (NLP) o Trattamento Automatico del Linguaggio è una disciplina che ha dato vita ad un'ampia gamma di tecnologie. Esso ha lo scopo di sviluppare strumenti automatici per l'elaborazione di testi scritti in linguaggio naturale, potendo così realizzare analisi variegate ed estrarre proprietà che sarebbe oneroso se realizzate da esseri umani. Esso si compone di due marco-componenti, come illustrato in Figura 1.

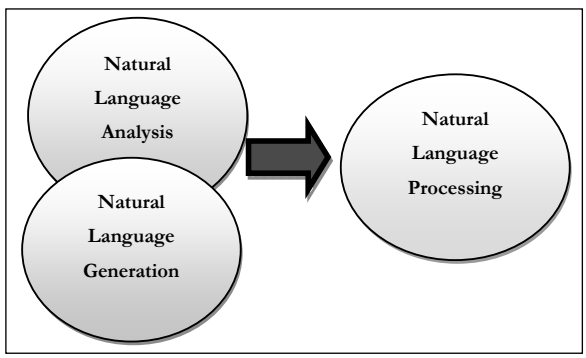

Fig. 1. Sottocategorie NLP.

- Natural Language Analysis: data una frase ha l'obiettivo di darne una rappresentazione della sua analisi.
- Natural Language Generation: data una grammatica di una lingua, ha lo scopo di produrne frasi in senso compiuto.

Il Natural Language Processing è il processo di estrazione di proprietà e caratteristiche semantiche da espressioni del linguaggio umano, tramite l'elaborazione da parte di un calcolatore elettronico. Data la sua complessità, il processo di elaborazione si suddivide in tre fasi fondamentali:

1. Analisi lessicale: si occupa della suddivisione di un'espressione linguistica in token che nell'ambito del linguaggio scritto sono identificati da entità come simboli, punteggiature, parole, ecc.
2. Analisi Sintattica: in questa fase viene analizzata la sequenza di token in modo da ottenere la struttura sintattica.
3. Analisi semantica: questa fase si occupa dell'assegnazione del significato alla struttura sintattica.

Queste fasi, per quanto riguarda lo studio informatico, sono divise nei passi illustrati di seguito.

Tokenizzazione

Tramite questa operazione il testo viene suddiviso in token; un token per la lingua italiana può essere definito come una sequenza di caratteri delimitata dagli spazi, con l'eccezione per i segni di punteggiatura che sono uniti alle parole, come l'apostrofo che compare tra due parole diverse.

Part of Speech Tagging

Part of Speech Tagging (PoS Tagging) è il processo di associazione di etichette alle parole del testo che corrispondono a particolari categorie lessicali. Esistono due approcci nell'ambito del NLP per implementare il PoS Tagging, il primo si basa sulla definizione di regole per l'identificazione di singole classi, il secondo si basa su tecniche di machine learning. Questi ultimi, adoperati in letteratura, utilizzano un approccio di addestramento supervisionato, ma per funzionare necessitano di insiemi di documenti annotati detti corpora. Molti strumenti di tipo statistico fanno uso di corpora per ottenere funzioni di probabilità e regole a supporto del PoS Tagging. Questo tipo di approccio presenta delle difficoltà quando si devono classificare parole sconosciute, cioè che non erano presenti nel training set e per cui non si hanno dati statistici sulla loro appartenenza a determinate classi.

Chunking

Il chunking è l'analisi di una proposizione, che è formata in forma semplice da un soggetto e un predicato. Il soggetto è tipicamente un "sintagma nominale", mentre il predicato è un "sintagma verbale" costituito da un verbo insieme a zero o più complementi e avverbi. Un Chunk è formato da uno o più token adiacenti.

Clause Identification

Nella elaborazione del linguaggio naturale, il Clause Idetification suddivide una frase in proposizioni, dette "clause". Un'estensione del Clause Identification, permette prima di isolare le proposizione, e poi di attribuire ad esse il proprio ruolo (principale, subordinata relativa, subordinata finale, ecc.).

Logic Analysis

L'analisi logica permette di assegnare ai singoli componenti di una frase, le proprie categorie logiche e permette di distinguere tra Soggetto, complemento e predicato nominale e o verbale.

Anaphora Resolution

Anaphora Resolution o risoluzione delle anafore consiste nel trovare a cosa si riferisce un'anafora. In linguistica un'anafora è una istanza

di un'espressione che si riferisce ad un'altra. Un esempio di anafora è presente nella frase: "Maria ha letto il libro che ha comprato ieri". In questa frase esiste un'anafora che fa corrispondere la parola *che* a *libro*.

Keyphrasing

Keyphrasing è l'identificazione delle parole chiavi all'interno di un documento, come titoli, sottotitoli, ecc.

Lemmatization

Lemmatization o lemmatizzazione è l'operazione che riconosce il lemma di una parola a partire dalla sua forma flessa. Il lemma è la forma di citazione di una parola. Ad esempio la forma flessa "fiori" avrà come lemma "fiore". Per i verbi invece, il lemma è la forma all'infinito del verbo.

Word Sense Disambiguation

Esiste un numero elevato di parole che assumono un significato diverso in base al contesto nel quale sono adoperate. Il compito del Word Sense Disambiguation è quello di attribuire a queste parole il proprio senso specifico.

Named Entity Recognition

Il named Entity Recognition consiste nella classificazione e nell'individuazione di stringhe di testo che si riferiscono a persone, organizzazioni, luoghi, date e tempi.

Full Coreference Resolution

Il full Coreference Resolution consente di determinare se due espressioni nel linguaggio naturale si riferiscono alla stessa entità. Ad esempio: "Presidente della Repubblica Italiana" → "Giorgio Napolitano".

In Figura 2 il processo di NLP è indicato nei suoi componenti principali.

2.1. Text Mining

Il Text Mining (Feldman e Sanger, 2007) può essere definito come l'analisi delle informazioni contenute in documenti scritti attraverso i metodi di Data Mining. Attraverso l'applicazione di alcuni strumenti di analisi si cerca di identificare ed esplorare schemi rilevanti all'interno dei documenti obiettivo. Con il termine Data Mining ci si riferisce ad un insieme di tecniche e metodologie che hanno come obiettivo l'estrazione di informazioni sconosciute a partire da ingenti volumi di dati. La differenza principale tra il Text Mining e il Data Mining è che il primo analizza testi, che sono generalmente non strutturati, mentre il secondo, analizza dati generalmente strutturati in database relazionali. Molta parte del processo di sviluppo di applicazioni di Text Mining coincide con il processo di Data Mining, ma si differenziano sensibilmente nel preprocessing. Mentre nel Data Mining questa fase è orientata alla selezione dei dati da usare e alla loro normalizzazione, nel Text Mining si pone l'attenzione sull'estrazione di caratteristiche rappresentative del documento. Nella pratica si cerca di trasformare i dati (testo) non strutturati in dati strutturati e, a tal fine, nel Text Mining si utilizzano molte tecniche prese da altre discipline come quello della ricerca di informazioni, l'estrazione di informazioni e di Natural Processing Language (NLP). Il processo di sviluppo di Text Mining si suddivide nei seguenti passi: parsing, pattern recognition, analisi sintattica e semantica, clustering. Le particolarità del linguaggio naturale con cui sono scritti i documenti e la sua naturale ambiguità, presentano alcuni problemi per il Text Mining. Le questioni che devono essere prese in considerazione durante questo processo sono le seguenti:

1. Stop List: Utilizzare una stop list che contiene parole con alta frequenza o che devono essere ignorate. Questa pratica non è molto utile nel Text Mining dal momento che le parole comuni possono essere utili per comprendere la semantica di un segmento di testo.
2. Lemmatizzazione: questa tecnica consente di invalidare la semantica di una porzione di testo.
3. Dati inesatti: La correzione di errori di scrittura, la sostituzione di acronimi e abbreviazione può essere parte di un processo di Text Mining.
4. Disambiguazione di senso delle parole: esistono due tipi di approcci, supervisionato

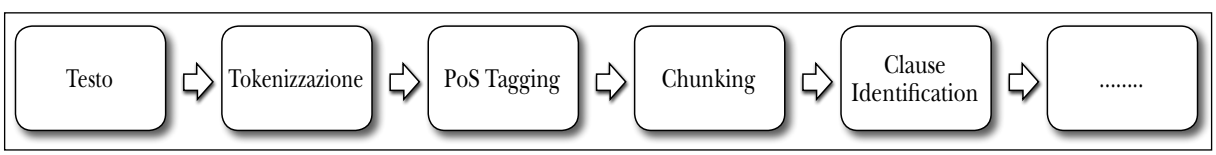

Fig. 2. Processo NLP.

e non. Quello supervisionato è spesso associato all'utilizzo di dizionari. Nell'approccio non supervisionato i differenti significati delle parole non sono conosciuti.

5. Tagging: se lo scopo dell'analisi è l'estrazione di informazioni dettagliate, è opportuno operare un'analisi grammaticale dei token. Dopo la suddivisione del testo in frasi, si può procedere con un'assegnazione del ruolo grammaticale (PoS), attribuendo ad ogni token il ruolo di nome, verbo, aggettivo ecc. Il Penn Treebank e il Turin University Treebank (TUT) sono, per esempio, collezioni di testi annotate attraverso tag che esprimono relazioni grammaticali.
6. Locuzioni: Esistono gruppi di parole che hanno senso solo se considerate nel loro insieme. La semantica di una locuzione non è uguale alla semantica delle sue parti, quindi è necessario che l'analisi sintattica gestisca le locuzioni che compaiono nel testo.
7. Analisi sintattica: l'ordine in cui le parole sono presenti nel testo deve essere preso in considerazione.
8. Tokenizzazione: Un testo può essere diviso in paragrafi, periodi, frasi di ogni lunghezza.

2.2. Information Retrieval

L'information Retrieval (Manning et al., 2008) è una disciplina scientifica con esperienza in materia di accesso e gestione dei contenuti. Nel 1968 è stata definita nel seguente modo:

> "*L'information Retrieval è un campo relativo alla struttura, analisi, organizzazione, memorizzazione, ricerca e recupero di informazione*"[1].

Una definizione più recente di Information Retrievel parla di: "ricerca di materiale di natura non strutturata presente all'interno di una vasta collezione che soddisfa un bisogno informativo". Negli anni si è assistito allo sviluppo di un Information Retrieval più sofisticato, rendendo i sistemi più sensibili al significato delle parole, considerando l'ordinamento delle parole nell'interrogazione, tenendo conto di eventuali informazioni sul feedback dell'utente, valutando l'affidabilità delle fonti, eseguendo operazioni sui testi come indicizzazione, lo stemming e la lemmatizzazione e categorizzando automaticamente i documenti. L'information Retrieval permette di individuare, dato un insieme ampio di documenti, solo quei documenti che soddisfano i nostri criteri di ricerca. Un'applicazione pratica delle tecniche di IR su collezioni di testi è il motore di ricerca.

2.3. Text Indexing

L'indicizzazione, nei motori di ricerca è il processo di raccolta, parsing, e immagazzinamento di documenti di testo. La progettazione di metodi di indicizzazione è un problema interdisciplinare, poiché coinvolge aspetti legati non solo all'informatica in generale, ma anche alla linguistica o alla psicologia cognitiva. I motori di ricerca più comuni si concentrano sull'indicizzazione della componente testuale dei documenti, ma è possibile anche l'indicizzazione di immagini e audio tramite il pattern recognition. L'utilizzo di un indice consente di recuperare documenti rilevanti all'interno di una collezione con un notevole risparmio di tempo, altrimenti, per recuperare i documenti sulla base di una query occorrerebbe una scansione completa di tutta la collezione. Nella realizzazione di un processo di indicizzazione bisogna tener conto di alcuni aspetti legati al suo utilizzo pratico, come ad esempio:

- Criteri di merging: bisogna considerare come le informazioni vengono aggiunte all'indice di volta in volta.
- Metodi di memorizzazione: scelta della struttura dati utile per l'immagazzinamento dell'indice (RAM/disco).
- Velocità di lookup: quanto velocemente un termine può essere reperito, aggiornato, o rimosso all'interno dell'indice.
- Robustezza ai guasti: è molto importante considerare quanto un sistema di indicizzazione è in grado di tollerare dei problemi al livello dell'hardware sottostante.

Il metodo più utilizzato per indicizzare grosse collezioni di documenti è l'indice inverso. Si fornisce, per ogni elemento x del vocabolario V considerato, l'insieme delle occorrenze x nella collezione di documenti. I termini x rappresentano item lessicali, cioè parole singole o espressioni composte che si decide di trattare come blocchi unici. Il vocabolario V è ciò che rimane dopo la fase di preprocessing che si occupa di rimuovere le stop words, stemming/lemmatizzazione, ecc.

Le fasi del processo di indicizzazione automatica che devono essere attuate in sequenza sono (Figura 3):

1. Analisi lessicale e selezione delle parole
2. Eliminazione delle stop-words
3. Riduzione delle parole originali alle rispettive radici semantiche
4. Eventuale pesatura degli elementi dell'indice (significatività)
5. Creazione dell'indice.

L'indicizzazione automatica serve, quindi, principalmente per produrre delle analisi rapide dei

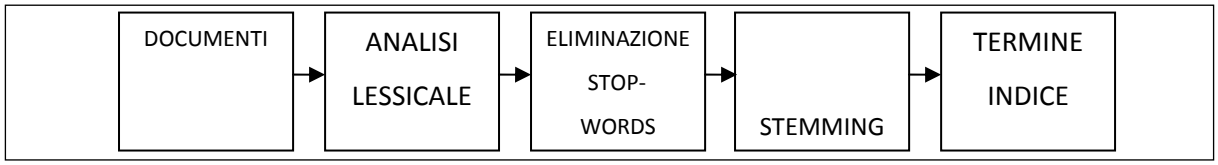

Fig. 3. Processo di indicizzazione.

testi raccolti, per esempio attraverso la individuazione di parole chiave di un testo.

2.4. Latent Dirichlet Allocation (LDA)

LDA (Blei et al. 2003) è un modello con variabili latenti (variabili delle quali non si è in grado di rilevare le realizzazioni), secondo cui le parole di ogni documento sono realizzazioni di misture di argomenti (topic). In particolare, ogni argomento si definisce attraverso una distribuzione di probabilità di tipo multinomiale su un vocabolario prefissato (l'insieme delle parole utilizzabili nel corpus). Ogni argomento è compatibile con ogni documento del corpus cioè, ogni documento può trattare di qualsiasi argomento, ma la sua frequenza all'interno del documento varia stocasticamente tra i documenti, poiché tale frequenza – che viene vista come un vettore in cui l'i-esimo elemento rappresenta la rilevanza del topic i all'interno del documento – si assume essere realizzazione di una variabile casuale di Dirichlet. Questo modello non consente di prevedere correlazione tra gli argomenti.

2.5. Opinion Mining e Sentiment Analysis

La possibilità degli utenti di esprimere attraverso il web opinioni che riguardano qualsiasi argomento (servizi, prodotti, politica, ecc.) ha portato alla nascita di strumenti informatici capaci di poter analizzare il Sentiment che un documento esprime. Il Sentiment è definito come il sottoinsieme di termini soggettivi che possono avere orientamento positivo, neutro o negativo. L'analisi automatica di opinioni, il sentiment e la soggettività di un determinato testo sono noti, rispettivamente come Opinion Mining, Sentiment Analysis e Subjectivity Analysis.

L'opinion Mining (OM) è stato introdotto nel 2003 in una pubblicazione di Dave (Dave et al., 2003) per descrivere il processo di estrazione automatica di opinioni dal testo. Questa disciplina si concentra sulla individuazione e caratterizzazione dell'opinione che il documento stesso esprime e non sull'argomento di cui parla.

Il termine Sentiment Analysis (SA) (Nasukawa e Yi, 2003) è stato introdotto al fine di descrivere il processo in grado di valutare automaticamente la polarità espressa da una serie di documenti forniti. Mentre la Opinion Mining si focalizza sul riconoscimento di opinioni espresse in un dato testo rispetto ad attributi specifici, la Sentiment Analysis, invece, è focalizzata sulla classificazione di determinati documenti in accordo con la polarità che essi esprimono (positiva o negativa). La SA è stata oggetto ed è tuttora di un ampio interesse di ricerca e ha portato alla nascita di svariati rami della medesima materia, ognuno dei quali si occupadi affrontare un determinato problema:

- Classificazione delle opinioni: La Sentiment Analysis si occupa della classificazione dei testi secondo due livelli di dettaglio.
  - Classificazione a livello di documento, che consiste nell'assegnare una categoria positiva o negativa a un testo che esprime opinioni.
  - Classificazione a livello di frase, consiste nella classificazione di frasi in soggettive e in oggettive e successivamente, per le soggettive, in positive, negative o neutrali.
- Sentiment Analysis basata su feature: La classificazione a livello di documento o di frase fornisce un'indicazione sul giudizio degli utenti ma non è in grado di esprimere informazioni dettagliate. Infatti, una frase con polarità negativa non è necessariamente indice di una visione totalmente negativa. Questo problema può essere scomposto in due fasi: identificare l'oggetto del discorso (la feature) e determinare la polarità dell'opinione espressa verso quella feature. L'oggetto è un insieme di componenti o di attributi. Formalmente un oggetto O è un'entità associata con una coppia (T,A), dove T è una gerarchia o tassonomia dei componenti e A è un insieme di attributi di O. Si utilizza il termine feature per identificare componenti e attributi senza fare alcuna distinzione.
- Frasi comparative: Sono frasi che esprimono una relazione tra uno o più soggetti, generalmente con aggettivi, o avverbi, comparativi o superlativi. Queste frasi devono essere trattate in maniera differente dalle altre frasi in

quanto hanno una forma sintattica e semantica differenti.

- False opinioni: L'identificazione di frasi pubblicitarie e false opinioni, cioè messaggi pubblicitari camuffati in modo da essere interpretati come vere e proprie opinioni, è una sfida per il futuro poiché attualmente non esistono algoritmi idonei per risolvere questo tipo di problema.

Le dimensioni che riguardano le opinioni definiscono gli aspetti della soggettività sul quale è possibile concentrare il lavoro. Tre dimensioni relative all'opinione sono state principalmente studiate in letteratura:

– Soggettività. l'analisi di soggettività è l'attività volta a riconoscere se una determinata entità testuale contiene espressioni soggettive. In pratica questo tipo di analisi ha il compito di individuare quali sono le frasi che contengono opinioni e quali, invece, forniscono solo informazioni oggettive.
– Polarità. l'analisi di polarità è volta a stabilire la polarità (in termini di orientamento positivo o negativo) espressa da un soggetto testuale. In pratica questo tipo di attività è volta a identificare quali sono le opinioni positive e quali quelle negative.
– Forza. l'analisi di forza è orientata al riconoscimento dell'intensità degli elementi soggettivi contenuti in una determinata entità testuale. Fondamentalmente l'attività di analisi della forza è effettuata per identificare l'intensità con cui è espressa un'opinione in modo da collocare tale opinione in una scala di valori. Si prendano ad esempio le due frasi: "Il governo è un po' instabile" e "Il governo è terribilmente instabile"; è ovvio che la seconda frase esprime una polarità negativa maggiore rispetto alla prima e, quindi, un orientamento più chiaro. L'analisi di forza è, dunque, utilizzata per rispecchiare questa classificazione. L'analisi di forza permette di effettuare una comparazione tra diverse opinioni e quindi stabilire quale tra queste ha la maggiore positività o negatività.

2.6. Matrici di co-occorenza

All'interno di un testo, le parole non compaiono isolate, ma compaiono in un contesto di altre parole. Il contesto precedente e successivo di una data parola che si esamina darà informazioni linguistiche determinanti sugli usi della parola, e consente di individuare sequenze di parole che occorrono con maggiore abitualità. Per estrapolare da un testo parole e i rispettivi co-testi si utilizzano strumenti specifici: le concordanze. Le concordanze sono delle presentazioni dei dati di un testo, con l'indicazione della frequenza con la quale la parola occorre e il contesto precedente e successivo (co-testo). Le principali funzioni delle concordanze sono:

– Osservare i diversi usi di una parola.
– Esaminare i diversi contesti (semantici o sintattici) in cui occorre una parola.
– Analizzare la regolarità con la quale una parola è accompagnata ad altre (prima o dopo).

Durante l'analisi di un testo è fondamentale studiare le collocazioni, o meglio espressioni composte da più di una parola grafica, che si comportano semanticamente e spesso morfo-sintatticamente come un solo lessema. Identificare sequenze di due o più parole fortemente associate è utile a moltissimi fini: estrazione di collocazioni, espressioni idiomatiche e nomi propri. La misura di associazione più semplice è la frequenza. Le collocazioni indicano dunque quanto frequentemente due o più parole occorrono insieme e sono presenti nella lingua parlata e scritta di tutti i giorni. Esempi di collocazioni sono: "Presidente della Repubblica", "forza pubblica","capo dello Stato".

Esistono diversi gradi di intensità nel legame tra due o più parole che co-occorrono e diversi modi per definire l'intensità di questi legami e anche quanto frequentemente due o più parole si trovano affiancate nei testi. La misura di associazione classica in linguistica computazionale è la Mutual Information (MI)[2]:

$$MI(w1,w2)= [\log2\ P(w1,w2)[/[P(w1)*P(w2)]$$

Questa formula in teorie dell'informazione, quantifica l'informazione supplementare sulla possibile presenza di w2 come seconda parola del bigramma una volta che sappiamo che la prima parola è w1. Esistono due maniere intuitive di pensare alla MI ignorando la trasformazione logaritmica:

– La MI è il rapporto tra la probabilità di occorrenza di un bigramma in un corpus stimata empiricamente (contando il numero di volte in cui il bigrammacapita nel corpus) e la probabilità teorica che ci si aspetteremmo se le due parole che lo compongono fossero indipendenti (ottenuta dal prodotto delle probabilità empiriche delle due parole).
– La MI può essere vista come il rapporto tra la probabilità della seconda parola nel bigramma dato che a è nota la prima parola e la pro-

babilità della seconda parola indipendente dal contesto.

Dalle due interpretazioni si evince che più alto è il valore della MI più è plausibile che le due parole siano associate.

3. Linguistica computazionale e analisi del discorso

Considerate la maggiori classi di strumenti software della linguistica computazionale, si analizza ora come si possa definire un processo automatico per l'analisi del discorso.

Gli strumenti identificati nel lavoro consentono di definire una tool-chain in grado di:

- estrapolare da una raccolta di commenti, un insieme di termini più probabili identificati da un peso.
- Individuare categorie morfosintattiche quali pronomi, soggetti, predicati verbali, complemento oggetto, complemento d'agente, verbi modali.
- Determinare le parole che co-occorrono con maggiore frequenti all'interno di un corpus.

Vediamo come il processo può essere strutturato.

Indicizzazione del testo: fornisce la possibilità di indicizzare un testo, in linea generale consente di individuare topic all'interno di forme grafiche (termini-parole). Questo strumento è utile per classificare il contenuto di un testo o, se trasformato in forme geometriche (spazi e vettori) può consentire raffronti rapidi tra i contenuti di testi, anche di grandi dimensioni.

Analisi di testo. Tramite la scelta di opportune euristiche. Esso consente di:

- individuare, attraverso l'uso dei pronomi, il grado di coinvolgimento emotivo del soggetto che prende parte ad un processo comunicativo;
- conoscere il soggetto al quale un utente si riferisce ed individuarne il sentimento positivo, negativo, neutro attraverso le categorie "soggetto" e "complemento oggetto";
- individuare la forma attiva all'interno di una frase di senso compiuto mettendo in evidenza la persona che compie l'azione;
- individuare l'uso dei verbi (modo, tempo e forma), come ad esempio l'uso del condizionale può segnalare che l'enunciato in cui ricorre è un ipotesi anziché una vera e propria affermazione, infatti viene impiegato quanto il parlante non vuole assumere in pieno la responsabilità di affermare qualcosa, l'uso del congiuntivo può caratterizzare atti comparativi come quelli di augurarsi o auspicare qualcosa;
- distinguere, attraverso l'uso di un verbo modale, il significato e il significante di una frase. Il verbo "volere" può essere usato per presentare cortesemente come espressione di un desiderio quello che in realtà è una richiesta o addirittura un comando, l'uso del verbo dovere implica un obbligo e l'uso del verbo potere implica una possibilità.

Calcolo delle co-occorrenze. Consente di individuare i concetti principali contenuti nel corpus, attraverso l'individuazione delle principali associazioni esistenti fra la forma grafica che costituisce il corpus. Inoltre, mette a disposizione una funzione di calcolo delle co-occorrenze che permette di conteggiare quante volte una serie di forme, considerate a due a due, si presentano vicine tra loro nel corpus. Oltre a questi risultati consente di ottenere output più analitici, come il numero di word analizzate, numero di token distinti e numero di cooccorrenze distinte. Questa analisi può essere utile per focalizzare in modo più specifico il contenuto di un testo o per classificare i paradigmi espressivi.

4. Conclusioni

Con questo lavoro si è voluto delineare una tool chain che possa automatizzare e specializzare le attività di analisi del discorso, che sono ancora attività semi-automatiche.

La Linguistica Computazionale, infatti, sta sviluppando strumenti che, se integrati in modo efficace in un processo unitario, possono consentire allo studio del discorso una profondità attualmente raramente raggiunta.

Bibliografia

Blei D., Ng A.Y., Jordan M.I., *Latent dirichletallocation*, J. Mach. Learn. Res., 3:993-1022, 2003.

Church K., Hanks P., W*ord association norms, mutual information, and lexicography*, ACL 1989, pp. 76-83.

Dardano M., *Il linguaggio dei giornali italiani*, Laterza, Bari, 1973.

Dave K., Lawrence S., Pennock D.M., *Mining the peanut gallery: opinion extraction and semantic classification of product reviews*, in WWW '03 Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519-528.

Feinerer I., *Introduction to the tm Package Text Mining in R*, The Comprehensive R Archive Network, 2007.

Feldman R., Sanger J., *The Text Mining Handbook-Advances approaches in analyzing unstructured data*, Cambridge University Press, 2007.

Fowler R., Hodge B., Kress G., Trew T., *Language and Control*, Routledge and Kegan Paul, London, 1979.

Gerald S., *Automatic information Organization and Retrieval*, McGraw-Hill Inc, 1968.

Grisham R., *Computational Linguistics*. Cambridge University Press, 1986.

Kanayama H., Nasukawa T., *Fully automatic lexicon expansion for domain-oriented sentiment analysis*, in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pp. 355-363, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Manning C., Ragnavan P., Schutze H., *Introduction of Information Retrieval* Cambridge University Press, Cambridge, 2008.

Meyer C.F., *English Corpus Linguistics*, Cambridge University Press, 1992.

Nasukawa T., Yi J., *Sentiment analysis: capturing favorability using natural language processing*. In Proceedings of the 2nd international conference on Knowledge capture, K-CAP '03, pp. 70-77, New York, NY, USA, 2003. ACM.

Pang B., Lee L., Vaithyanathan S., *Thumbs up? sentiment classification using machine learning techniques*, in Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing. pp. 79-86. Philadelphia, 2002. Association for Computational Linguistics.

Pang B., Lee L., *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*, in Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics, pp. 271-278. Barcelona, ES, 2004. Association for Computational Linguistics.

Ratnaparkhi A., *A simple introduction to maximum entropy models for natural language processing*. Technical Report 97-08. Institute for Research in Cognitive Science, University of Pennsylvania, 1997.

Sinclair J., *Corpus, Concordance, Collocation*, 1991. [Oxford University Press, 1991].

Note

[1] Gerald Salton - Automatic information Organization and Retrieval - McGraw-Hill Inc. 1968.

[2] K. Church & P. Hanks Word association norms, mutual information, and lexicography. ACL 1989.