ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

CAMPUS DI CESENA
SCUOLA DI INGEGNERIA E ARCHITETTURA

Corso di Laurea in Ingegneria Elettronica, Informatica e Telecomunicazioni- ambito Informatica

ELABORAZIONE DEL LINGUAGGIO NATURALE NELL' IA E TECNOLOGIE MODERNE: SENTIMENT ANALYSIS COME CASO DI STUDIO

Elaborata nel corso di: Fondamenti di Informatica B

Tesi di Laurea di:
GIOVANNI CIANDRINI

Relatore:
Prof. ANDREA ROLI

Co-relatori:
Prof. FEDERICO CHESANI

ANNO ACCADEMICO 2014-2015
SESSIONE I

PAROLE CHIAVE

Intelligenza Artificiale

NLP

Sentiment Analysis

Approccio semantico

Approccio statistico

A tutte le persone più importanti della mia vita che mi hanno sostenuto sempre

Indice

Introduzione

L'informatica e le sue tecnologie nella società moderna si riassumono spesso in un assioma fuorviante: essa, infatti, è comunemente legata al concetto che ciò che le tecnologie ci offrono può essere accessibile da tutti e sfruttato, all'interno della propria quotidianità, in modi più o meno semplici.
Anche se quello appena descritto è un obiettivo fondamentale del mondo high-tech, occorre chiarire subito una questione: l' informatica non è semplicemente tutto ciò che le tecnologie ci offrono, perchè questo pensiero sommario fa presagire ad un'informatica "generalizzante"; l' informatica invece si divide tra molteplici ambiti, toccando diversi mondi inter-disciplinari, e non possiamo affrontare un percorso che riguarda una sua tecnologia se non teniamo conto di questa ipotesi fondante. L'importanza di queste tecnologie nella società moderna deve spingerci a porre domande, riflessioni sul perchè l'informatica, in tutte le sue sfaccettature, negli ultimi decenni, ha portato una vera e propria rivoluzione nelle nostre vite, nelle nostre abitudini, e non di meno importanza, nel nostro contesto lavorativo e aziendale, e non ha alcuna intenzione (per fortuna) di fermare le proprie possibilità di sviluppo. Occorre essere sempre stimolati all'idea che capire questi meccanismi è fondamentale per essere protagonisti nella nostra società , e che anche se non si è esperti o amanti di certe tematiche, non è un impresa impossibile, se vengono seguiti dei modelli precisi e una metodologia ordinata di analisi.
Per introdurre a un trattato scientifico che riguarda un ambito specifico inerente al mondo informatico occorre senz'altro capire come nel tempo si è sviluppato questo ambito all'interno del proprio contesto, per poter comprendere le basi che lo caratterizzano, ma è di vitale importanza a un certo punto distaccarsi dal passato e guardare al presente e al futuro, approcciandoci con occhio critico e con un buon bagaglio di conoscenze preliminari alle tecnologie specifiche che ci offre la modernità inerenti a quell'ambito.
In questo trattato ci occuperemo di definire una particolare tecnica moderna relativa a una parte di quel mondo complesso che viene definito come Intelligenza Artificiale. L'intelligenza Artificiale(IA) è una scienza che si è sviluppata proprio con il progresso tecnologico e dei suoi potenti strumenti, che non sono solo informatici, ma soprattutto teorico- matematici (probabilistici) e anche inerenti l'ambito Elettronico-TLC (basti pensare alla Robotica): ecco l'interdisciplinarità. Il lettore che si approccia per la prima volta a livello tecnico al concetto di Intelligenza Artificiale e che magari non ha una definizione chiara del concetto, deve affrontare questo percorso liberandosi

da tutti i pregiudizi, con il solo pensiero che l'IA rappresenta a tutti gli effetti una vera Scienza, cercando di cogliere nel primo capitolo del documento i concetti fondamentali per crearsi un idea organica in testa. Concetto che è fondamentale per poi affrontare il nocciolo del percorso presentato nel secondo capitolo del documento proposto: i due approcci possibili, semantico e probabilistico, verso l'elaborazione del linguaggio naturale(NLP), branca fondamentale di IA. Per quanto darò un buono spazio nella tesi a come le tecniche di NLP semantiche e statistiche si siano sviluppate nel tempo, verrà prestata attenzione soprattutto ai concetti fondamentali di questi ambiti , perché, come già detto sopra, anche se è fondamentale farsi delle basi e conoscere l'evoluzione di queste tecnologie nel tempo, l'obiettivo è quello a un certo punto di staccarsi e studiare il livello tecnologico moderno inerenti a questo mondo, con uno sguardo anche al domani: in questo caso, la Sentiment Analysis(capitolo 3). Sentiment Analysis (SA) è una tecnica di NLP che si sta definendo proprio ai giorni nostri, tecnica che si è sviluppata soprattutto in relazione all'esplosione del fenomeno Social Network, che viviamo e tocchiamo costantemente. L'approfondimento centrale della tesi verterà sulla presentazione di alcuni esempi moderni e modelli di SA che riguardano entrambi gli approcci(statistico e semantico), con particolare attenzione (nel quarto capitolo) a modelli di SA che sono stati proposti per Twitter in questi ultimi anni, valutando quali sono gli scenari che propone questa tecnica moderna, e a quali conseguenze contestuali (e non) potrebbe portare questa particolare tecnica. L'obiettivo principale, dunque, è cercare di accompagnare il lettore attraverso argomenti che tecnicamente possono sembrare abbastanza complicati, e che a volte prevedono l'uso di termini molto sofisticati, ma che sono essenziali per poter comprendere a pieno l'importanza di questa tecnica e delle sue ripercussioni su diversi ambiti, mettendo in evidenza sempre i modelli e i concetti fondamentali con una metodologia ordinata di analisi, cercando infine di invogliarlo ad applicare questa metodologia di ragionamento e questo tipo di approccio anche ad ambiti scientifici diversi da quello preso in esame in questo elaborato.

Capitolo 1

Natural Language Processing in IA

Illustriamo in questo capitolo un'idea organica di NLP all'interno del mondo "Intelligenza Artificiale", tenendo presente del contesto in cui questa tecnica si è sviluppata.

1.1 Il mondo IA

In questo capitolo partiremo definendo in maniera organica l'Intelligenza Artificiale, di cui NLP ne rappresenta una particolare branca, definizione che non può non tenere conto del background culturale e del contesto in cui essa si è sviluppata (pensiamo allo sviluppo e alla crescita delle conoscenze tecnologiche negli anni '70,'80); solo dopo aver definito e contestualizzato il mondo IA potremo avvicinarci a NLP, riassumendo in breve la sua storia, e presentando i concetti base e le tecnologie che riguardano l'elaborazione del linguaggio naturale, in particolare i diversi approcci che vengono utilizzati per definirle. In questo modo, in accordo con il nostro obiettivo, potremo avvicinarci con questo bagaglio di conoscenze ad affrontare le tecnologie moderne che riguardano questo ambito, dando una valutazione critica e organica del loro impatto nella società odierna.

1.1.1 Intelligenza Artificiale e test di Turing

Occorre subito specificare che il concetto IA non ha una definizione precisa, bensì molteplici definizioni che dipendono da quale approccio viene utilizzato per andare a descrivere questo concetto, approcci che erano senza distinzioni nei primi anni '70,'80, quando IA nasceva effettivamente : l'approccio umano, incentrato ai processi di pensiero e ragionamento, testando la somiglianza dell'elaboratore a un essere umano, e l'approccio razionale, incentrato sul comportamento e sulla razionalità, testando invece la razionalità dell'elaboratore. E' proprio negli anni '80-'90, con lo sviluppo degli strumenti a nostra disposizione, che arriviamo da Ingegneri a intraprendere

la strada verso un approccio puramente razionale nel definire l'IA, attraverso l'introduzione degli agenti razionali [11], unità elementari alla base di IA, allontanandosi da quell'approccio umano basato soprattutto sui comportamenti degli esseri umani, che porterà a ciò che verrà poi etichettato come scienza cognitiva. In parole semplici, la differenza tra i due approcci verte nel cambiamento del requisito di partenza : se nei primi tempi veniva approfondito il concetto di IA partendo dall'ipotesi che ci si trovava ad analizzare un essere con dei comportamenti simili e paragonabili a quelli di un essere umano, concentrandoci solamente sul corretto uso dell'inferenza, con la strada dell'approccio razionale invece siamo consapevoli di partire dall'ipotesi che abbiamo davanti a noi un elaboratore; non ci poniamo più dunque il problema di capire se ha comportamenti simili a un essere umano, ma generalizziamo l'approccio verso di esso , concentrandoci non solo sull'inferenza, poiché essa è solo uno dei molteplici meccanismi utilizzabili per arrivare alla razionalità. Gli elaboratori che abbiamo davanti prendono il nome di Agenti razionali , definiti semplicemente come un qualcosa che agisce, che fa qualcosa, che riesce a rappresentare la conoscenza e applicarvi un ragionamento, perché avere una buona idea del funzionamento del mondo non solo consente di apprendere passivamente ciò che caratterizza l'ambiente su cui si affacciano, ma permette loro di generare strategie più efficaci per interagire con esso: ecco la svolta fondamentale per IA.
Il concetto di razionalità in poche parole, intesa da Ingegneri , passa da fare la cosa giusta a fare qualcosa.

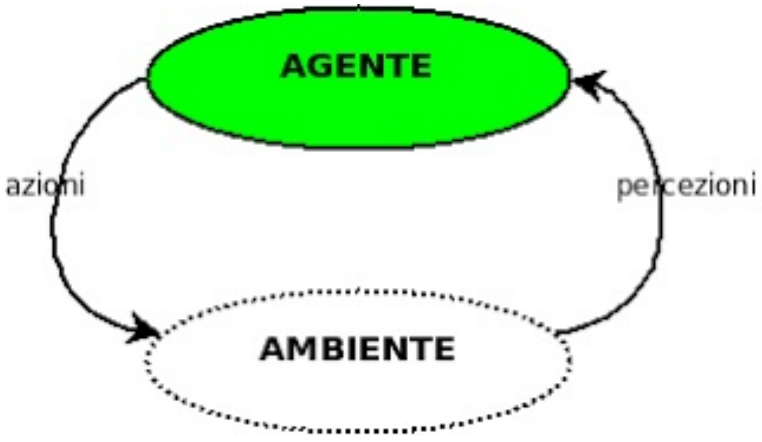

Figura 1.1: *Schema elementare dell'agente razionale, unità del mondo IA: dalle percezioni ricevute dall'ambiente, l'agente fa qualcosa su di esso attraverso delle azioni, che produrranno delle conseguenze contestuali sull'ambiente tali a volte da alterare e modificare le percezioni iniziali dell'agente stesso.*

Anche se come è stato detto il concetto di IA non è definibile univocamente perché dipende dal tipo di approccio che utilizziamo verso questo mondo (umano o razionale), possiamo individuare due fattori comuni che

ne hanno determinato senza meno lo sviluppo e lo stato dell'arte moderno: lo sviluppo delle tecnologie e delle conoscenze informatiche, elettroniche e non solo, ma soprattutto il punto di partenza del mondo IA, che può essere intravisto nel famoso test di Turing (1950), considerato il crocevia concettuale per fornire una soddisfacente definizione operativa dell'intelligenza.
Quando un calcolatore è intelligente? Turing non suggerisce una lista di caratteristiche sine qua non un elaboratore di informazioni può essere considerato razionale o meno, bensì egli fornisce un test basato sull'impossibilità di distinguerlo da entità che lo sono senza dubbio: gli esseri umani.
In sostanza, il test di Turing prevede tre mondi (A,B,C) divisi tra di loro e senza alcun tipo di conoscenza reciproca, dove nel mondo A abbiamo sicuramente un esaminatore umano, e nei mondi B e C sappiamo esserci un elaboratore e un essere umano, senza che l'esaminatore umano conosca a priori quale mondo rispettivo occupano questi due elementi. L'esaminatore umano fa una sola azione: può fare delle domande in forma scritta e mandarle ad entrambe i mondi, ricevendo una risposta; nel momento in cui l'esaminatore non riuscirà a capire se le risposte provengono da una persona oppure da un elaboratore, allora l'elaboratore in questione avrà passato il test. Come prima cosa, è importante constatare che l'ambiguità tra gli approcci di IA (umano e razionale) dei primi tempi discende proprio dall'impostazione di questo test, che invita a nascondere l'ipotesi di partenza nella definizione di razionalità (sto parlando con un calcolatore, o con un essere umano?). Questo tipo di test invita successivamente a due questioni importanti: è assolutamente matematico l'assioma test passato, computer intelligente?. Ma soprattutto: quanto può essere grosso il lavoro che sta dietro alla programmazione di un elaboratore in grado di passare questo test? Mentre per la prima domanda rimandiamo ad altri testi per un'analisi più approfondita riguardo le tematiche inerenti il mondo IA e il relativo stato dell'arte, affrontiamo la seconda domanda riformulandola da Ingegneri, progettisti : quali devono essere le capacità che l'elaboratore deve possedere per passare questo test? Gli esperti di IA hanno elencato sei diverse capacità da tenere in considerazione per poter progettare e programmare un calcolatore(agente razionale) in grado di poter passare questo test:

- Interpretazione del linguaggio naturale (l'agente deve comunicare con l'esaminatore umano)
- Rappresentazione della conoscenza (l'agente deve memorizzare quello che sa)
- Ragionamento automatico (l'agente deve utilizzare la sua conoscenza per rispondere e trarre conclusioni)
- Apprendimento (l'agente deve adattarsi a nuove circostanze)
- Visione artificiale (l'agente deve percepire gli oggetti)

- Robotica (l'agente deve poter manipolare gli oggetti e spostarsi fisicamente)

Nonostante le ultime due capacità non riguardano esattamente il test di Turing in questione ma più precisamente un test di Turing totale (dove l'esaminatore testa anche la capacità percettive e di movimento del soggetto), notiamo comunque che progettare un sistema in grado di riuscire a passare questo test richiede la progettazione di diverse capacità, ognuna delle quali diventerà nel tempo una particolare branca del variegato mondo IA. La potenza di questo test la vediamo proprio durante tutti questi sessant' anni in cui esso è rimasto comunque significativo: in tutti questi anni i ricercatori non hanno mai fatto tanti sforzi per progettare un sistema capace di passare il test in questione, ma hanno invece speso tutte le loro energie (coadiuvati da un progresso tecnologico sempre più incessante) allo studio dei principi alla base del concetto intelligenza fornito da questo test, definendo e studiando le diverse componenti IA che derivano dalle capacità elencate proprio qui sopra. Il test non fu importante quindi semplicemente per il suo significato, ma per quello che ha portato nello studio di queste tematiche: è grazie al test di Turing che il mondo IA oggi è quello che conosciamo, ma allo stesso momento paradossalmente nessun calcolatore è riuscito mai a passare questo test.
Ed è proprio dalla prima delle capacità elencate, l'interpretazione (ed elaborazione) del linguaggio naturale (Natural Language Processing: NLP), che noi partiremo per affrontare il percorso che svolgeremo nei prossimi capitoli.

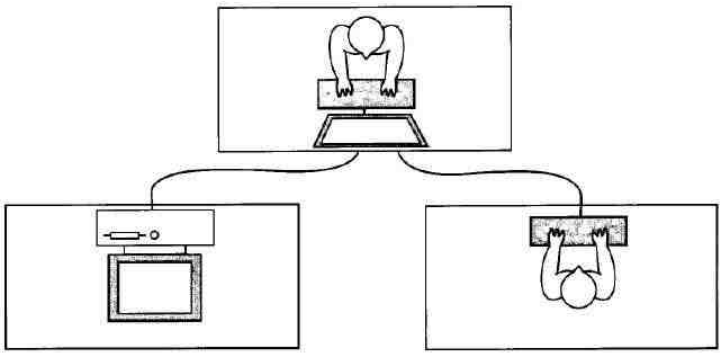

Figura 1.2: *Rappresentazione esemplificativa del test di Turing.*

1.2 Primi approcci a NLP

L'obiettivo finale di questa sezione sarà andare a definire, in maniera rigorosa e organica, il concetto di NLP all'interno dell'Intelligenza Artificiale, relativamente alla quale abbiamo discusso e esaminato le caratteristiche in maniera concisa ma completa nella precedente sezione.

Ora, per poter arrivare a dare un senso all'elaborazione del linguaggio naturale, teniamo solamente presente quello che abbiamo dedotto sinora dal test di Turing, associando NLP alla capacità di un elaboratore di comunicare con un essere umano, e cerchiamo, attraverso dei brevi cenni storici, di poter capire effettivamente nel tempo come si sono sviluppate queste tecniche legate al concetto NLP, in relazione allo sviluppo delle tecnologie e delle conoscenze relative a IA: cerchermo nella sottosezione seguente di ripercorrere molto brevemente la sua storia e le relative tecniche, per poter poi nella seconda sottosezione analizzare il concetto e, tenendo in considerazione il suo sviluppo nel tempo, dare una definizione rigorosa, andando poi alla fine del capitolo a elencare qualche esempio e tecnica concreta, che riprenderemo poi più avanti nel nostro percorso.

1.2.1 Brevi cenni storici

Ripercorrendo la storia di NLP fino ai giorni nostri, possiamo dividere il suo sviluppo in quattro diverse fasi, ognuna delle quali ha prodotto lavori e ricerche fondamentali per lo sviluppo dello stato dell'arte moderno, caratterizzate rispettivamente dalla traduzione automatica, dall' influenza di IA, da un'adozione di uno stile logico-grammaticale e dall'uso massivo di dati linguistici e l'introduzione dei Big Data (vedremo più avanti l'importanza di questo concetto)[8].

- **Prima fase: Fine anni '40 - Fine anni '60: Machine Translation (MT)**
  La prima fase fu caratterizzata soprattutto dalla traduzione automatica, branca della Linguistica Computazionale che studia la traduzione dei testi da una lingua naturale a un'altra attraverso programmi informatici (Google Translate, ndr). Le ricerche inerenti a NLP cominciarono ufficialmente nei primi anni '50, attraverso dei primi rudimentali esperimenti di traduzione automatica dall'Inglese al Russo, finanziati in dimostrazioni di IBM nel 1954, ma è alla fine degli anni '50 che NLP cominciò a vedere gli albori, quando fu collegata da Minsky all'information retrieval(IR), insieme di tecniche che si occupano di gestire la rappresentazione, la memorizzazione, l'organizzazione e l'accesso ad oggetti contenenti informazioni. Questa fase fu dominata da un crescente ottimismo e entusiasmo, proprio perché in un epoca dove ancora le risorse e le tecniche computazionali erano molto povere (non esistevano linguaggi ad alto livello, e le macchine erano caratterizzate da un accesso veramente minimo alle risorse e una quantità ridicola di storage), si riuscirono comunque ad affrontare i primi problemi relativi alla semantica, sintattica e all'ambiguità del linguaggio, fornendo e completando dizionari e regole di traduzione.
  Le tecniche di NLP iniziarono a collegarsi fortemente a queste regole di traduzione, ma ancora risultavano distanti dal mondo IA.

- **Seconda Fase: Fine anni '60 - Fine anni '70: Intelligenza Artificiale (IA)**
  La seconda fase di NLP fu caratterizzata dal sempre più forte accostamento al mondo IA, e quindi vediamo più enfasi verso la conoscenza del mondo e dell'ambiente circostante (importanza degli agenti razionali) e verso il ruolo della costruzione e della manipolazione del significato della rappresentazione. I dizionari, le regole di traduzione sviluppate nella prima fase, furono ridimensionate dallo stesso Minsky (1968) per diventare input linguistici in dei primi sistemi semplici capaci di interpretare questi input e in questo modo avvicinarsi a un corretto uso dell'inferenza. SHRDLU(Winograd, 1973) e LUNAR(Woods,1978) furono i discendenti di questi primi sistemi, che però mostravano capacità ancora migliori di processare task NLP con uno stile procedurale. In questa fase è da tener soprattuto conto del cambio di direzione inerente al significato di semantica dei task NLP: nel 1980, R. Schank, in accordo con le teorie che stavano alla base dei primi sistemi capaci di interpretare e elaborare quelle regole che erano state trovate nella prima fase di NLP, lavorò in maniera tale da creare modelli in cui la semantica non fosse più legata a una concezione (tipica delle prime analisi linguiste) in cui si andava a valutare singolarmente le proposizioni logiche all'interno di frasi e periodi, ma arrivò a definire una semantica general purpose: questo concetto di organizzazione a larga scala della semantica denota le interazioni tra tutti gli elementi che fanno parte dell'universo del discorso, e può dare così un grande contributo al supporto dell'inferenza, soprattutto per quanto riguarda dialoghi e discorsi prolungati, e non semplici proposizioni logiche. NLP diventa semantic-driven , con la concezione di una semantica "general purpose".

- **Terza fase: Fine anni '70 - Fine anni '80: Analisi Logico-Grammaticale**
  Se la seconda fase di NLP fu caratterizzata dall'influenza dell' IA e dal concetto di semantica in un significato più allargato, la terza fase di NLP vide un rientro in campo da parte della figura dei linguisti relativamente a queste tecniche, con l'introduzione di una nuova analisi logico-grammaticale nella rappresentazione della conoscenza e nella costruzione di sistemi capaci di gestire task NLP: infatti, il limite della fase semantica general-purpose veniva intravisto nel fatto che vennero dati nuovi modelli generali del concetto di semantica, ma le regole linguistiche-grammaticali erano ancora regole particolari e legate al singolo periodo, frase, contesto. Quindi vennero proposti nuovi modelli generali di analisi logico-grammaticale, raggruppando queste regole in un numero ben definito di tipi di grammatica, per esempio funzionale, categorica, e in delle strutture generali: soltanto in questo modo potevano essere orientate verso la computabilità come dei

principi astratti, generali, supportando anche algoritmi importanti di parsing, riferito al processo che analizza un flusso continuo di dati in ingresso (input) in modo da determinare la sua struttura grazie ad una data grammatica formale; in questa fase vennero creati i primi parser , programmi che eseguivano questo compito.

- **Quarta fase: Anni '90: Statistic Natural Language Processing (SNLP)**
  L'approccio linguista che caratterizzò la terza fase fu di grande influenza nella quarta e ultima fase di NLP, che vede una grande svolta proprio durante gli anni '90. I modelli logico-grammaticali e il concetto di semantica general-purpose, la costruzione di parser in grado di tradurre e comprendere dizionari anche molto consistenti, alberi lessico-grammaticali in grado di gestire contesti molto ampi di traduzione, diedero adito a un nuovo approccio per poter gestire, manipolare una grande quantità di dati e di informazioni, che con gli albori di Internet videro un primo ingresso nel mondo della tecnologia: i cosiddetti Big Data. Sebbene i modelli per gestire task NLP attraverso un approccio semantico general-purpose erano molto astratti, e quindi potevano comunque affrontare una gran quantità di flusso di informazioni, si vide proprio in questi anni il limite di questo approccio, che era comunque legato (seppur attraverso dei modelli generali) alla conoscenza del singolo dato, task, information: per questo motivo, nell'ultima fase di NLP nacque un nuovo approccio capace di affrontare, elaborare e interpretare attraverso un calcolatore questa grande quantità di informazioni: arriviamo a un approccio statistico/probabilistico per NLP (Manning e Schuetze, 1999), che caratterizzò i nuovi sistemi di NLP e i nuovi parser dell'ultima decade, applicando i principi base della teoria della probabilità a questi sistemi NLP. Arriviamo proprio in questa decade e con quest'approccio a definire nuove fondamentali tecniche attraverso un approccio probabilistico di NLP in grado di sviluppare capacità molto importanti: estrazione dell'informazione, information retrieval(motori di ricerca), sono solo due esempi della potenza di quest'approccio, che vedremo caratterizzare tutt'oggi il contesto moderno di queste tecniche.

Vedremo nell'ultima sezione del primo capitolo qual è lo stato moderno di NLP ai giorni nostri, ma prima intendiamo dedicare una sottosezione alla definizione formale e organica di NLP, avendo nel nostro bagaglio tutti i passaggi più importanti dell'evoluzione di questo concetto negli ultimi 50 anni, in relazione allo sviluppo tecnologico contestuale.

1.2.2 Linguistica Computazionale e IA: definizione di NLP

Nella conclusione della sottosezione inerente a IA, sono state messe in risalto le caratteristiche e le proprietà fondamentali che denotano, delineano il concetto di intelligenza associato a un calcolatore: in particolare, prendiamo come punto focale la prima di queste caratteristiche, l'interpretazione del linguaggio naturale, collegata quindi alla capacità che deve possedere l'agente razionale per comunicare con l'essere umano(che è l'esaminatore nel test di Turing).
Cosa intendiamo per interpretare, elaborare, capire il linguaggio naturale? Occorre definire un punto di partenza per la nostra analisi, e occorre inoltre cercare un collegamento saldo tra NLP e il concetto di linguaggio. In prima analisi, riflettiamo per esempio sulle forze fondamentali della natura: la forza gravitazionale, elettromagnetica, nucleare debole e forte, che hanno una natura molto diversa tra loro, sono accomunate da una caratteristica fondante: sono tutte forze d'interazione, e quindi intendiamo da questo seppur banale collegamento alla realtà, che il problema fondamentale, anche in natura, è quello di definire un concetto di interazione, delineando i soggetti di quest'interazione e il significato vero e proprio del concetto a differenza del contesto. Avvicinandoci ora all'ambito informatico, vediamo per esempio che l'interazione acquisisce un ruolo significativo nel mondo dei Sistemi Distribuiti, dove un sistema Software funziona lavorando su macchine fisiche diverse, e il problema dell'interazione tra i diversi ambienti diventa una delle principali questioni da gestire, non solo a livello di progetto del sistema, ma soprattutto a livello di modellazione e di analisi, nella produzione del suddetto software. Tutto ciò per mettere in evidenza il punto di partenza: in ogni problematica tecnica, partiamo dall'affrontare la questione dell'interazione.
Nell' IA i soggetti dell'interazione sono facili da individuare: abbiamo a che fare sicuramente con un' interazione uomo-macchina, che porterà dietro tutte le problematiche del caso (elencate anche dal test di Turing), e che sicuramente dovrà essere affrontata esponendo modelli e tecniche che riguardano l'interazione tra essere umani e elaboratori. Trovati i soggetti dell'interazione, è più complicato definire il vero e proprio modello di interazione tra esseri umani e calcolatori: dobbiamo cercare delle tecniche che riescano a far comunicare due mondi in apparenza completamente diversi, nonostante l'elaboratore abbia con sè le caratteristiche di un agente razionale; come fare a trasformare il linguaggio umano, con tutte le sue problematiche, le sue ambiguità, in una serie di dati comprensibili da un elaboratore tramite regole formali (linguaggio macchina), considerando anche che questa trasformazione dev'essere reversibile (andata e ritorno, l'essere umano deve poter capire dati elaborati da un calcolatore). Inseriamo ora, in prima battuta, l'elaborazione del linguaggio naturale: essa è definita proprio come tecnica capace di produrre modelli che hanno il ruolo di mediare

tra il mondo umano e il mondo degli elaboratori a livello di comprensione e generazione del linguaggio naturale. Ecco il collegamento tra NLP e il concetto vero e proprio di linguaggio: i programmi NLP devono poter riuscire a elaborare espressioni proprie del linguaggio umano, che possono essere scritte in diverse espressioni della lingua (dialetti, lingue, registro scritto, parlato); la comprensione di NLP è vincolata quindi alla conoscenza di diversi campi dello studio del linguaggio. Il problema di definire NLP quindi passa attraverso la conoscenza e le problematiche che riguardano lo studio del linguaggio vero e proprio: ad esempio fonetica, semantica, morfologia, sintassi... Per questo motivo, quando si parla di NLP, si tende ad associare molto spesso questo concetto ad una branca che si è sviluppata a partire dagli anni '50, non molto distante ma diversa dal concetto IA, che tocca in maniera molto forte queste tematiche: la Linguistica Computazionale, che si concentra sullo sviluppo di formalismi descrittivi del funzionamento di una lingua naturale, tali che si possano trasformare in programmi eseguibili dai computer. I problemi che affronta la linguistica computazionale consistono nel trovare una mediazione fra un oggetto di studio in costante evoluzione (il linguaggio umano) e le capacità di comprensione della macchina, limitate a quanto può essere descritto tramite regole formali; con NLP si tenta di dare quindi proprio questo modello di interazione, cercando di formulare approcci (che vedremo nel proseguo del percorso) in grado di determinare tecniche capaci di mediare il problema di comprensione e di gestione del linguaggio naturale.
Considerato il collegamento molto forte tra NLP e la Linguistica Computazionale, il ruolo di NLP come mediazione a livello di linguaggio tra elaboratori e esseri umani, e l'importanza del suo collegamento con le tecniche vere e proprie dello studio del linguaggio (vedremo più avanti come sarà molto importante nelle tecniche moderne di NLP il ruolo e l'importanza della figura dei linguisti), proviamo a dare una definizione organica del concetto:

L'elaborazione del linguaggio naturale *è una branca, un campo di studi e di ricerca che si divide tra IA e Linguistica Computazionale, e che fa riferimento al processo di trattamento automatico mediante un calcolatore elettronico delle informazioni scritte o parlate in lingua naturale, ponendosi come vero e proprio modello di mediazione nell'interazione uomo-macchina: la complessità che sta alla base di questo processo, dovuta alle caratteristiche intrinseche di ambiguità del linguaggio umano, è affrontata attraverso un duplice meccanismo che da una parte suddivide questo processo in fasi diverse, con analisi lessicale, grammaticale, sintattica, semantica(influenza della Linguistica Computazionale), e da un'altra parte associa a queste analisi tecniche in grado di implementare al meglio questo modello di mediazione(influenza di IA).*

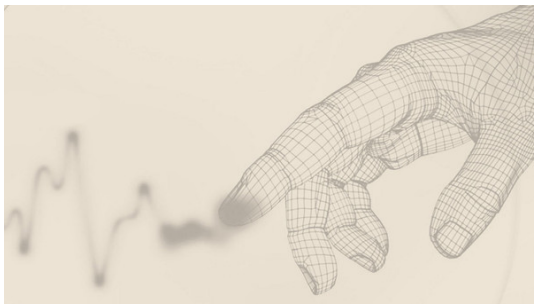

Figura 1.3: *Il ruolo fondamentale di NLP è quello di mediazione tra linguaggio umano(indefinito e ambiguo) e linguaggio macchina(definito e con regole formali*

1.3 Stato dell'arte moderno di NLP

Dopo aver discusso ampiamente sulla definizione e sul significato contestuale dell'elaborazione del linguaggio naturale, e averne ripercorso brevemente le sue evoluzioni storiche nei campi IA e Linguistica Computazionale, interiorizzando la sua importanza in relazione a tecniche che possiamo già intuire avere influenze molto significative in un contesto sociale dove l'impatto tecnologico è rilevante (come ad esempio il contesto storico attuale della società, sempre più tecnologica e informatizzata a tutti i livelli), terminiamo questo capitolo mostrando lo stato dell'arte di NLP ai giorni nostri, riassumendo anche attraverso qualche esempio e riferimento il suo impatto nella società moderna. Ritengo che questa sezione, seppur breve e comunque di carattere ancora introduttivo, sia molto importante per capire e per affrontare il proseguimento del trattato, dato che poi ci avvicineremo nel prossimo capitolo a un'analisi molto più tecnica di NLP e dei suoi approcci, valutando attraverso esempi anche tecniche moderne, per poi andare a sfociare nei capitoli successivi a una tecnica moderna particolare, nucleo della tesi: Sentiment Analysis.
Ai giorni nostri NLP è in continua espansione e grazie alle nuove tecnologie informatiche ricopre tutt'ora un ruolo fondamentale ad esempio per quanto riguarda le sue espressioni in campo di Information Retrieval(Google, Bing, Yahoo..) e nel campo della traduzione automatica (Google Translate, dizionari online..); molti progressi sono stati fatti nel campo della sintassi, migliorando i programmi parser attraverso strutture logiche sempre più ben definite, e inoltre la potenza sempre più elevata degli elaboratori ha facilitato molto l'esecuzione e la computazione di algoritmi anche pesanti, non consumando le risorse. Ovviamente siamo ancora lontani da sistemi perfetti in grado di lavorare al 100% delle possibilità e garantire una conoscenza globale dell'ambiente sul quale si affacciano e dei dati che hanno a disposizione, basti pensare ai cosiddetti "problemi intrattabili", ma è facile pensare che NLP occupi nella società moderna una posizione rilevante,

cosa che non poteva essere se la storia di NLP fosse stata diversa o fosse stata scollegata dal contesto informatico in espansione. Ripercorrendo la storia di NLP attraverso le sue diverse fasi, vediamo che da una primissima fase di semplice MT e di semplici ricerche di dati siamo passati attraverso uno sviluppo tecnologico e ricerche in questo settore a modelli sempre più general-purpose, che affrontassero e che dessero regole il più generali e riusabili possibili, come nella produzione di un qualsiasi sistema software, perché è di questo che stiamo parlando: parlare di tecniche di costruzione di un sistema NLP si avvicina molto di più alla produzione di un sistema software di quanto crediamo, ma è essenziale definire dei modelli e degli approcci , dato che questi sistemi fanno riferimento a una razionalità, a una conoscenza globale del mondo intorno a sè che non è prevista da un semplice sistema software: in questo modo, con l'introduzione di un approccio duale a quello semantico (probabilistico-statistico), ci rendiamo conto di riuscire a modellare e programmare algoritmi e interi sistemi riutilizzando modelli e teorie prettamente matematiche impiantate in un contesto e in un problema informatico di elaborazione di grandi quantità di informazioni. Vedremo che la Sentiment Analysis, tecnica di elaborazione del linguaggio naturale che si sta definendo proprio nel nostro contesto, e che dipende direttamente dalle nuove tecnologie informatiche di grande interesse (Social Network), riprenderà proprio questi concetti appena espressi: oggi le tecniche di NLP, realizzate attraverso un approccio definito e dei modelli general-purpose che si sono sviluppati in questi 50-60 anni della sua esistenza, cercano di essere impiantate in tecnologie informatiche moderne, cercando di sfruttare la potenza di queste tecniche.
Unire il vecchio al nuovo, unire il bagaglio di conoscenza legato a NLP e ai suoi modelli alle nuove infrastrutture e strumenti che ci offre la tecnologia, sapersi adattare al cambio repentino del contesto tecnologico e del progresso moderno, sarà proprio il tema che verrà ripreso nella conclusione, che in un certo senso ridefinisce il ruolo di un ingegnere informatico all'interno di un'azienda.

Capitolo 2

Approccio statistico e approccio semantico

Esaminiamo e apprendiamo in questo capitolo i concetti chiave di NLP, studiandone i possibili approcci tecnici, per poter essere in grado di ritrovare questi concetti su tecniche moderne di IA.

2.1 Concetti chiave e modelli generali per NLP

In questa sezione andremo a riprendere il concetto di elaborazione del linguaggio naturale come l'abbiamo definito nel primo capitolo del trattato, cercando di valutare prima le problematiche intrinseche di questa tecnica, e quindi l'esigenza di fornire un percorso capace di stabilire modelli generali in grado di affrontare queste difficoltà. Partiremo poi nel definire i concetti tecnici chiave nell'elaborazione del linguaggio naturale, valutando nelle successive sottosezioni gli elementi fondamentali dell'elaborazione del linguaggio naturale, fornendo un flow definito e modelli base che verranno poi affrontati in seguito da due filoni di ricerca basati su approcci duali, statistico e semantico, che vedremo nella prossima sezione.

2.1.1 Valutazione delle problematiche di NLP

Per poter gestire il ruolo di mediazione legato al concetto di NLP per come l'abbiamo definito nel precedente capitolo, e' naturale pensare che occorre gestire svariate problematiche legate al concetto che risolvere task NLP significa andare ad affrontare l'ambiguità del linguaggio parlato. Attraverso tecniche e modelli svariati, più o meno performanti, occorre definire prima modelli per gestire queste interazioni, e poi in ultima analisi implementare tecniche in grado di seguire questi modelli. Quello appena descritto è il problema iniziale relativo alla condizione di NLP, ovvero legato intrinsecamente alla sua definizione, e al suo ruolo nel mondo dell' IA: analizzando

però il flusso che attraversa NLP nel ruolo di mediatore, ci rendiamo conto che questa problematica è divisa a sua volta in due sfaccettature. Se da una parte il mediatore NLP dev'essere in grado di riuscire a tradurre il linguaggio naturale in dati formali, ci sarà anche un processo di ritorno, in cui il mediatore dev'essere in grado di generare da dei dati formali delle informazioni esprimibili in linguaggio naturale. Per questo motivo le problematiche dei task NLP si dividono in due grandi categorie: problemi di interpretazione e problemi di generazione. I problemi di interpretazione e di comprensione riguardano tutti quei problemi che richiedono processi in grado di partire dal linguaggio naturale, effettuare delle rimozioni di ambiguità, e riuscire a costruire dati formali in grado di essere interpretati con facilità da un elaboratore e nelle operazioni di calcolo; per quanto riguarda la seconda categoria, i problemi di generazione, abbiamo invece tutti quei problemi che richiedono processi che comprendano la generazione (a partire da dati formali) di dati comprensibili all'uomo, attraverso ad esempio capacità di scegliere una struttura del testo, di scegliere una costruzione sintattica adeguata, di generare un intonazione vocale adatta, etc..

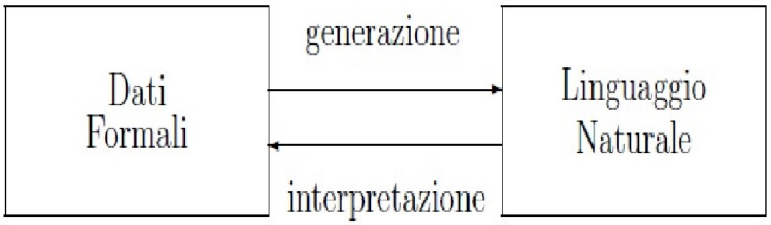

Figura 2.1: *Le due grandi categorie di problemi dei task NLP: problemi di interpretazione e problemi di generazione*

Linguistica computazionale, Intelligenza Artificiale, Computer Science: NLP è a tutti gli effetti un agente interdisciplinare, coinvolto in maniera più o meno profonda in tutti questi mondi, ma la vera difficoltà nel trattare NLP si può vedere nell'ambiguità presente a livelli differenti del linguaggio stesso (problemi di interpretazione). Analizzare il linguaggio naturale per poterne dare un'interpretazione significativa e non ambigua significa analizzare l'ambiguità di un concetto, un periodo, un discorso, in diversi livelli. Proviamo a valutare con un esempio questo concetto appena espresso, di importanza fondamentale: se vogliamo interpretare la frase Inglese " I made her duck " , per interpretare questo periodo ho bisogno di disambiguare almeno tre differenti livelli: duck viene inteso come il verbo "to duck" , oppure

si riferisce a “anatra“ ? Il verbo “made“ ha significato di “fare, creare“ , oppure di “cucinare“? E inoltre: her e duck, fan parte della stessa struttura sintattica? In questi tre livelli di analisi si annidano i primi esempi di soluzioni che proporranno modelli che poi osserveremo più a fondo: il Part of Speech Tagging risolverà la prima ambiguità, il Word Sense Disambiguation e il Probabilistic Parsing le ultime due. Come analizzeremo successivamente, questi problemi vengono oggi risolti con soluzioni ad alto livello relative a tecniche di Intelligenza Artificiale e di Machine Learning, ma occorre mettere in evidenza che tutte queste soluzioni partono comunque dal modellare prima a basso livello queste tipologie di problemi, che troviamo soprattutto a livello di linguaggio, per poi sfruttare questi modelli e proporre tecniche avanzate in grado di ottimizzare queste soluzioni. Questi concetti base che andremo a fornire nella prossima sottosezione saranno poi investiti da due approcci, statistico e semantico, di cui valuteremo i diversi impatti senza però andare a cercare una strada migliore tra i due (anche se nell'ultima decade i sistemi di apprendimento statistico hanno avuto i risultati migliori nella risoluzione di questi task) , ma cercando di evidenziare come entrambi gli approcci siano validi in termini di ricerca di soluzioni per i task NLP, facendo riferimento a titolo di esempio anche a qualche progetto. In particolare, definiremo dunque nelle sottosezioni seguenti elementi in grado di affrontare un percorso composto da: analisi morfologica, modellazione del linguaggio, parsing sintattico, part-of-speech tagging, traduzione statistica e semantica lessico-computazionale. Solo dopo questo processo “a basso livello“ saremo in grado di astrarre e valutare soluzioni “ad alto livello“ per affrontare il problema dell'interpretazione del linguaggio naturale, ed esaminare i due approcci nella risoluzione a livello di software di queste problematiche.

2.1.2 Analisi Morfologica e parsing sintattico

Il legame di NLP con il campo della Linguistica Computazionale è veramente forte, e infatti vediamo subito che la figura dei linguisti in questo ambito è di vitale importanza: partiamo infatti nell'affrontare il problema di modellare NLP dal problema di modellazione del linguaggio e di Analisi Morfologica del linguaggio.
L'elemento che sta alla base di questa prima fase sono le parole [10]: esse sono l'unità fondamentale nella nostra analisi, rappresentano il blocco base del linguaggio, compongono qualsiasi tipo di linguaggio umano (parlato, scritto..), e rappresentano soprattutto le più piccole forme del linguaggio che possono essere enunciate autonomamente e avere un contenuto pragmatico e semantico(dotate di un significato). In generale, esse a loro volta sono composte da morfemi, il cui concetto rappresenta il più piccolo elemento di una parola dotato di significato che non può essere suddiviso ulteriormente; dunque, studiare le unità fondamentali del linguaggio significa andare a studiare i morfemi che compongono le parole. Questi morfemi si dividono a loro

volta in due tipologie : morfemi radice (i morfemi principali delle parole), e i morfemi che rappresentano suffissi, prefissi, collegati con le parole stesse. Non perdiamo però di vista il contatto con la nostra analisi principale: è fondamentale avere chiaro questa seppur generica introduzione inerente all' analisi morfologica del linguaggio, per poter essere in grado di affrontare il prossimo step, ovvero una prima trasposizione di questo scenario verso il mondo informatico, cioè collegare il concetto di analisi morfologica a un qualche processo informatico: è qui che nasce il concetto di Parsing, inteso come processo che analizza un flusso di dati continuo in input, in modo da determinare una sua struttura definita data una grammatica formale. Il nostro problema, dunque, sarà quello di costruire un Parsing Sintattico, in grado di fare questa prima analisi morfologica delle parole, e fare detecting dei vari morfemi che eventualmente le compongono, dando una prima forma di struttura definita (formale); lo schema è quello classico di un sistema, dove inseriamo le parole come degli input veri e propri e il sistema di parsing rappresenta la scatola nera che è in grado di elaborare questi input e restituire in output dati formali che verranno dati in pasto a un elaboratore [5]. Definiamo ora tre passi fondamentali che compongono il percorso (bilaterale) per risolvere il problema di costruzione di un sistema del genere, legato ovviamente al problema principale di disambiguare le parole in relazione al contesto semantico. Nel processo di Parsing, in un primo livello (Livello superficiale), abbiamo bisogno di riuscire a riscontrare e formalizzare in maniera astratta le regole grammaticali con i quali i morfemi sono combinati per generare la parola in questione (per esempio, in inglese, la parola:
cities = city + s); parte qui il vero detecting dei morfemi all'interno delle parole. In un secondo livello (Livello intermedio), modelliamo questi morfemi trovati per cogliere il loro valore semantico all'interno della parola (ad esempio, nell'esempio precedente, la "s" trovata rappresenta un plurale:
city + s = city + PL); in questa fase, che rappresenta il cuore dell'attività del Parsing, avremo bisogno di una modellazione particolare attraverso macchine a stati finiti (FST), dipendenti da grammatiche formali specifiche. Infine arriviamo all'ultimo livello, il livello lessico, in cui abbiamo quindi fatto parsing sintattico della parola per valutarne la formalità, e abbiamo tutte le informazioni per mettere questa word in pasto a un elaboratore che lavora con dati formali, e sarà in grado di comprenderla. In maniera duale e simmetrica, andremo a comporre il processo di creazione della parola a partire da un dato formale.

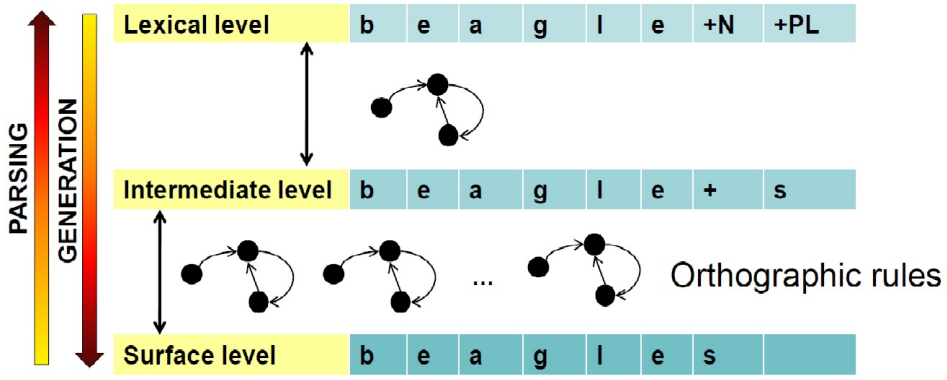

Figura 2.2: *Schema puntuale del processo di Parsing e del processo di Generazione: la parola "BEAGLES"*

Dopo aver valutato tecnicamente questi concetti inerenti all'interpretazione di una singola parola, e aver visualizzato bene il contesto in cui ci stiamo muovendo, finalmente abbiamo raggiunto l'obiettivo di aver trasposto in maniera esaustiva il problema linguistico dell'interpretazione sintattica verso un processo di parsing sintattico puramente informatico, in grado di interpretare semanticamente una singola word estrapolandola da un contesto più ampio; ora dovremo provare nella prossima sottosezione a capire in quale maniera riuscire a lavorare con più entità, con più parole, studiare le varie combinazioni tra esse, per avvicinarci sempre più a interpretare e analizzare un "periodo", e non una singola parola che lo compone.

2.1.3 Modelli di linguaggio, probabilità e CFG

Anche se dovremo a un certo punto (come prima) collegarci a un processo informatico, capiamo proprio nel momento in cui dobbiamo affrontare la combinazione dell'analisi di più parole l'importanza del concetto di probabilità collegato a queste tecniche. Valuteremo più avanti in concreto il vero impatto dell'approccio probabilistico e i suoi veri punti di forza, ma occorre iniziare a determinare i legami e le interazioni chiave che esistono a basso livello tra NLP e la probabilità; per questo motivo inseriamo qui, in un contesto ancora "comune" tra approccio semantico e approccio statistico, alcuni concetti che rappresentano le fondamenta di NLP per entrambi gli approcci. Siamo ora in grado di processare un testo a livello morfologico, e quindi cerchiamo di passare alla fase successiva, ovvero la gestione di un periodo, di una frase composta da più parole, partendo dal definire dei veri e propri modelli di linguaggio. Introduciamo il concetto fondamentale di

Corpus, che rappresenta una grande collezione di testi generica, scritti in linguaggio naturale (da umani per gli umani), come ad esempio le migliaia di pagine che compongono il World Wide Web, e che quindi fornisce un insieme di parole che possono essere interpretate. Un modello di linguaggio definisce una distribuzione di probabilità su questo insieme di parole, potenzialmente infinito; in pratica, a differenza del modello di linguaggio specifico scelto, da esso viene associata una certa funzione f(w) a ogni parola del lessico, e in base al modello di linguaggio, alle dimensioni del corpus, ai concetti fondamentali di statistica e probabilità (che in questo contesto rappresentano la stessa cosa) e infine all'apprendimento, si riesce a costruire a basso livello tutta quella struttura in grado di poter sostenere poi ad alto livello tecniche e algoritmi di estrazione dell'informazione, traduzione automatica e information retrieval. I modelli di linguaggio associati a una generica parola(w) del lessico sono modelli n-gramma, in particolare:

- Modelli uni-grammi, in cui f(w) = P(w) , probabilità associata a una parola
- Modelli bi-grammi, in cui f(w) = P(w;w-1), probabilità associata a una parola e alla precedente
- Modelli N-grammi, in cui f(w) = P(w; w-(N-1)), probabilità associata a una parola e alle N precedenti

Ovviamente la scelta del modello di linguaggio sarà importante per ottenere un livello più o meno raffinato di interpretazione del linguaggio e di disambiguazione semantica, ma come già detto abbiamo bisogno anche di un corpus con certe dimensioni per avere la possibilità di poter sfruttare al meglio questi modelli, che alla fin fine esprimono delle funzioni (matematiche) che hanno bisogno di molti "valori" sul quale lavorare per funzionare al meglio, e far funzionare gli algoritmi che le sfruttano direttamente: questo concetto, andando sempre più ad alto livello, sta alla base della cosiddetta "machine learning" e dei sistemi software che sviluppano la parte di apprendimento automatico delle macchine, che riconduce un pò tutto quello che stiamo valutando al mondo informatico e alle tecnologie moderne (l'impatto dell'IA nel nostro contesto moderno).
Come stimare la funzione probabilità P(w) , e associarla a una parola? E' essenziale ed è mio obiettivo, come già espresso più volte, cercare di capire in questa parte relativamente tecnica i concetti fondamentali che poi saranno ripresi nell'approfondimento centrale della tesi; ma in questo momento abbiamo bisogno di fare un ulteriore collegamento, seppur molto generale, per trovare delle regole fondamentali di probabilità, proprio perché queste regole verranno riprese nel momento in cui andremo a trasporci verso il mondo informatico, valutandone il significato in relazione all'approccio specifico. Dobbiamo quindi in pratica collegare i modelli a N-Grammi mostrati a un' espressione formale, e per farlo abbiamo bisogno di definire il processo

Markoviano, come un processo nel quale la probabilità di transizione che determina il passaggio ad uno stato di sistema dipende unicamente dallo stato di sistema immediatamente precedente (proprietà di Markov) e non dal come si è giunti a tale stato. Questo concetto è fondamentale per introdurre la catena di Markov, processo che gode della proprietà di Markov appena descritta, con spazio degli Stati discreto (numerabile). Ora, partendo dagli uni-grammi, scrivendo una prima formula relativa a una particolare funzione di probabilità, dovremo ovviamente andare a combinare le varie probabilità di tutte le N parole che abbiamo a disposizione nel corpus: per poter generalizzare e semplificare questo calcolo, andiamo a scrivere la catena attraverso un'approssimazione; a questo punto, approcciandoci verso i bigrammi, dovremo risolvere questa formula sfruttando il concetto di frequenza dei bigrammi nel nostro corpus. In poche parole, questo processo che racchiude queste formule matematiche, mi da la conferma che la funzione di probabilità associata a un bigramma dipende essenzialmente dalla frequenza con cui il bigramma si presenta all'interno del corpus; ecco il perché un corpus più ampio mi da più possibilità di disambiguare il linguaggio naturale e effettivamente associare la parola giusta, in dipendenza delle parole precedenti. Per questo, i bigrammi possono essere rappresentati come catene di Markov: questo concetto è indipedente dall'approccio poi che verrà fatto per implementare una tecnica specifica di NLP, ma mi da solamente il collegamento tra un modello di linguaggio (che viene implementato attraverso algoritmi) e una formula matematica che viene ripresa dalla teoria della probabilità: vedremo più avanti, in alcune analisi di tecniche concrete, come abbiamo bisogno di riprendere questo concetto e magari ridefinirlo, proprio nel momento in cui ci andiamo a trasferire da queste basi matematiche alle vere tecniche informatiche che le implementano.

Per chiudere questa sezione, introduciamo molto velocemente un ultimo concetto che verrà poi ripreso subito nella prossima sezione in chiave probabilistica: le Context-Free-Grammar(CFG), ovvero le grammatiche formali

$$P(w_1 \ldots w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)\ldots P(w_n \mid w_1^{n-1})$$

$$P(w_n \mid w_1^{n-1}) \approx P(w_n \mid w_{n-1})$$

$$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

$$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}$$

Figura 2.3: *1: probabilità unigrammo; 2: probabilità di un bigrammo; 3: probabilità di un bigrammo con frequenza relativa; 4:probabilità N-grammo*

(struttura astratta che descrive un linguaggio formale in modo preciso, associando attraverso regole formali alcune parole a un determinato alfabeto), definite da 4 tuple : G = (N,T,P,S). Le 4 tuple rappresentano nell'ordine: l'insieme di simboli non-terminali (N), l'insieme di simboli terminali (T), l'insieme delle regole (P) e l'insieme dei simboli di partenza (S). E' importante iniziare a capire il ruolo di queste grammatiche proprio perché danno quelle regole di traduzione e interpretazione tali da trasformare periodi,frasi in dati formali, associandoli a strutture definite dipendenti dalle proprietà del linguaggio a cui appartengono, di cui la grammatica specifica ne ha dettato le regole di traduzione: arriviamo a una struttura "ad albero" dei periodi. Insomma, relativamente a quella scatola nera che mi rappresentava il sistema di parsing sintattico definito in precedenza, le grammatiche CFG rappresentano l'insieme di regole fondamentali con cui questa scatola nera funziona e elabora le frasi che riceve in input, dandone una struttura formale e sintattica. Dunque, dopo aver finalmente definito NLP nel primo capitolo, e aver interiorizzato il ruolo che occupa all'interno di IA, abbiamo ora studiato ed elencato tutti quegli strumenti tecnici che rendono NLP strettamente collegata sia al mondo del linguaggio in sè (parsing sintattico relativo a una parola), sia in maniera più generale strettamente collegata a concetti probabilistici/matematici (modellazione del linguaggio relativo a un corpus, e gestione di periodi, frasi attraverso grammatiche). Andremo dunque ora, con questi concetti bene in testa, ad approfondire in maniera prima teorica e poi concreta i due possibili approcci verso NLP, andando poi nei prossimi capitoli a definire una tecnica moderna(Sentiment Analysis) e riprendendo nell'ultimo capitolo questi concetti di NLP collegandoli a questa tecnica.

2.2 Approccio probabilistico e PCFG

A questo punto, dopo aver definito le grammatiche non contestuali (CFG), siamo in grado di caratterizzare una stringa come appartenente o no a un determinato linguaggio, attraverso regole di parsing: ma quale potrebbe essere un modo per poter decidere se un intero periodo possa appartenere a un determinato linguaggio? Allarghiamo dunque i soggetti del nostro discorso: dobbiamo tradurre in maniera informatica questo passaggio, cercando di capire in quale modo riusciamo a interpretare il linguaggio naturale, dopo aver definito basi comuni nella definizione di grammatiche, di necessità di parsing, di traduzione. Arrivati a questo punto intravediamo due possibili approcci che si sono sviluppati negli anni per poter affrontare e risolvere questo problema: l'approccio probabilistico e l'approccio semantico. In questa sezione analizzeremo il primo approccio, il filone statistico-probabilistico (che qui vogliono dire la stessa cosa), che ha riscontrato un grande successo negli ultimi anni, soprattuto in relazione all'esplosione dei Big Data. Un modello probabilistico del linguaggio definisce una distribuzione di proba-

bilità su un insieme (potenzialmente infinito) di stringhe: nella sottosezione precedente abbiamo definito modelli che funzionano attraverso funzioni probabilistiche, come i modelli unigrammi, bigrammi, n-grammi, e abbiamo determinato che questi modelli sono in grado di interpretare abbastanza fedelmente il linguaggio determinando, attraverso la frequenza relativa associata a una cosiddetta catena di Markov, delle funzioni di probabilità all'interno del proprio corpus: ma dato che realizzare un corpus infinito è impossibile, in quale modo posso generare un modello potenzialmente infinito se ho a disposizione n-word limitate? E' fondamentale capire, in questo contesto, che quello che deve cambiare sono le basi, e che il concetto di CFG definito precedentemente sarà ridefinito e verranno messe a punto le PCFG(Grammatiche non contestuali probabilistiche), che sono a tutti gli effetti CFG con un nuovo, fondamentale requisito: esse associano una probabilità a ogni regola di riscrittura [9]. In poche parole, associamo un'ulteriore probabilità relativa alle regole che compongono i nodi dell'albero sintattico; in questo modo, con le PCFG abbiamo modo di avere conoscenza non solo sulle word interne al corpus, ma anche al di fuori di esso: su queste nuove basi, possiamo intravedere finalmente tecniche ad alto livello, che vedremo nella prossima sottosezione. Riscontriamo però due limiti che devono essere risolti: anche se le PCFG, per come le abbiamo definite, sembra riescano a costruire alberi sintattici completi(seppur sempre dipendenti da una stima, da una probabilità, e non da una certezza assoluta), ci rendiamo conto che formalizzare un albero sintattico non vuol dire che siamo riusciti a cogliere il significato del periodo, la semantica del contesto: e quindi per questo motivo andremo anche a parlare di un ulteriore approccio, semantico, in grado di affrontare questo discorso, e vedremo i pro e i contro anche di questo approccio. Il secondo limite che dobbiamo essere in grado di superare è relativo alle catene di Markov: seppur esse rappresentino una modellazione concreta e abbastanza completa di un primo approccio al linguaggio naturale, è fondamentale in questo contesto probabilistico ridefinire queste catene di Markov, come abbiamo ridefinito le CFG: è qui che iniziamo a parlare di Hidden Markov Models (HMM), o catene di Markov nascoste, il cui significato può risultare ambiguo, ma che se viene contestualizzato in questo discorso, riusciamo veramente a capire quali sono le basi di questo approccio probabilistico. HMM sono effettivamente catene di Markov, che però nascondono lo stato attuale, associando una funzione di probabilità relativamente allo stato in cui si possono trovare: perché, in questo passaggio di ulteriori ridefinizioni, vogliamo rinunciare a un'informazione che con le semplici catene di Markov visibili avevamo? Le HMM sono in grado di associare probabilità a particolari eventi che vengono generati in ogni possibile stato: la gestione di questa duplice funzione di probabilità (stati, eventi), è in grado di modellare un contesto molto più ampio, rispetto a ciò che facevano le semplici catene di Markov. Allargare il contesto, passare da CFG a PCFG, da catene di Markov visibili a invisibili, è il passaggio fondamentale che caratterizza l'approccio probabilistico: PCFG e HMM sono le valvole

da gestire per un modello ad alto livello, e vedremo come nella Sentiment Analysis si passa attraverso questi concetti per poter definire modellazioni probabilistiche di questa tecnologia. Concludiamo dicendo che comunque PCFG e HMM sono sì elementi fondamentali, ma pur sempre caratterizzati da limiti: l'approccio probabilistico, che è in continua evoluzione, ha avuto perfezionamenti e sono stati introdotti nuovi modelli al posto di HMM, soprattutto relativamente alla tecnica di POS-Tagging, che non andremo ad analizzare dettagliatamente qui. Il lettore deve essere consapevole del ruolo fondamentale svolto dalle PCFG e da HMM nell'approccio probabilistico, per poter percorrere criticamente il percorso che verrà affrontato successivamente nel quarto capitolo.

2.2.1 Information Retrieval, Estrazione dell'informazione e traduzione automatica

In questa sottosezione illustreremo brevemente le tecniche informatiche che si sono sviluppate nell'ambito dell'approccio probabilistico, ovvero nell'ambito di una modellazione attraverso PCFG del linguaggio naturale, costruendo funzioni di probabilità a due livelli per poter essere in grado di manipolare un grande flusso di dati. La fortuna dell'approccio probabilistico sta proprio in questo: queste tecniche, che si sono sviluppate soprattutto negli anni '80-'90 con l'avvento del WWW e dei Big Data, sono ai giorni nostri considerate quasi scontate, e ci soffermeremo poco sul loro funzionamento: quello che il lettore deve tenere presente comunque, è il continuo collegamento tra le basi probabilistiche (ridefinite nella precedente sottosezione) e le tecniche implementative che sfruttano queste basi, tecniche che tendono a variare alcuni punti in questi modelli per poter ottenere vantaggi in un certo senso rispetto ad un altro. L'information retrieval (IR), che è la prima tecnica che affrontiamo, consiste nel trovare i documenti rilevanti per le necessità informative di un utente: è un naturale sottoinsieme di NLP, perché esso ha a che fare con una certa funzionalità di NLP. Questa tecnica, perfezionata nel corso degli anni, è fondamentale al giorno d'oggi, basti pensare alla fortuna dei motori di ricerca. L'obiettivo principale di questa tecnica, che è in grado in poche parole di costruire veri e propri sistemi capaci di fare information retrieval, è proprio quello di avere un ritorno ad-hoc dell'informazione: l'utente, attraverso una query, descrive l'informazione che desidera avere, e il sistema IR dev'essere in grado di far avere all'utente una lista inerente all'informazione richiesta dall'utente, o soddisfacendo esattamente le richieste dell'utente(matching esatto), oppure dando in risposta documenti che il sistema ha valutato avere una buona influenza nella query formulata dall'utente (matching stimato, grande flusso di dati). E' molto importante anche capire, valutare, se un sistema IR può avere delle buone prestazioni oppure no: i due parametri che è possibile ottimizzare all'interno di algoritmi IR riguardano la precisione, che misura la proporzione di una parte rispetto a un totale, e la copertura, che misura quanta parte del

corpus non è stata inclusa nei risultati. Andiamo anche a spendere due parole su altre tecniche di NLP probabilistico: l'estrazione dell'informazione, che fa riferimento al processo mediante cui si inseriscono dati in un database esaminando un testo e cercando le occorrenze di una particolare classe di oggetto o evento all'interno di questo DB, e non da meno la traduzione automatica, che fa riferimento all'operazione di traduzione da un testo in lingua naturale a un altro (obiettivo, sorgente), e che forse è una delle tecniche più complicate in questo ambito: si pensi alla parola "hard", che in inglese ha una certa valenza contestuale, ma che in italiano può essere tradotta "forte" oppure "difficile", a seconda del contesto in cui si trova. Allargare il discorso in questo momento diventa fondamentale, e notiamo che l'approccio probabilistico, sebbene determini un passaggio importante e abbastanza lineare nella nostra analisi, può risultare stretto in certi versi, ma sicuramente risulta l'approccio più semplice, lineare, fedele, con il quale gestire una grande quantità di dati in entrata.
Per riassumere: l'approccio probabilistico, caratterizzato a basso livello dal passaggio alle PCFG e agli HMM(e non solo), fornisce tecniche in grado di gestire con una certa fedeltà e linearità grandi moli di dati (come può essere il corpus che rappresenta il linguaggio naturale), associando funzioni di probabilità alle word e alle stesse regole di traduzione: queste tecniche, quali Information Retrieval, Estrazione dell'informazione e Traduzione Automatica, fanno parte di NLP dai primi anni '60-'70, ma ottengono molto successo soprattutto durante gli anni '90, con l'ascesa del WWW e dell'informazione che viaggia sul web: sebbene l'approccio probabilistico abbia molti limiti(e uno di questi è proprio l'impossibilità di associare una traduzione semantica esatta a una traduzione sintattica esatta), esso rappresenta un' arma vincente per l'implementazione di sistemi in grado di gestire le problematiche di NLP e interpretare il linguaggio naturale. Non andiamo ora a mostrare qualche esempio concreto di tecnica probabilistica NLP, ma avremo modo (soprattutto nell'ultimo capitolo) di valutare criticamente una tecnica moderna che segue un modello probabilistico.

2.3 Approccio semantico

Dopo aver valutato e presentato l'approccio probabilistico, andiamo a contestualizzare il secondo possibile approccio all'interpretazione del linguaggio naturale : l'approccio semantico.

2.3.1 Valutazioni e critiche dell'approccio: WSD

La semantica fa riferimento alla parte di linguistica che si occupa del piano del significato di un periodo. E' molto complicato arrivare a delineare una definizione rigorosa del concetto di significato, già a partire da un piano linguistico: un significato può essere rappresentato da un'idea, un'immagine

mentale, e può essere connesso alle sensazioni che esso dà luogo, ma allo stesso momento invece può essere rappresentato attraverso un collegamento oggettivo, cosa rappresenta questo significato a livello di modello[3]. Queste problematiche che stanno alla base di un'interpretazione semantica di un contesto reale, sono problematiche che si ripercuotono fino ad arrivare a problemi che dovremo affrontare nella modellazione di questo mondo, per poter arrivare a definire tecniche come abbiamo fatto nell'approccio statistico. L'approccio semantico si prefigge di sconfiggere quel limite che è imposto da un approccio puramente probabilistico, ovvero quello di non avere certezza assoluta della traduzione semantica di un periodo come lo si ha generalmente di una traduzione sintattica: il limite dell'approccio probabilistico sta nel fatto che tutto ciò che viene svolto a basso livello da algoritmi probabilistici è quello di sottostare a funzioni matematiche (che vengono modellate all'interno di tecniche con PCFG) che descrivono il funzionamento e le regole nella gestione di svariate mole di dati: le formule matematiche non si possono occupare di significato.
Come riuscire a ingegnerizzare il concetto ambiguo di "significato", e interpretare semanticamente il linguaggio naturale? Questo task è tuttora irrisolto nel mondo dell'NLP e in generale nel mondo dell'IA; riuscire a trovare un modo semplice e lineare di modellare questo problema per poter arrivare a una disambiguazione assoluta del significato di un periodo è praticamente impossibile. Quello che possiamo fare è sempre lo stesso ragionamento: proviamo ad allargare i soggetti del nostro discorso. Se è praticamente impossibile modellare regole in grado di tradurre a pieno il significato di singoli periodi, cerchiamo di utilizzare nel nostro discorso regole associate a un intero corpus, sorpassando quegli algoritmi o codici relativi a singoli periodi, word: nell'area dell'apprendimento semantico, ci sono tantissimi usi di tecniche basate sul corpus: alcuni ricercatori hanno usato tecniche empiriche per indirizzare un task complicato a un'interpretazione semantica, cercando di sviluppare accurate regole in grado di dare la giusta interpretazione semantica a un contesto, oppure abbiamo una metodologia empirica nell'affrontare questo tipo di problematiche, producendo dei veri e propri parser (come facevamo nell'approccio probabilistico). Ma la grande novità dell'approccio semantico sta nel Word Sense Disambiguation(WSD), tecnica che caratterizza quest'approccio il cui principale obiettivo è quello di identificare il corretto significato di una parola in un certo contesto[7]. Il funzionamento riprende un pò anche il discorso fatto nelle precedenti sottosezioni per la fase di traduzione NLP: oltre alla parola in input (descritta in linguaggio naturale), al "parser semantico" viene data l'informazione di quale "Part of Speech" (parte del discorso) la parola fa parte (è un nome? è un verbo? un aggettivo?); in output, avremo semplicemente che ogni occorenza di una generica word avrà il proprio tag relativo al suo significato. In poche parole, stiamo arrivando piano piano alla definizione di un processo di analisi semantica che finalmente astrae da algoritmi o codici relativi a un singolo periodo, a una singola word, ma che vanno a investire un intero cor-

pus: però è da evidenziare che mentre nell'approccio probabilistico abbiamo associato funzioni matematiche per poter effettuare questo passaggio e allargare il discorso, qui dobbiamo per forza definire attraverso nuovi modelli un modo di affrontare il linguaggio in maniera più precisa, e per forza di cose più complicata: è qui che entrano in gioco tecniche "machine-learning" , in grado di avvicinare l'elaboratore a un essere pensante e in grado di apprendere automaticamente strategie per poter affrontare problematiche sempre nuove. Il WSD rappresenta proprio la scatola nera semantica che abbiamo definito nelle precedenti sottosezioni quando eravamo ancora in un analisi comune ai due approcci: Supported Vectors, Tag, parti del discorso, rappresentano elementi a cui il WSD deve fare riferimento per poter arrivare a utilizzare tecniche basate sul corpus, arrivando a definire un vero e proprio "parser semantico".

Anche se in realtà il discorso fatto può risultare abbastanza complicato, è importante tenere presente che in una valutazione critica di questo approccio dobbiamo avere in mente che il limite imposto da un approccio probabilistico può considerarsi quasi superato, ma che per andare a realizzare tecniche implementative, sistemi software in grado di riuscire a modellare il problema del WSD(e del semantic parsing), abbiamo un lavoro molto oneroso da fare, a differenza di sistemi che rispecchiano un approccio probabilistico, in grado di collegarsi comunque costantemente a formule e quindi modelli lineari di traduzione. Come in ogni contesto, può risultare favorevole o meno l'utilizzo di un approccio o di un altro a seconda del funzionamento o della tipologia di tecnologia che si vuole realizzare: definiti gli obiettivi, e conoscendo entrambi gli approcci, possiamo criticamente decidere in quale contesto può essere vantaggioso usare il filone probabilistico o il filone semantico. Interiorizzati entrambi gli approcci che si sono sviluppati negli anni relativamente al problema dell'interpretazione del linguaggio, e studiati gli elementi cardine sul quale andare a modellare e implementare sistemi, possiamo finalmente passare a studiare l'impatto di questi concetti

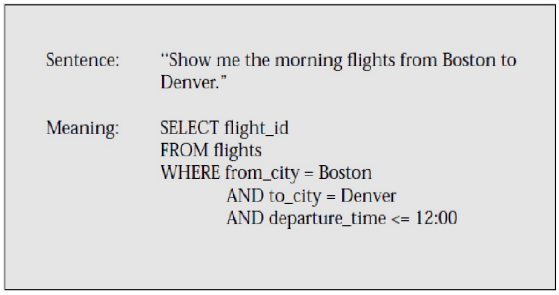

Figura 2.4: *Un esempio di semantic parser presa da una query relativa a un DB*

nel contesto sociale moderno, dopo aver illustrato a titolo informativo un esempio di sistema semantico-NLP.

2.3.2 BabelNet e BabelFly

In questa sottosezione vorrei illustrare a titolo informativo due sistemi basati sui concetti appena espressi, ovvero tecnologie che si occupano di gestire sistemi in grado di interpretare semanticamente il linguaggio naturale. Roberto Navigli, Professore dell'Università della Sapienza di Roma, responsabile del dipartimento di Linguistica Computazionale, gestisce da più di 5 anni un progetto chiamato "multiJedi", inerente proprio alla creazione di risorse lessicali a larga scala e alla comprensione di testi in diverse lingue. Nell'ambito di questo progetto, Navigli ha contribuito insieme ad altri ricercatori a sviluppare varie tecnologie e sistemi software che andassero proprio a riprendere le basi che abbiamo discusso prima inerenti a NLP, come approccio semantico, comprensione del linguaggio, andando a costruire tecnologie user-friendly in grado di gestire ad alto livello questi argomenti: in particolare, citiamo BabelNet e BabelFly. BabelNet è una rete semantica computazionale con copertura a larga scala, che riprende ingressi lessici da WordNet, uno dei sistemi di computazione linguistica più importanti creato negli anni '90, e ingressi enciclopedici da Wikipedia. BabelNet ha una copertura a larga scala, copre 271 linguaggi diversi, e gestisce più di 300M di relazioni semantiche: gestisce le entries (ingressi) che vengono da WordNet e da Wikipedia, cercando di modellarle sottoforma di concetti e di nomi-entità. Quello che rimane, ancora, è un forte carattere di ambiguità, legato all'ambiguità di queste entries (per l'ingresso "calcio" intendiamo la sostanza chimica, o lo sport?)

BabelFly è un sistema che racchiude disambiguazione multilingua e collegamenti diretti alle entità: con la stessa interfaccia utente di BabelNet, esso rappresenta la parte in grado di disambiguare le word in entrata sulla rete di BabelNet, con la possibilità di linkarsi direttamente all'entità di

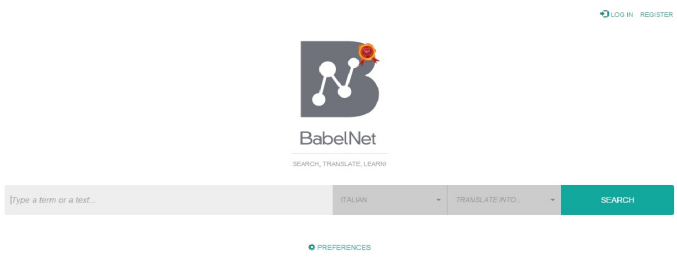

Figura 2.5: *UI di BabelNet*

cui l'utente fa richiesta. Questo processo di disambiguazione ovviamente procede per step, partendo da una signature semantica, per poi passare al trovare ogni possibile significato della word(ambiguità), connessione dei vari significati trovati, estrazione di un grafo logico, e infine selezionare i significati più inerenti. A livello di API, troviamo nel sito diverse tecnologie con le quali costruire query in grado di interagire con il sistema BabelNet: HTTP query, JAVA query, SPARQL query. Nella libreria inerente a Java, ad esempio, vediamo che sono presenti tre classi principali:

- BabelNet: la classe principale, che rappresenta l'entry point sulle risorse BabelNet: questa classe è implementata con il Singleton Pattern
- BabelSynset: classe che rappresenta un set di lessici multilingua che hanno una caratteristica comune: ad esempio, potrei essere interessato ad avere in mano tutti i termini che fanno parte di un certo discorso
- BabelSense: è la classe che rappresenta la singola word presente nello specifico BabelSynset

Con queste API gerarchiche, riusciamo a interfacciare (attraverso IDE come Eclipse o NetBeans) sistemi software in grado di interagire con BabelNet [1]. Inutile ribadire qui che con l'approccio a oggetti (Java, Scala) abbiamo l'enorme possibilità di fare query riusabili, estendibili, e quindi rimanere sempre a un buon livello di astrazione dalla tecnologia basso livello. Nell'ultima figura, visualizziamo un esempio d'uso in cui ritrovo l'ID del wikidata(dato di wikipedia) per ogni BabelSense presente in un BabelSynset.

```java
import it.uniroma1.lcl.babelnet.BabelNet;
import it.uniroma1.lcl.babelnet.BabelNet;
import it.uniroma1.lcl.babelnet.BabelSense;
import it.uniroma1.lcl.babelnet.BabelSynset;
import it.uniroma1.lcl.babelnet.BabelSynsetID;
import it.uniroma1.lcl.babelnet.InvalidBabelSynsetIDException;
import it.uniroma1.lcl.babelnet.data.BabelSenseSource;
import java.io.IOException;

public class Example {
    public static void main(String[] args) throws IOException, InvalidBabelSynsetIDException {
       BabelNet bn = BabelNet.getInstance();
       BabelSynset by = bn.getSynset(new BabelSynsetID("bn:00000288n"));
       for (BabelSense sense : by.getSenses(BabelSenseSource.WIKIDATA)) {
           String sensekey = sense.getSensekey();
           System.out.println(sense.getLemma() + "\t" + sense.getLanguage() + "\t" + sensekey);
       }
```

Figura 2.6: *Usage example*

Capitolo 3

Sentiment Analysis e IA

Dopo una breve contestualizzazione dello scenario moderno rappresentato da un'economia data-driven, andremo a definire in questo capitolo il ruolo della Sentiment Analysis in relazione ai concetti di NLP e IA descritti precedentemente, e le sue principali caratteristiche.

3.1 Economia data-driven e Big Data

Andiamo ora a descrivere in particolare lo scenario moderno su cui l'Intelligenza Artificiale opera attraverso le sue caratteristiche, descrivendo anche in maniera generale le tecnologie informatiche che ne fanno parte[4].
Nel contesto moderno informatico siamo dominati da un concetto di economia data-driven, ovvero una nuova modalità di approccio verso la realtà attraverso la gestione di un flusso di dati in ingresso, che si presenta con una mole sempre più importante: i cosiddetti Big Data, che fanno riferimento a data-set in quantità di volume enormi derivanti da fonti diverse (social media, mobile, web), che arrivano alle aziende con velocità a cui non sono mai state abituate. Capiamo subito dunque che servono nuovi approcci all'analisi del concetto di dato: in questo scenario occorre analizzare il concetto di "dato" in maniera ingegneristica, soprattutto in relazione a ciò che ci offre la modernità, e solo dopo potremo descrivere in che maniera il fenomeno di economia data-driven sia in grado di generare nuovi mestieri che riguardano questo ambito, valutando il suo impatto profondo.

3.1.1 Ruolo dei dati e figura del Data Scientist

I dati e il software sono andati sempre di pari passo nella storia dell'informatica: il dato rappresenta l'unità fondamentale con la quale un sistema software può interagire, prendendolo in ingresso e analizzandolo attraverso meccanismi più o meno ingegnerizzati. Nel momento in cui, intorno agli anni '90, abbiamo avuto l'esplosione del fenomeno Web, abbiamo iniziato a capire quale fosse il fondamentale ruolo dei dati all'interno delle nostre analisi software, come ad esempio pagine web, pagine utenti, query su DB, con

l'esplosione di quel trend che venne chiamato "Web Analysing", fornendo una nuova figura nel mondo del lavoro. La rivoluzione tecnologica che domina la società moderna ha fatto sì che il progresso oggi fornisse nuovi scenari con i quali andare a interfacciarsi: tecnologie e app mobili, tecnologie social media, hanno dato quella spinta tecnologica in cui anche il concetto base di "dato" deve essere per forza ridefinito, riscrivendo di conseguenza il significato di software, che deve adeguarsi a questo nuovo concetto. Il software sta diventando sempre più interdisciplinare nel nostro contesto: negli ultimi anni, coadiuvato anche dalla grande ascesa dei social media (Facebook, Linkedin, Twitter) e da una sempre più informatizzazione a livello globale, come abbiamo discusso nell'introduzione del nostro percorso, esso sta decisamente invertendo la propria tendenza: da materia di nicchia, oggi il Software è diventato a tutti gli effetti una materia interdisciplinare. Bioinformatica (simulazione di calcolo su PC, esami in laboratorio), geografia(gps e software di gestione spazi, Google Maps), sono soltanto due degli esempi più lampanti in cui ci rendiamo conto che al giorno d'oggi il software ha un ruolo molto più predominante di quello che gli associavamo fino a non molto tempo fa. Possiamo senz'altro generalizzare e definire il software come uno dei vettori che guidano l'analisi della società che viviamo, perché esso stesso la costruisce: device, ambienti, infrastrutture che popolano la realtà, sono i nuovi sistemi software che la caratterizzano. I dati sono le informazioni che questi dispositivi scambiano con l'ambiente reale, e dunque riuscendo a interpretare questi dati attraverso le tecnologie informatiche che conosciamo, si può dare senz'altro un'interpretazione della realtà stessa: non stiamo parlando solamente di Data Mining, ovvero dell'interpretazione dei dati che ci arrivano dai dispositivi, ma addirittura di Reality Mining, ovvero interpretazione della realtà stessa. E' in questo contesto che nasce la figura di "Data Scientist", che rappresenta un esperto del settore, che non solo è in grado attraverso competenze informatiche di dare un significato a questi dati, ma che avendo anche competenze di Business Intelligence, riuscirà a dare un significato contestuale per aiutare aziende o imprese a sviluppare nuovi meccanismi per primeggiare sul mercato: un mercato che non è più solo fatto da umani, ma che vede come protagonisti soprattutto elaboratori, devices, agenti razionali. E' proprio questa la nuova frontiera dell'intelligenza artificiale: i dati sono diventati l'interfaccia del mondo reale. Oltre alle considerazioni fatte, bisogna anche capire che le tecnologie moderne modificano in maniera definitiva i concetti di tempo e spazio. Nell'era moderna tutto accade in tempo reale, e un gran flusso di informazioni viaggia online costantemente: ma il tempo reale lato macchina è un concetto profondamente diverso dal tempo reale lato utente (concepito dagli esseri umani). Inoltre, non esiste più oramai la netta divisione "realtà virtuale" da "realtà fisica", perché ci rendiamo sempre più conto che i dispositivi e i vari dati che attraversano il mondo online al giorno d'oggi influenzano eccome la realtà fisica (dispositivi wearable, gps, tecnologie opinion mining..); per questo motivo, possiamo tranquillamente affermare che il codice fa muovere la realtà in

cui noi viviamo, con i suoi tempi e spazi, e ha il potere di modificare lo spazio fisico in cui è ambientato. Pensiamo ad esempio a una grossa catena di vestiti che mi propone un'offerta nel momento in cui io sto transitando proprio davanti al negozio; oppure, a tutti quei nuovi metodi di "analisi predittiva", che in base al controllo costante dei dati generati da un certo cliente (ad esempio relativamente al mondo dell'utenza telefonica), siamo in grado di capire attraverso alcuni suoi comportamenti se questo cliente ha intenzione di disdire il proprio abbonamento telefonico oppure no, e in questo modo riuscire ad anticipare una disdetta che a posteriori sarebbe molto più complicato gestire. Il mercato sta cambiando, ed è essenziale capire che ci saranno nuovi lavori legati a questi fenomeni, che quindi è essenziale saper analizzare: umani e algoritmi sono i nuovi protagonisti in simbiosi tra loro del mercato. Un esempio particolare in cui è possibile riscontrare questi concetti è il mondo dei Social Network e la loro moderna ascesa, definiti come servizio di rete sociale che consente la gestione dei rapporti sociali, facilitando la comunicazione e la condivisione di informazioni digitali.
E' in questo contesto che vediamo proprio il binomio umani-agenti razionali protagonisti di questi fenomeni, governati sempre da quei concetti IA che abbiamo spiegato abbondantemente nel capitolo precedente: ed è in questo contesto che vediamo le possibilità migliori per il futuro.

3.1.2 Intelligenza semantica

Ma quali sono in concreto i nuovi scenari che questo fenomeno del data-driven porta nella nostra società? Oltre al discorso che riprenderemo nelle vere e proprie conclusioni di questo trattato, inerente a una nuova ridefinizione del ruolo dell'ingegnere informatico all'interno delle aziende, ci basti pensare che in questo nuovo contesto molte azioni e molte operazioni che sono state sempre svolte con una certa naturalità da esseri umani, saranno svolte in maniera funzionale da agenti razionali. Non sarà più necessario ad esempio conoscere un posto o chiedere indicazioni per raggiungere una certa località, abbiamo i software di navigazione (es. Google Maps); oppure ad esempio non ci sarà più bisogno di svolgere tradizionali e onerose ricerche di mercato per valutare l'opinione dei consumatori su di un prodotto, avremo tecnologie ad-hoc che svolgeranno questo compito. Tutto ciò è solamente possibile grazie al nuovo ruolo dei dati nel contesto moderno: i dati sono una valvola di traduzione della realtà, e saperli gestire attraverso competenze informatiche adeguate (infrastrutture, linguaggi di programmazione, sistemi relazionali classici e non relazionali, e ovviamente conoscenze IA), fornisce una potentissima capacità: la capacità di "interpretare" la realtà. Le aziende devono cambiare il loro modo di decidere in base a questi nuovi scenari: per essere protagonisti nel mercato, occorre sempre più abbandonare vecchi preconcetti e passare invece a fare decisioni real-time, analisi predittiva e intercettazione dei commenti, per migliorare sempre più velocemente i propri prodotti e avere un ruolo predominante nel nuovo mercato che si sta for-

mando, dominato dal binomio uomo-agente razionale. Ed è proprio qui che nasce il fenomeno principale del nostro discorso: la Sentiment Analysis (o Opinion Mining), che rappresenta in poche parole l'attività di identificazione, elaborazione e classificazione di informazioni legate ad un brand o ad un argomento attraverso software di elaborazione del linguaggio e linguistica computazionale allo scopo di determinare l'attitudine di chi ha pubblicato e la polarità contestuale del contenuto (positiva, neutra, negativa): ecco uno dei nuovi scenari che noi andremo a sviluppare più dettagliamente nel prossimo capitolo. Più in generale, parliamo proprio di una "Intelligenza Semantica", ovvero della possibilità di tradurre i dati e dargli un significato contestuale per facilitare il lavoro di promozione di un prodotto da parte dell'aziende: ad esempio, se in un forum inerente "Impianti fotovoltaici" riesco a captare le opinioni e le preferenze degli utenti rispetto all'argomento(interesse al design, interesse ai consumi, interesse alle spese), io riesco a capire che il cliente X è più interessato a una certa caratteristica rispetto a un'altra, alla quale ad esempio è interessato il cliente Y. In poche parole, l'economia data-driven, oltre a generare un fenomeno molto grande che ridefinisce il significato stesso di dato e software, produce all'interno delle aziende nuovi tipi di lavoro legati a un nuovo mercato dominato da uomini e agenti razionali, che generalmente si combinano nell'obiettivo di perseguire una certa Intelligenza Semantica: per rimanere dominanti in questo nuovo mercato, le aziende hanno bisogno di investire in queste nuove figure. Sentiment Analysis rappresenta proprio un'attività legata all'intelligenza semantica, che viene sviluppata soprattutto all'interno di Social Network e forum, dove vediamo confluire molti utenti e quindi clienti.

3.2 Sentiment Analysis

Arrivati a questo punto, e dopo aver definito il contesto e l'ambiente a cui facciamo riferimento, andiamo a dare una definizione rigorosa del fenomeno Sentiment Analysis. Definiamo Sentiment Analysis (o Opinion Mining) la tecnica che è in grado di catturare la vantaggiosità di un insieme di documenti tramite tecniche di NLP: questa tecnica fa una sorta di classificazione a seconda della polarità di questi documenti, commenti, dati(positiva, negativa, neutra), riuscendo a interpretare, determinare come sta andando un certo prodotto o un certo brand in relazione alle opinioni degli utenti. E' naturale intendere che la Sentiment ha numerosissime applicazioni nel mondo soprattutto dei Social Network, dove abbiamo la possibilità di avere un enorme flusso di utenti e dati sui prodotti; ma allo stesso momento la ritroviamo nel mondo dei blog, dei forum, insomma in relazione a tutto ciò che è raggiungibile online (Big Data, Internet Of Things). La Sentiment Analysis, se modellata bene, può rappresentare un enorme strumento di vittoria nel mercato di oggi, perché rappresenta una nuova frontiera nel mondo dell'economia data-driven, ma anch'essa avrà i suoi punti di forza e i suoi punti

deboli, che sono in costante fase di sviluppo e studio: la Sentiment Analysis è ancora una tecnica in fase di cantiere, di sviluppo, come spiegheremo più avanti.

3.2.1 Caratteristiche e problematiche

La Sentiment Analysis è caratterizzata da molti fattori, e alcuni di essi rendono questa tecnica molto difficoltosa. Ad esempio, consideriamo la caratteristica principale della Sentiment Analysis: la possibilità di catturare la polarità di un certo commento, che significa essere in grado di interpretare la polarità di un periodo: come fare? Potremo dire ad un primo livello di analisi che potrebbero esistere delle parole chiave nel linguaggio naturale che determinano senz'altro un'evidente polarità, positiva ad esempio con gli aggettivi "bello", "meraviglioso", "funzionale", oppure negativa con "orribile", "brutto", "disprezzo". Dunque, potremo semplicemente dire che per fare detecting di una certa polarità basta semplicemente fare riferimento a una lista di parole chiavi, che sono legate a un trend positivo, negativo, neutro, e controllare la loro presenza nel dato che vuol'essere interpretato. In realtà, anche se volessimo ridurre in maniera banale il problema del detecting della polarità di un certo commento, facendo riferimento agli studi condotti dal ricercatore Pang[2], uno dei massimi esponenti di questo fenomeno, troviamo diversi problemi nel decidere di comune accordo quali parole effettivamente rappresentino una polarità positiva e quali termini rappresentino una polarità negativa. Nell'esempio proposto nella pagina successiva, vediamo due uomini che propongono una lista di parole relative a un corpus che rappresentano per loro delle polarità rispettivamente negative e positive, e mettiamo in confronto l'accuratezza (precisione) di questi set con un set elaborato attraverso un'analisi statistica che abbiamo discusso nei capitoli precedenti da parte di un agente: notiamo che abbiamo una precisione più elevata (69 per cento), e inoltre vediamo che le parole che fanno riferimento alle polarità sono diverse. Per questo ci rendiamo conto che non basta solamente una semplice formulazione di una lista di parole che fanno riferimento a sentimenti positivi, negativi, per risolvere questo task: avremo senz'altro bisogno di modelli, e questi modelli dovranno prendere in esame senz'altro un intero corpus di dati, e quindi capiamo l'importanza di trasporre la Sentiment Analysis attraverso dei modelli il più possibile basati sul corpus. Ovviamente avremo anche altri problemi relativamente alla Sentiment Analysis, basti pensare alla classificazione dei documenti, e all'attribuzione di commenti a un utente specifico, determinando il comportamento di questo utente per prevedere ad esempio alcune sue mosse e interessi(predictive analysing, o ad esempio gli algoritmi di Facebook, Twitter, che analizzano le immagini postate dagli utenti, cercando di capire i loro interessi etc.) In rete mostriamo veramente chi siamo attraverso i dati che scambiamo attraverso i Social Media.

	Proposed word lists	Accuracy (%)
Human 1	positive: *dazzling, brilliant, phenomenal, excellent, fantastic* negative: *suck, terrible, awful, unwatchable, hideous*	58
Human 2	positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting* negative: *bad, cliched, sucks, boring, stupid, slow*	64
Statistics-based	positive: *love, wonderful, best, great, superb, still, beautiful* negative: *bad, worst, stupid, waste, boring, ?, !*	69

Figura 3.1: *Esempio di Pang: elaborazione umani-agenti lista parole chiave nel detecting di una polarità*

3.2.2 Ruolo di IA e NLP nella Sentiment Analysis

Diventa fondamentale arrivati a questo punto, dopo aver capito lo scenario moderno data-driven e aver inteso a fondo il senso e il significato della Sentiment Analysis, capire in quale modo alcune di quelle tecniche elencate nel capitolo precedente relativamente ai problemi di NLP vengono riprese e modellate in questo scenario: vedremo poi, nell'ultimo capitolo, in che maniera valutare Sentiment Analysis su Twitter attraverso gli approcci probabilistici e semantici, quali modelli vengono proposti e in che maniera possiamo sfruttare elementi a nostro vantaggio, valutando criticamente questi modelli. Uno dei primi problemi , oltre il detecting della polarità(associata ad un approccio statistico e alla frequenza relativa discussa in ambito NLP), è quello di riconoscere quale parte del commento rappresenta l'elemento soggettivo(e quindi l'opinione diretta dell'utente) e quale l'elemento oggettivo: proviamo a fare riferimento quindi a tutti quei discorsi fatti sul Part Of Speech (parte del discorso) in ambito NLP, per poter capire quale parte di periodo rappresenta la parte soggettiva e quale la parte oggettiva dell'analisi. Questo problema per niente banale ha visto molti ricercatori studiare alcune possibilità: Hatzivassiloglou e Wiebe (2000) diedero una prima analisi al fenomeno perseguendo l'obiettivo di giudicare se la parte di un commento fosse soggettiva o no attraverso gli aggettivi che comparivano nel commento; attraverso molti progetti, essi definirono il detecting della soggettività di un commento attraverso caratteristiche chiave , legate al detecting di questi aggettivi. Ma nel caso in cui devo interpretare un commento neutrale, nel quale dunque non riscontro questi aggettivi, come faccio a capire che effettivamente è un commento neutrale e quindi una parte soggettiva, invece che una parte oggettiva? Cioè, come faccio a distinguere un commento neutrale da una parte oggettiva (quindi una parte che non riguarda un opinione), se in entrambi i casi non ho aggettivi a cui fare riferimento? Ecco dove ripren-

diamo per filo e per segno il discorso svolto nel secondo capitolo: abbiamo bisogno di allargare i soggetti del discorso; non possiamo generalizzare tutta la problematica del detecting delle varie parti di un commento attraverso il collegamento con singoli aggettivi, singole word, ma abbiamo il bisogno di avere la possibilità di riconoscere parti intere di un discorso, tradurre periodi, per poter effettivamente avere una precisione alta e un ambiguità relativamente bassa. Ecco ancora una volta l'importanza di avere modelli basati su un intero corpus, ed ecco l'importanza di avere approcci NLP in grado di generalizzare e allargare il discorso: POS Tagging, Part of Speech, sono tutte tecniche che non abbiamo descritto in maniera dettagliata, ma di cui abbiamo capito il senso e il significato in relazione al loro sviluppo: è proprio in questo contesto che queste tecniche verranno utilizzate a basso livello, andando a gestire le caratteristiche chiave della Sentiment. E' importante, a titolo di esempio, capire anche che l'interpretazione di periodi, di parti del discorso, ci permettono di fare classificazione dei commenti, e quindi ogni documento analizzato fa parte di una certa materia soggettiva: in questo modo, algoritmi di ricerca sul mercato potranno lavorare con documenti che fanno tutti parte di un certo argomento, facilitando di gran lunga queste ricerche, velocizzando il lavoro di sviluppo all'interno dell'azienda, e ridefinendo anche il ruolo del customer-care, che può effettivamente runtime capire valutazioni, critiche, problemi dei singoli utenti e andare a risolvere questi problemi nella maniera più efficiente possibile. La Sentiment Analysis, la cui modellazione definitiva è ancora in fase di cantiere, rappresenta senz'altro uno degli spiragli futuri dell'Intelligenza Artificiale applicata nel contesto dell'economia data-driven, che deve maturare in ambito aziendale.

Capitolo 4

Sentiment Analysis e Twitter

Andiamo in questo capitolo conclusivo a descrivere il fenomeno Sentiment Analysis all'interno di Twitter, illustrando e valutando due diverse tipologie di modelli proposti nell'affrontare questo tipo di tecnica, criticando costruttivamente i pregi e i difetti di questi modelli.

4.1 Il ruolo della Sentiment Analysis in Twitter

In questa prima sezione andiamo a ripercorrere brevemente la storia della Sentiment Analysis in relazione allo sviluppo dei suoi modelli all'interno di Twitter, dopo aver contestualizzato l'ascesa di questo Social Network nella nostra società. Questa analisi getterà le basi ai due modelli che presenteremo nelle successive sezioni di questo capitolo.

4.1.1 Contesto di Twitter e prospettive

Twitter è un servizio gratuito di social networking e microblogging, creato nel marzo 2006 dalla Obvius Corporation di San Francisco, ed è ad oggi una delle rete sociali più usate di tutto il mondo. Principale antagonista del colosso Facebook, è utilizzata da milioni di utenti che ogni giorno la usano per condividere e visualizzare le informazioni più disparate: opinioni, informazioni, commenti, e molto altro. Il social network è strutturato in maniera tale che è caratterizzato dalla pubblicazione di brevi messaggi, denominati "tweets", che possono contenere opinioni, immagini, riferimenti e link multimediali; è possibile seguire un gruppo di utenti oppure venire seguiti, attraverso i gruppi sociali "Followers" e "Following", definiti per ogni utente del servizio. Twitter nasce come servizio di comunicazione unilaterale, in cui un utente scrive e un altro legge, ma nel tempo ha stabilito anche una sorta di "conversazione": attraverso un apposito tag (@) è possibile menzionare un altro utente, e attraverso un altro tag(#) è possibile creare un topic, un argomento, un dibattito inerente a una qualsiasi tematica: il

Figura 4.1: *Simbolo di Twitter: è in questo servizio che troviamo il territorio migliore per fare Sentiment Analysis*

concetto di hashtag all'interno di Twitter è fondamentale per lo sviluppo della Sentiment Analysis su questo servizio.

4.1.2 Sentiment Analysis in Twitter

Considerando gli enormi numeri di Twitter, come i più di 200 milioni di utenti iscritti, tra cui ovviamente anche svariate aziende e marchi noti, oltre 500 milioni di tweets scambiati ogni giorno e poco meno di 6000 tweets ogni secondo, è facile intendere che c'è un enorme potenziale informativo in questa mole di dati: sicuramente sarà molto difficile estrapolare, interpretare questi dati, considerando la loro quantità. In relazione al concetto espresso nel precedente capitolo, questi dati possono essere visti al giorno d'oggi come la valvola che definisce la realtà, e interpretare queste informazioni significa interpretare la realtà: in particolare, attraverso la Sentiment Analysis, interpretiamo le opinioni degli utenti relative a un topic, forum, argomento. Twitter e i Social Network in generale infatti sono ottimi ambienti per condurre indagini di vario tipo, proprio grazie alla semplicità del loro funzionamento: l'utente può in ogni momento scrivere ciò che gli passa per la mente, e grazie alla condivisione tematica, in Twitter l'utente si ritrova a interagire con parenti e/o colleghi, e inoltre può esprimere opinioni e valutazioni inerenti a un certo topic o a certi prodotti all'interno del loro contesto, scambiando opinioni con altri utenti sui medesimi argomenti. Da queste opinioni personali possono uscire senz'altro informazioni utili a determinare il grado di soddisfacimento dei clienti; la novità è che queste informazioni, se tradotte e interpretate con sistemi ad hoc, possono essere maneggiate in maniera real-time, facilitando il lavoro di indagine e velocizzando le decisioni dell'azienda sul mercato: più veloci, più completi, più profitti. Con precise API (che è possibile scaricare online), è possibile

recuperare i Tweet, e dopodichè, attraverso sistemi in grado di utilizzare e implementare quelle tecniche di NLP di cui abbiamo parlato prima, siamo effettivamente in grado di interpretare commenti e fare detecting delle opinioni espresse dagli utenti. Ora che abbiamo interiorizzato il perché sia così importante sviluppare tecniche di NLP in relazione a questo fenomeno di Sentiment Analysis all'interno di Social Network come Twitter, facilitando la supremazia e il controllo di un'azienda sul mercato, sviluppando analisi predittiva e velocizzando tempi decisionali relativi alle opinioni degli utenti, dobbiamo valutare criticamente in quale maniera ad alto livello approcciarci a questo fenomeno: quali vantaggi mi può dare un approccio semantico a differenza di uno probabilistico, o viceversa? Quale potrebbe essere il modello giusto per definire la Sentiment Analysis, che tuttora rimane una tecnica che non viene definita attraverso un modello univoco, ma con svariati approcci? Il nucleo di queste valutazioni è da trovare nelle prossime due sezioni, in cui illustreremo le caratteristiche di due approcci di Sentiment Analysis su Twitter, uno classico (analisi semantica), e uno invece relativo a una ricerca svolta nel 2014, probabilistico e molto più moderno.

4.2 Sentiment Analysis semantica in Twitter

L'approccio semantico che caratterizza attraverso alcuni modelli la Sentiment Analysis in Twitter parte da una certa direzione di ricerca: ci preoccupiamo di identificare, all'interno di Twitter (ma il discorso vale anche per altri tipi di blog, forum, etc.), quali sono le peculiarità semantiche da aggiungere a un modello di identificazione semantica, come ad esempio il ruolo di hashtag, di ripetizioni di carattere, emoticon: ci preoccupiamo di trovare un nuovo set di caratteristiche che derivino dalla rappresentazione concettuale semantica delle entità che appaiono nei tweets. Queste caratteristiche fanno riferimento ai concetti semantici che rappresentano le entità estratte dai tweets. La ragione per la quale un approccio semantico di questo tipo parte proprio da questa operazione sta nel fatto che certe entità e concetti tendono ad avere una più forte consistenza con sentimenti positivi e negativi: conoscere queste relazioni ci aiutano senz'altro a determinare la polarità e i sentimenti semantici relativi alle entità, alzando di gran lunga la precisione e l'accuratezza dell'indagine. L'obiettivo principale che dobbiamo sempre avere in testa infatti è quello di ricercare la miglior precisione e accuratezza nell'interpretazione di queste entità: conoscendo i loro legami semantici, siamo di gran lunga avvantaggiati all'interno degli algoritmi di NLP. Esistono diversi tool che sono in grado di estrapolare queste entità e interpretare i dati in questo modo: AlchemyAPI, Zemanta e OpenCalais. Nel lavoro proposto da Hassan, He [6], andiamo ad analizzare proprio attraverso un esperimento l'efficienza di questi tool e il miglioramento dell'accuratezza grazie ai modelli semantic-purpose che li caratterizzano, in relazione a tre diversi data-set forniti in input:

- Stanford Twitter Sentiment Corpus(STS): Dataset che è formato da 60K tweets casuali, caratterizzati da emoticon, alcune negative e alcune positive;

- Health Care Reform (HCR): Dataset che è formato da commenti contenenti l'hashtag # HCR (riforma della salute), dove sono stati eliminati commenti neutrali ma solamente raccolti trend positivi e negativi

- Obama-McCain Debate (OMD): Dataset che è formato da tutti quei tweets inerenti allo scontro per le presidenziali USA tra Obama e McCain: anche qui sono stati eliminati commenti neutrali o misti.

Dopo aver definito più o meno le caratteristiche del nostro data-set che abbiamo in ingresso, andiamo ora a definire la strada per applicare una metodologia di approccio semantica al problema della traduzione di questi dati (massimizzando la precisione e l'accuratezza dell'indagine Sentiment). Come detto prima, i concetti semantici delle entità che vengono estrapolati dai tweets possono essere usate per misurare la correlazione generale tra gruppi di entità (ad es: l'insieme dei prodotti Apple), con una polarità data; ma dobbiamo capire a questo livello, che l'analisi semantica non va solamente a descrivere le entità visibili dal tweet in questione, ma si prefigge di andare oltre, stabilendo tutte quelle connessioni semantiche che fan sì di riuscire a costruire legami tra entità in grado di descrivere la polarità anche di entità non ancora mai incontrate, ma che fanno parte di un certo gruppo semantico. Esempio lampante: se della frase "Finally, I got my iPhone!" , noi riusciamo a estrapolare semanticamente la parola iPhone, che da sola non esprime essenzialmente una polarità, ma che sapendo che è collegata direttamente al gruppo "Prodotti Apple", e che questi prodotti sono stati classificati in generale ricchi di polarità positive (su algoritmi e stime probabilistiche), allora viene associato generalmente il termine iPhone a un trend positivo: ecco il cuore dell'analisi semantica e del modello proposto in questo documento, dove abbiamo la ricerca di caratteristiche semantiche in grado di approcciarsi con la Sentiment Analysis.

Quindi, oltre ad avere i data-set in ingresso nello spazio originale di input, nell'analisi semantica abbiamo bisogno di un informazione aggiuntiva che riguarda le relazioni contestuali semantiche delle entità: come le inseriamo nella nostra analisi? Facciamo riferimento a due metodi, aumento e interpolazione. Con il metodo dell'aumento, noi andiamo a inserire nello spazio input tutti quei concetti e tutte quelle caratteristiche semantiche, aumentando di gran lunga la grandezza del vocabolario e del corpus in generale: anche se ciò, come accennato prima, può essere un fattore positivo (più termini nel corpus, più machine learning), in relazione alla Sentiment Analysis ci rendiamo conto che la precisione può diminuire di gran lunga, e aver utilizzato approccio semantico non ci è servito quasi a nulla. Per questo viene molto meglio utilizzare il metodo dell'interpolazione; non an-

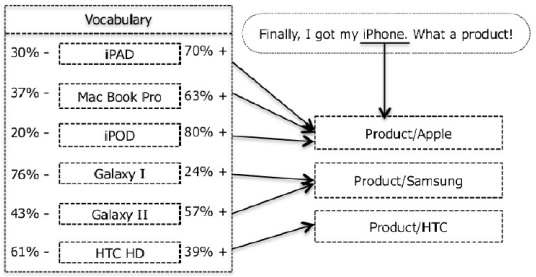

Figura 4.2: *Schema del cuore dell'analisi semantica di un tweet*

diamo direttamente ad aggiungere tutti quei concetti semantici all'inizio della nostra analisi (in input), ma andiamo a definire attraverso una formula precisa la regola sotto la quale devono essere trovati e modellati questi concetti (runtime), che riguardano concetti semantici, sequenze di Part Of Speech, argomenti di sentiment; le API che dovranno implementare questo approccio dovranno essere molto elaborate, ma almeno in questo modo non perdiamo nulla in quanto precisione e accuratezza dell'analisi. In particolare, riprendendo i data-set illustrati precedentemente, applicando questi due metodi e anche un terzo, il replacement, che consiste semplicemente nel rimpiazzare l'entità di un tweet con il suo valore semantico, riscontriamo ciò che viene illustrato nella tabella in figura.

Interpretando questo schema, possiamo fare valutazioni critiche di vario tipo: concludiamo in generale che è bene investire in un'analisi semantica quando l'azienda o un marchio vuole interpretare, con una precisione alta (magari facendo dipendere dei costi di manutenzione, sviluppo, in relazione

Method	STS	HCR	OMD	Average
Semantic replacement	74.10	61.35	71.25	68.90
Semantic augmentation	77.65	63.65	72.70	71.33
Semantic interpolation	**83.90**	**66.10**	**77.85**	**75.95**

Figura 4.3: *Schema tabella data-set*

a questi dati), anche un numero minore di tweet; ma ci rendiamo conto che il limite grosso sarà implementare sistemi in grado di gestire, oltre la mole di dati provenienti dalla realtà, anche tutte quelle connessioni semantiche di cui dobbiamo tenere traccia nel processo di interpretazione delle opinioni.

4.3 Sentiment Analysis probabilistica in Twitter: modello basato sul corpus

A questo punto, dopo aver mostrato rapidamente un modello di approccio semantico alla Sentiment Analysis su Twitter, cercando di dare le basi per effetture analisi critiche inerentemente a quest'approccio, andiamo a valutare una metodologia duale, riprendendo anche concetti descritti già nel secondo capitolo dell'elaborato: approccio statistico-probabilistico. Valutiamo un modello che investe questo approccio proponendo un modello portato avanti da ricercatori e ingegneri italiani(Vanzo,Croce, Basili) [12], che cercano di allargare il discorso e definire un modello basato sul corpus, e non sul singolo dato,tweet. Molti degli articoli scientifici, tecnici, inerenti alla Sentiment Analysis (periodo 2012-2015), sono stati molto incentrati sull'idea che il sentimento degli utenti fosse una funzione di un singolo tweet. Sulla questione di ridurre il sentimento a una funzione possiamo essere più o meno d'accordo, ricordando che l'approccio probabilistico prevede proprio l'utilizzo di funzioni per descrivere frequenze relative, per trovare stati e eventi ed essere in grado di dare un'interpretazione ai dati del linguaggio naturale. Ma capiamo che ci possono essere molti svantaggi a ridurci all'analisi di un solo tweet: questo perché in realtà il tweet, che può essere stato filtrato ovviamente da algoritmi ad hoc, è un elemento di un contesto più ampio, uno scenario di cui fa parte (ad esempio di un topic comune); perché gettare via l'informazione del contesto che comunque siamo sempre in grado di reperire, pensando che non ci serva a niente? Con questo modello capiamo l'importanza di riuscire ad associare modelli non a singoli tweet, ma ad interi contesti di cui i tweet fanno parte: quello che viene cercato di fare è di delegare il problema del riconoscimento e dell'interpretazione della polarità attraverso una classificazione che sta sopra un flusso di tweets. In quale modo riusciamo a fare questo, e soprattutto perché capiamo che utilizzare un approccio probabilistico su un singolo tweet sia poco preciso? (Obiettivo fisso: accuratezza e precisione). Capiamo con un semplice esempio che l'interpretazione della polarità di un tweet può essere molto ambigua, se non viene messa in relazione al contesto, allo scenario da dove viene fuori. Valutiamo il tweet "Sono d'accordo con te riguardo le sostituzioni", inerentemente al topic inerente una partita di calcio specifica (linkata con il simbolo dell'hashtag #): con un'analisi sommaria, potremo quasi associare una polarità positiva a questo commento; ma, ricostruendo lo scenario e il contesto(qualche tweets precedente), capiamo che magari questo tweet è di risposta a un tweet negativo (in disaccordo con le sostituzioni): ecco

l'impossibilità di delegare un approccio statistico a un singolo tweet. Per riuscire a svincolarci da questo problema, cerchiamo di porci un duplice obiettivo: per prima cosa, arricchiamo la rappresentazione contestuale di un tweet recuperando anche l'informazione dello scenario, argomento, di cui fa parte, e come seconda cosa introdurre una nuova classificazione più complessa che lavora su un'intera sequenza di tweet(e non un singolo dato). Per fare questa classificazione complessa abbiamo un solo metodo: ridefinire il ruolo di un vettore di interpretazione, che inizialmente era in grado di collegare una singola entità, tweet, ad un topic, ma che in questo contesto rivaluta le sue caratteristiche; il vettore è unico ma sviluppa la rappresentazione delle entità in modo autonomo. Senza andare tanto in profondità a questo argomento, diciamo semplicemente che sfruttando la ridefinizione del SVM e l'allargamento del modello a investire tutto il contesto risulta la potenza moderna di questa tecnica, che trova nell'approccio probabilistico uno strumento magari non precisissimo, ma di cui gli studi stanno arrivando proprio a una precisione elevata, uno strumento sicuramente affidabile e di cui la traduzione è poco onerosa a livello computazionale, tecnologico. Per concludere, in questo secondo scenario, ridefiniamo il vettore SVM che contiene diverse rappresentazioni di entità: in questo modo, ricostruiamo dal singolo tweet un contesto intero, attraverso ovviamente strumenti e formule basso livello, che vengono implementati da tecniche già analizzate nel capitolo secondo; a livello di precisione non siamo al massimo, ma siamo in grado abbastanza velocemente di gestire tutti i dati provenienti dalla realtà.

Conclusioni

Durante il percorso trattato all'interno di questo documento, il mio obiettivo è stato quello di mostrare i collegamenti che uniscono due mondi (intelligenza artificiale e economia data-driven), che in apparenza crescono e maturano singolarmente, ma che a un certo punto vengono uniti per l'ascesa di nuove, fondamentali tecniche analitiche, come la Sentiment Analysis. Nell'ultimo capitolo capiamo l'importanza di definire modelli per questa tecnica: siamo in una realtà nuova, dominata da dati che costituiscono la realtà stessa, e interpretare questi dati significa interpretare l'intero contesto sociale moderno: per questo motivo è fondamentale dare una concretizzazione e sfruttare i vantaggi di una tecnica come la Sentiment Analysis, che si prefigge l'ambizioso obiettivo di andare a ridefinire l'intero mondo del customer-care. Abbiamo valutato che approccio semantico e approccio statistico, che sono le basi di traduzione del linguaggio naturale in generale, applicati in questo contesto risentono degli stessi effetti che subiscono in un contesto non data-driven, e quindi possiamo valutare attraverso i punti forti e i punti deboli la riuscita o meno di questi approcci. In generale, essendo la Sentiment Analysis ancora in fase di cantiere come tecnica, non esiste l'approccio o il modello migliore di altri: semplicemente, possiamo fare delle valutazioni e accettare alcuni limiti a differenza di altri, utilizzando un contesto semantico o probabilistico, limiti che sapevamo già esistere a livello teorico di traduzione del linguaggio. Noi non cambiamo nulla della teoria riguardante le tecniche di NLP (secondo capitolo del trattato): semplicemente prendiamo tutto il blocco NLP, definito prima come mediatore in un piano orizzontale all'interno del primo capitolo, e definito poi verticalmente nel secondo capitolo attraverso tutti i suoi piani di astrazione, e lo inseriamo in questo contesto data-driven, dopo aver definito nel terzo capitolo cosa implica una concezione dell'economia e del contesto sociale dominata dai dati e dal software. Il quarto capitolo è il cuore di questa analisi, e vediamo che mettendo il blocco NLP in un contesto nuovo, moderno, dinamico, andiamo a descrivere nuove tecniche, e soprattutto legandoci alle nuove infrastrutture informatiche, come i Social Network, andiamo a ridefinire il mercato intero, legandoci alle opinioni degli utenti e gestendo una nuova tecnica che ridefinisce questo ambito. Dobbiamo trovare ancora il modello migliore per descrivere questa nuova tecnica in Twitter: ma, in generale, abbiamo capito che (soprattutto nell'approccio probabilistico), conviene legarci a un intero contesto, a un intero scenario, e non al singolo tweet, per

andare a descrivere quelle caratteristiche semantiche che facilitano l'ascesa della precisione e dell'accuratezza di un analisi di opinioni. Infine, questo trattato fornisce un ampio respiro verso il futuro e verso una ridefinizione del ruolo dell'Ingegnere Informatico all'interno dell'azienda, con la nascita di fondamentali nuovi mestieri: il Software non è più argomento di nicchia e delegato in uno spazio virtuale, esso esce dai computer ed entra nel mondo reale (attraverso devices), non andiamo più quindi a mettere il reale nel virtuale (come nei primi siti web anni '90), ma facciamo l'esatto contrario, andando a popolare il mondo reale di software. Una volta gli esperti del settore siti-web erano i Web Analytics, ma ora che il software entra dentro la realtà e viviamo nuove tecnologie, avremo App Analytics per esempio, esperti di social business e social intelligence (basti pensare al mining delle immagini caricate sui Social Media usato per fini pubblicitari), predictive analytics (in grado di monitorare il comportamento di un cliente e anticipare le sue mosse), e infine data-scientist(esperti dell'interpretazione dei dati che compongono la realtà). Tutte queste professioni stanno maturando grazie all'inserimento del Software all'interno della realtà: ridefiniamo il ruolo dell'Ingegnere Informatico all'interno dell'azienda proprio perché, se l'azienda vuole primeggiare sul mercato, e il mercato è popolato da dati e codice, allora significa che l'azienda deve essere in grado di interpretare questo codice, che non è più fine a sè stesso, ma diventa il territorio per fare più in generale Business Analysis. Grazie all'abilità di utilizzare più linguaggi di programmazione, unita all'abilità di gestire database, e rimanendo nel nostro caso grazie all'importanza della conoscenza di tecnologie IA, l'Ingegnere Informatico è l'unica figura aziendale in grado di saper interpretare questi dati, e quindi la realtà. Essere in grado di cambiare velocemente strategie e adattarsi a nuovi scenari tecnologici sarà l'arma vincente per ogni azienda sul mercato moderno. Per essere padroni sul mercato, occorre essere padroni sulla realtà; per essere padroni sulla realtà, occorre essere padroni sui dati; per essere padroni sui dati, occorre essere padroni delle tecnologie informatiche.

Ringraziamenti

Vorrei spendere due parole per ringraziare tutte le persone che mi hanno sostenuto in questo percorso molto impegnativo ma estremamente gratificante e che non mi hanno mai fatto mancare il proprio appoggio.
In particolare, dedico questa tesi a tutta la mia famiglia e parenti, a mia mamma Cinzia, mio babbo Franco e mio fratello Andrea, che amo con tutto me stesso e da cui ho ricevuto un costante sostegno in questo difficoltoso ma soddisfacente percorso universitario, con un fortissimo sostegno morale ed economico per tutta questa fase di studio.
Ringrazio di cuore la mia fidanzata Giulia, che arricchisce le mie giornate, che mi è stata vicino sempre e soprattutto in questo ultimo periodo di studio-lavoro, non facendomi mai mancare il suo enorme affetto.
Ringrazio tutti i miei amici, i più intimi e meno, sia i compagni di corso universitario che mi hanno fatto compagnia in questi anni, facendomi maturare come ragazzo prima e come studente poi, e ovviamente tutti i miei amici di Pesaro, che sono il sale della mia vita.
Ringrazio infine il mio relatore Andrea Roli e il mio correlatore Federico Chesani, che mi hanno aiutato a sviluppare bene questo trattato e hanno sempre mostrato enorme disponibilità, e che insieme ad altri docenti di Cesena mi hanno saputo far appassionare ad argomenti entusiasmanti e veramente interessanti, offrendomi anche la possibilità di poter cercare autonomamente la mia strada per il futuro, verso un ricco e nuovo scenario lavorativo.
Con la speranza di proseguire il mio percorso di studi con la scuola Magistrale, mando un abbraccio a tutti coloro che mi vogliono bene.

GRAZIE!

Giovanni Ciandrini

Bibliografia

[1] JAVA API ON BABELNET: http://babelnet.org.

[2] Pang B. and Lee L. Opinion mining and sentiment analysis. 2008.

[3] G. Berruto. *Corso Elementare di Linguistica Generale*. Utet Università, 2006.

[4] Accoto C. Reality mining: dai big data alla social intelligence, le nuove professioni della data-driven economy. 07/05/2015. Università di Ingegneria e Scienze Informatiche, Cesena.

[5] Fiore E. Aspetti e problematiche del parser del linguaggio naturale, 2000. http://www.di.unipi.it/cappelli/seminari/fiore1.pdf.

[6] Hassan and He. Semantic sentiment analysis of twitter. 11-15/11/2012. The 11th International Semantic Web Conference (ISWC 2012), Boston, USA.

[7] Zelle J. and Tou Ng H. Corpus-based approaches to semantic interpretation in natural language processing. *AI Magazine Volume 18 Number 4 (AAAI)*, 1997.

[8] Karen Sparck Jones. Natural language processing: A historical review. 2001.

[9] Schutze H. Manning D.C. *Foundations of Statistical Natural Language Processing*. The MIT press, 1999.

[10] Navigli R. Natural language processing: Introduction. 27/04/2015. Università di Bologna.

[11] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2009.

[12] Croce D. Vanzo A. and Basili R. A context-based model for sentiment analysis in twitter. 2014. Dipartimento di Enterprise Engineering, Università di Tor Vergata, Roma.