Analisi e Annotazione Automatica dei Testi Letterari

R. Basili, A. Moschitti, and M. Pennacchiotti

(†) Dipartimento di Informatica, Sistemi e Produzione
{basili,moschitti,pennacchiotti}@info.uniroma2.it

Sommario In questo lavoro viene introdotto un approccio alla analisi dei testi letterari secondo una prospettiva computazionale ispirata ai modelli di apprendimento automatico propri dell'Intelligenza Artificiale. Si propone quindi un modello per la rappresentazione ed analisi dei fenomeni narrativi trattabile al calcolatore ed un processo intelligente per la annotazione narrativa automatica. Il sistema risultante può essere addestrato alla simulazione dei comportamenti di un critico che renda disponibili alcuni esempi in forma di annotazioni di tipo narrativo di frammenti testuali. Esso si configura quindi come un *agente letterario* che, adattando il proprio comportamento alle decisioni del critico, ne automatizza la analisi soggettiva. L'approccio proposto si basa sulla ipotesi che le proprietá testuali osservabili sono fortemente correlate con i concetti narrativi da apprendere. Nel presente lavoro è stato studiato l'impatto che diverse tali proprietá hanno sulle capacitá di generalizzazione dell'agente. Le misure della accuratezza raggiungibile sono state effettuate su una specifica opera letteraria, *Gli Indifferenti* di Alberto Moravia. Nella pur preliminare fase presente, i risultati ottenuti confermano, raggiungendo livelli elevati di accuratezza ($> 80\%$), le ipotesi del lavoro aprendo cosí nuove prospettive alle tecnologie di supporto all'indagine dei testi letterari.

1 Introduzione e Motivazioni

È stato spesso osservato ([1]) che l'analisi dei testi letterari assistita al calcolatore è ancora basata sullo studio delle *concordanze* cosi come esso è stato tradizionalmente inteso sin dal tredicesimo secolo. Cio' che viene usualmente trascurato in tale prospettiva è l'enorme insieme di informazioni che i processi di elaborazione testuali oggi offrono in modo più o meno trasparente per l'utente finale: strutture di memorizzazione dei dati, indici, rappresentazioni intermedie, elementi ipertestuali indotti o esplicitamente forniti dalle edizioni digitali. Come ricordato in [1], tali evidenze costituiscono un altro oggetto,

un nuovo *mostro letterario* (come il centauro *Cheiron*), macro unitá dotata di nuove significazioni artistiche ed ermenutiche proprie. È infatti ovvio che la rappresentazione digitale, i suoi metadati e i suoi derivati (indici, alberi di rappresentazione sintagmatica, riferimenti ipertestuali interni o riferiti a sorgenti esterne) costituisce un nuovo insieme di evidenze, ad amplificazione e completamento delle tradizionali *concordanze*. Ciononostante, tali evidenze non appaiono valorizzate a sufficienza negli studi e nelle tecnologie oggi disponibili a sostegno della analisi letteraria. Allo studioso ne è reso di fatto impossibile l'uso e la conseguente tesaurizzazione nell'insieme delle possibili interazioni critiche con l'opera originaria .

Un vero approccio moderno non puo' sottrarsi alla sfida di integrare il *mostro* e le sue parti in un processo interattivo tra lo studioso ed il sistema di calcolo, mirato ad amplificarne le interazioni con l'opera letteraria e ad accompagnare attivamente l'evolvere dei suoi percorsi critici. È chiaro che tutto ciò richiede tecnologie e architetture informatiche rinnovate rispetto alla tradizionale visione offerta dai cataloghi e dalle biblioteche digitali.

La vera sfida a lungo termine di questa ricerca è la definizione di paradigmi e ambienti per la analisi dei testi narrativi al calcolatore che esibiscano alcune proprietá fondamentali:

- *orientati al sistema di calcolo*, cioé basati su astrazioni esplicite (cioé rappresentazioni) di concetti e definizioni narrative, poste al servizio di meccanismi automatici di inferenza propri dei processi di analisi critica (ad es. analogia)
- *interattivi* con lo studioso, tramite paradigmi di eleborazione intelligenti e proattivi, in grado di intraprendere iniziative autonome sull'opera letteraria in cooperazione con il critico (ad es. la scoperta, proposta e validazione di nuove correlazioni funzionali e formali tra porzioni non ancora studiate del testo)
- *multifunzionali* e *integrati*, in grado cioé di sostenere un processo incrementale di raffinamento delle conoscenze interne del sistema in parallelo con la evoluzione della analisi dello studioso.

Tali paradigmi assicurerebbero la progettazione di nuove piattaforme ed ambienti di studio ad alto livello di astrazione ed in grado di amplificare la capacitá di analisi/sintesi letteraria di singoli o comunitá di studiosi.

Non v'è dubbio infatti che l'influenza delle nuove tecnologie in altri settori artistici (quali ad esempio, la composizione musicale o le arti visuali) è di gran lunga maggiore, e, di fatto, portatrice di significativi stimoli allo studio ed ai processi creativi in tali ambiti, ove confrontata con il limitato impatto nel campo letterario. La musica contemporanea, e non solo quella elettronica, vede nel calcolatore un imprescindibile strumento di analisi e sostegno al processo creativo ([2]). Da un lato, lo sviluppo di linguaggi di rappresentazione del dato sonoro sono da tempo oggetto di studio ed essi vantano consolidate tradizioni tecnologiche. E' altresí vero che le tecnologie di analisi e comprensione linguistica hanno prodotto risultati significativi in modelli computazionali di straordinaria ricchezza e flessibilitá ([3,4,5,6,7,8]). Piuttosto l'utilizzo di questi ultimi, concentrato su problemi applicativi più semplici (come il miglioramento delle tecnologie di ricerca automatica dell'informazione come i motori di ricerca nel Web) ha in generale trascurato l'ambito del linguaggio nella sua incarnazione più alta: il testo letterario. È chiaro che la complessitá del testo letterario è maggiore di diversi ordini di grandezza. La relativa maturitá dei processi automatici di analisi testuale consente pero' oggi di invertire la tendenza e progredire lungo la strada della ricerca di modelli più sofisticati per i fenomeni semantici di incrementale ricchezza. Modelli espliciti per la semantic lessicale infatti oggi rendono più vicina la rappresentazione e la gestione di forme più astratte di conoscenze riguardo al testo ed ai suoi diversi livelli interpretativi.

Le descrizione dei fenomeni narrativi vista come progressivo arricchimento ed astrazione dei fenomeni di semantica testuale è l'obbiettivo della linea di ricerca in cui questo lavoro si inserisce. La capacitá dei sistemi di elaborazione dell'informazione di costruire e gestire rappresentazioni di crescente complessitá attraverso livelli di progressiva astrazione ha caratterizzato l'intera storia dell'informatica costituendono probabilmente il fondamento disciplinare e, al contempo, la ragione del suo successo (cioé la supremazia rispetto ad altre tecnologie). Ai processi indipendenti di gestione del sistema hardware si è nel tempo sostituito quella di *macchina astratta* (hardware e software al contempo) che ha stimolato enormi progressi nell'area dei sistemi operativi, dei linguaggi di programmazione e nella gestione dei dati. In ambito narrativo, una analoga prospettiva sug-

gerisce che le capacitá di organizzazione, annotazione ed analisi testuale possano un giorno confluire nella analoga nozione di *macchine astratte* dedicate alla elaborazione ed alla comunicazione narrativa (*macchine narrative*) caratterizzate da astrazioni concettuali delle interpretazioni e dei modelli propri del critico.

In questo lavoro abbiamo inteso condurre una esplorazione di questa prospettiva concentrandoci sugli scopi e sui modi con cui tali astrazioni possano essere realizzate e sulle implicazioni possibili riguardo al loro uso (ad esempio il controllo delle interazioni con l'utente, critico letterario o studioso): quali fenomeni osservabili nel *mostro* letterario possono trovare un'utile collocazione nel progresso che lo studioso compie rispetto all'opera originaria? Quali iniziative possono essere prefigurate in analogia, concorrenza e sostegno al lavoro dello studioso? E quali percorsi frutto di questa interazione sono resi possibili a posteriori al fruitore finale dell'opera?

In questa analisi preliminare abbiamo inteso studiare, secondo una prospettiva induttiva ed orientata all'apprendimento automatico, una rappresentazione esplicita di fenomeni narrativi nei termini di un modello esplicito delle conoscenze narrative riguardo ad un opera. Su tale modello è stato progettato un processo automatico di analisi semantica (un *agente letterario*), che esposto alle valutazioni assunte dal critico, osservate nelle annotazioni di un testo elettronico, generalizza tali scelte in un processo di apprendimento automatico per proporne di nuove in porzioni dell'opera non ancora analizzate dallo studioso. L'esplorazione, in questa fase della ricerca, si è concentrata sulla valutazione della natura e delle tipologie di evidenze testuali che consentono all'agente di acquisire un modello esplicativo adeguato del comportamento del critico, a partire dagli esempi forniti. La nozione di adeguatezza qui utilizzata è funzionale e non descrittiva: essa coincide con la misura di accuratezza con cui l'agente è in grado di riconoscere gli fenomeni narrativi in nuove porzioni (non analizzate) del testo in accordo con lo studiose. Questa analisi è stata condotta sull'opera di Alberto Moravia, *Gli Indifferenti* ([9]) che secondo le parole dell'autore è *Un romanzo con pochi personaggi, con pochissimi luoghi, con un'azione svolta in poco tempo. Un romanzo in cui non ci fossero che il dialogo e gli sfondi e nel quale tutti i commenti, le analisi e gli interventi dell'autore fossero accuratamente aboliti in una perfetta oggettivitá* ([10]).

La nozione di adeguatezza adottata fornisce evidenze empiriche oggettive e consente una verifica sperimentale rigorosa mirata alla comparazione sistematica di modelli diversi in termini delle osservazioni usate, cioé le proprietá testuali rese disponibil e dei metodi induttivi applicati. Gli indicatori quantitativi di valutazione delle prestazioni, largamente utilizzati negli studi di Intelligenza Artificiale e Trattamento Automatico della Lingua, sono stati qui assunti come parametri per la analisi comparativa tra i modelli diversi. A riguardo, sono state indagate diverse classi di informazione testuale (da quelle ortografiche a quelle morfologiche, sintattiche e semantiche) ed è stato verificato il loro impatto sulla accuratezza dei processi induttivi dell'agente risultante. La definizione del modello di rappresentazione, la progettazione dei processi di addestramento e una prima analisi dei risultati saranno discusse rispettivamente nelle sezioni 2, 3) e 4.

2 Una ontologia di elementi narrativi

Lo studio critico di un'opera letteraria generalmente parte dalla identificazione e dalla analisi di specifici fenomeni narrrativi, quali l'uso di schemi o di particolari forme descrittive. Uno strumento automatico in grado di cooperare con il critico nella analisi di *corpora* di larga scala dovrebbe identificare e incarnare un modello di tali aspetti: cio' che è necessario quindi è una rappresentazione dei fenomeni narrativi come un loro modello di conoscenza esplicito. Esso fornisce, da un lato, un vocabolario concettuale per la descrizione digitale dei fenomeni narrativi di interesse nell'opera e, d'altro canto, un modello di rappresentazione del mondo narrativo in grado di sostenere i processi autonomi di riconoscimento automatico, di apprendimento e di pianificazione. Per esempio, un tale modello potrebbe da un lato rappresentare ed organizzare gerarchicamente l'uso di un linguaggio figurativo e dall'altro fornire una base sistematica per il riconoscimento di tali figure in frammenti di testo nuovi e non precedentemente incontrati e studiati dal critico.

Il processo di definizione ambisce ad esplicitare i fenomeni narrativi attraverso il lavoro cooperativo tra il critico e l'ingegnere della conoscenza. Da una parte i critici individuano gli elementi narrativi

adeguati a catturare fenomeni e pratiche letterarie di interesse. Dall'altra gli ingegneri determinano le rappresentazioni e la loro organizzazione opportuna nel dominio digitale di un sistema di calcolo. Il risultato è una *ontologia di riferimento* per il sistema computazionale dedicato al riconoscimento e all'apprendimento autonomo nei testi letterari. La *ontologia* esprime ed esaurisce l'insieme dei concetti e delle relazioni del mondo oggetto della sua analisi e costituisce di per sé l'unico mondo al quale il sistema fará riferimento. Poiché la natura e la caratterizzazione testuale di una tale ontologia è spesso oggetto di un'analisi del critico in una prospettiva soggettiva, l'ontologia riprodurrá tale soggettivitá. La sua applicabilitá, almeno nelle fasi inziali, sará confinata ai testi alla base dell'ispirazione del critico: nel tempo però, cioé a valle dei processi di apprendimento abilitati dalla iniziale ontologia (*naive*) progettata, il sistema potrá evolvere modificandone concetti e relazioni di base, in un processo di retroazione ricorsiva orientato ad un confonto più adeguato ai testi ed ai fenomeni diverso da quelli inizialmente previsti.

L'informazione di base dell'analisi critica è il testo e quindi una rappresentazione dei fenomeni di interesse è fondata su una qualche nozione di unitá testuale, la cui grana dipende dalla natura e dagli scopi dell'analisi. La descrizione di un modello narrativo passa necessariamente attraverso l'arricchimento di tali unitá testuali tramite la dichiarazione esplicita (*annotazione*) delle proprietá di interesse. Esse caratterizzano e legano il modello decrittivo (concetti ed attributi ad essi associati) al dato testuale d'origine, incarnando quindi una visione strettamente empirica, cioé ereditata dal dato fondamentale dell'opera originaria. Tale modello, al contempo, ne determina gli aspetti artistici ed universali (ontologici) attraverso il sistema di attributi e proprietá istanziate in frasi o elementi lessicali di volta in volta designati.

È rilevante che l'insieme delle proprietá cosí rilevate formino un sistema di enti e relazioni tra loro che in una prospettiva del tutto induttiva istanzia coerentemente il modello astratto, ontologia, dei fenomeni narrativi di interesse: induttivamente cioé, le realizzazioni testuali di oggetti e proprietá astratte vanno a costituire il mondo di riferimento (ontologia appunto) del critico e, come conseguenza, determinano gli assiomi e le presupposizioni alla base dell'agire del risultante sistema automatico, l'agente letterario.

La descrizione arricchita di un testo è ottenuta tradizionalmente mediante annotazioni esplicite che estendono parole, frasi o paragrafi di un testo mediante un linguaggio strutturato detto *linguaggio di marcatura* (*mark-up*). Questo processo prende quindi complessivamente il nome di *annotazione* e un insieme di *annotazioni puntuali* ne costituiscono il risultato. Ogni annotazione puntuale esprime il tipo narrativo principale ed i suoi attributi e caratterizza cosí l'unitá testuale associata. Il frammento di testo annotato, quindi, combina un concetto narrativo (cioé il suo tipo ed una sua descrizione formalizzata in attributi predeterminati) e, al contempo, una sua esemplificazione (cioé la frase, il passaggio annotato). Nella prassi corrente ([XML]) il linguaggio d'elezione per le annotazioni di testi narrativi è *XML* (*eXtended Mark-up Language*), che costituisce, per la sua flessibiltá ed il suo grado di standardizzazione, un protocollo ad alta espressivitá e di larghissima condivisibilitá tra comunitá di utilizzatori diverse. Una organizzazione internazionale (Text Encoding Initiative, TEI) ha giá progettato e formalizzato l'insieme dei tipi descrittivi (etichette di concetti e attributi) che consentono la annotazione nei testi di informazione linguistica (e.g. morfologica o sintattica), filologica, storica ed editoriale. La relativa arbitrarietá che all'interno degli standard promossi da TEI è resa possibile, ha consentito di estendere nel nostro studio il livello semantico e narrativo che non è ad oggi oggetto di una standardizzazione stretta da parte della TEI. Il sistema progettato nel corso di questa ricerca quindi rappresenta uno dei rari tentativi nella definizione di alcune proprietá semantiche e narrative ed una estensione *ad hoc* delle linee guida di TEI che meglio riflette gli scopi della nostra ricerca.

La nozione di *annotazione narrativa specifica* utilizzata nel nostro lavoro si basa sulle seguenti ipotesi:

- essa copre un *frammento di testo contiguo* che costituisce l'ambito (o *span*) della annotazione; due annotazioni diverse possono essere l'una contenuta completamente nell'altra (*inclusione* tra ambiti), ma non determinano mai sovrapposizioni parziali del loro ambito
- una *annotazione* caratterizza una *porzione di testo* mediante un *tipo narrativo principale*, un potenziale *sottotipo* e *un insieme di attributi*, cioé espressioni atomiche (valori) che descrivono aspetti,

stati o variazioni del tipo principale
- Il *vocabolario* che caratterizza il sistema di tipi, sottotipi e attributi delle annotazioni è determinato inizialmente e soggiace a precise leggi logiche tali che
  - un *sottotipo* caratterizza tutti e soli gli elementi testuali che sono anche caratterizzati dal suo *tipo superiore*
  - un *attributo* determina una *caratteristica propria di un tipo* e assume valori in un *dominio* specifico
  - un *tipo* puo' essere descritto da un preciso sottoinsieme di attributi che, in taluni casi possono essere lasciati non specificati (*attributi opzionali*)

La definizione dell'apparato di tipi narrativi e dei loro attributi rende possibile l'attivitá di annotazione, cioé la creazione delle annotazione specifiche che caratterizzano un testo sorgente. In questa ricerca il testo sorgente considerato è l'opera de Gli Indifferenti di Alberto Moravia [9]).

2.1 Studio di un caso: *Gli Indifferenti* di Alberto Moravia

Del capolavoro de *Gli Indifferenti*, scritto da Moravia nel 1929, v'è poco da ricordare in uno studio sperimentale e orientato alla tecnologia, come questo. È casomai da sottolineare il suo stile lucido e razionale realizzato in un linguaggio semplice ed estremamente diretto. Per questo esso rappresenta un caso di studio interessante rappresentativo di molti temi che ricorreranno nel Moravia maturo: le generalizzazioni possibili infatti potranno essere strumento di studio per un corpus letterario più ampio, che includa i suoi lavori successivi.

Il romanzo ha una struttura teatrale: i dialogi e le azioni si svolgono in ambienti fortemente caratterizzati, stanze o spazi cittadini, con una attenzione estrema per le loro descrizioni oggettive che si soffermano su dettagli minuziosi, espressivi. La capacitá di isolare e riconoscere azioni, descrizioni, luoghi e personaggi assume quindi un ruolo importante poiché la loro funzione è centrale per la indagine sugli obbiettivi ed i percorsi narrativi della pur esile trama.

Per questo la *ontologia degli elementi narrativi* qui definita per Gli indifferenti si costituisce attorno a quattro classi principali di *descrizioni*. Ogni descrizione è denotata da una annotazione (etichetta

NAR (oggetto)) i cui tipi principali (attributo OGG) sono *Luogi esterni (le), Luoghi Interni (li), Personaggio Maschile (pm), Personaggio Femminile (pf)*. Ogni descrizione ha due proprietá distinte. La prima è la *tipologia* (attributo TIPO), cioé l'uso possibile di un linguaggio figurativo che caratterizza la descrizione. Se tale linguaggio viene usato in un framento allora la annotazione specifica del frammento avrá un valore *simbolico* (*s*) per l'attributo (TIPO=s). Al contrario la descrizione si dirá oggettiva (TIPO=o). La seconda proprietá è la *espressivitá* (EXP) che caratterizza la comlessitá della prosa e puo' essere *sintetica* (EXP=s) o *analitica* (EXP=a).

Una ulteriore dimensione di una descrizione è legata alle azioni descritte. Anzitutto una azione è descritta mediante un *punto di vista*: spesso l'autore presenta gli eventi di una descrizione attraverso lo sguardo di un personaggio. Infine la descrizione può riguardare *azioni* compiute dai personaggi o semplicemente stati di cose. Queste due informazioni caratterizzano la nozione di sottostruttura narrativa introdotta dalla etichetta opzionale SUB-NAR: una descrizione puo' quindi essere caratterizzata da una sottostruttura SUB-NAR (il cui ambito è completamente contenuto in quello della descrizione), i cui attributi sono la presenza di una azione (attributo AZIONE=s) ed il punto di vista (attributo PDV). I valori validi per l'attributo PDV sono tutti i personaggi dell'opera, laddove la assenza di tale informazione suggerisce l'autore come sguardo oggettivo, spettatore imparziale.

In sintesi, il modello di annotazione identifica una descrizione narrativa come una macro struttura testuale di tipo NAR, di cui il tipo descrittivo viene dichiarato mediante l'attributo OGG, e gli attributi TIPO e EXP definiscono rispettivamente la tipologia e l'espressivitá. Quando necessario, la descrizione viene decomposta in una sua sottostruttura mediante il tipo SUB-NAR che introduce le azioni descritte ed i punti di vista. Un esempio di frammento annotato che definisce due descrizioni narrative è mostrato in Fig. 1. La prima introduce in forma oggettiva e sintetica un personaggio maschile (Leo). La seconda descrive analogamente invece un personaggio femminile (Carla), attraverso lo sguardo di Leo.

Sicuro, rispose Leo accendendo una sigaretta; forse non mi vuoi? <NAR OGG=pm TIPO=o EXP=s>Curvo, seduto sul divano, egli osservava la fanciulla con una attenzione avida;</NAR> <NAR OGG=pf TIPO=o EXP=s><SUB_NAR PDV=Leo>gambe dai polpacci storti, ventre piatto, una piccola valle di ombra fra i grossi seni, braccia e spalle fragili, e quella testa rotonda cosí pesante sul collo sottile.</SUB_NAR></NAR>

Figura 1. Un esempiodi annotazione narrativa nel formalismo TEI da *Gli Indifferenti* (Capitolo 1).

La versione originale italiana del romanzo di Moravia è composto di sedici capitoli e di circa 91,000 parole. Un gruppo di critici[1] ha annotato interamente l'opera che nella sua versione finale è quindi stato digitalizzato interamente nel suo formato secondo lo standard TEI, ed è quindi accessibile e navigabile secondo le ususali applicazioni Web.

3 Apprendimento automatico e Riconoscimento di fenomeni narrativi

Nella letteratura di Intelligenza Artificiale il problema dell'apprendimento automatico è cosí definito: *Apprendere* per un algoritmo, o un sistema software, significa *migliorare la propria prestazione attraverso l'esperienza di un certo compito o processo* [11]. Tale miglioramento è caratterizzato come segue:

- apprendere significa *ottimizzare* una certa funzionalitá, ad esempio, la capacitá di riconoscere fenomeni o proprietá nei testi,
- apprendere corrisponde al miglioramento di uno o più *indici di prestazione*, ad esempio l'accuratezza del riconoscimento cioé il numero di decisioni generate automaticamente, coerenti con la corrispondente scelta del critico
- apprendere richiede una *esperienza*, ad esempio la interazione con un utente o con un maestro che fornisce esempi e controesempi

Gli approcci induttivi sottolineano l'ultimo aspetto, restringendo la nozione di esperienza alla esposizione ad esempi della funzione da apprendere, cioé un insieme rappresentativo di decisioni

[1] Una discussione più estesa del modello narrativo e della sua rappresentazione ontologica è discusso altrove in questo volume.

corrette. Negli algoritmi induttivi *basati su esempi*, il sistema automatico sviluppa quindi la funzione di decisione, detta *ipotesi*, a partire dai dati ricevuti ad esempio (addestramento). Il processo di induzione fornisce quindi una ipotesi che surroga il concetto da apprendere mediante l'interpretazione dei dati d'addestramento (ad esempio, la interpolazione o la generalizzazione). L'ipotesi ha lo scopo di massimizzare la qualitá (*correttezza* o *adeguatezza*) delle predizioni future possibili su dati non controllati (dati di prova o *testing*).

Questo paradigma è largamente applicato in scenari molto diversi tra loro che vanno dalla classificazione automatica di testi al riconoscimento di forme (per esempio, volti di persone) in immagini bidimensionali. È chiaro che un approccio induttivo richiede alcune fondamentali assunzioni:

- La inferenza del sistema (riconoscimento) puo' sempre essere ricondotta ad un passo di classificazione, cioé di selezione della classe opportuna all'interno di un numero finito di opzioni
- La esperienza del sistema si basa su esempi che possono essere descritti formalmente. In genere, una descrizione formale utilizza un insieme sistematico di *proprietà descrittive* che dovrebbero essere correlate con il concetto da apprendere. Nel caso della classificazione automatica di testi (per esempio classi di notizie d'agenzia) le proprietà irrilevanti possono essere la agenzia di stampa di provenienza o la data che, al contrario di altre informazioni (quali le parole utilizzate nel testo delle notizie) sono scarsamente correlate con le classi[2]. Nel caso del riconoscimento di volti proprietà irrilevanti possono essere la luminosità globale dell'immagine o il numero di *pixel* utilizzati.

Il romanzo annotato come descritto nella sezione 2 è stato usato per addestrare un'algoritmo di apprendimento automatico a riconoscere ed annotare i frammenti di pagine specifiche del romanzo. L'algoritmo induttivo basato su esempi è stato progettato ed addestrato attraverso i frammenti di testo (annotazioni specifiche) creati dagli esperti nell'opera. Associando l'informazione testuale ovvero

[2] In una stessa data ad esempio, una agenzia di stampa fornisce notizie di categorie diverse dalla cronaca, alla politica, allo sport: conoscenre la agenzia di provenienza non ci dice nulla sulla categoria opportuna di una notizia.

l'insieme delle proprietà del testo osservate nei frammenti già annotati, l'algoritmo apprende un loro modello (testuale) per annotare automaticamente, a sua volta, porzioni del testo non osservate in precedenza. I frammenti quindi costituiscono gli esempi di addestramento e la funzione ipotesi caratterizza le diverse nozioni narrative descritte nella ontologia.

Lo sviluppo della funzione ipotesi per un concetto è formulata sulla base della analogia tra le diverse annotazioni specifiche di quel concetto disponibili nell'addestramento. Tali analogie si stabiliscono tra porzioni di testo, l'unica evidenza disponibile al sistema nei testi che saranno analizzati in futuro. La qualità di tale osservazione sarà quindi un indice della accuratezza raggiunta dal sistema e, d'altro canto, la evidenza empirica di larga scala che convalida (o smentisce) il modello narrativo appreso.

É da sottolineare che nel nostro approccio viene resa possibile una interazione tra l'aggiustamento incrementale del modello narrativo, ad opera del critico, e la sua applicazione automatica a porzioni crescenti di testo: ció costituisce l'obbiettivo finale di questa ricerca, cioé la definizione di un paradigma interattivo per la analisi critica dei testi letterari. La scala di analisi raggiungibile in questo scenario, mediante la cooperazione e la proattività dell'agente letterario (rese possibili dall'apprendimento automatico), costituisce una novità rispetto agli studi correnti sulla digitalizzazione delle opere letterarie. Tali studi, spesso incentrati sulla definizione di forme di annotazione obbiettive (ad es. l'anno di pubblicazione o altre notizie di tipo storico-critico), sono essenzialmente orientati ad amplificare i processi di fruizione (ricerca e consultazione) ma non forniscono strumenti diretti di indagine letteraria, rimandando alla attività tradizionale di studio del critico il processo di analisi vero e proprio. Negli scenari previsti in questo lavoro, la scoperta di alcune dimensioni narrative interessanti è in parte demandata al livello di parziale interpretazione dell'agente letterario agente su una più large mole di osservazioni collezionabili. Questo amplifica non solo le possibilità di accesso del critico ma anche parte dei suoi processi analitici.

3.1 Analisi Narrativa e Categorizzazione Automatica

Poiché la nozione minimale di frammento narrativo è esemplificata attraverso una annotazione specifica il cui ambito puo' coincidere con un intero paragrafo, il problema induttivo posto nell'addestramento (cioé annotare i frammenti narrativi mai osservati in precedenza) puo' esser decomposto nei due seguenti sotto problemi.

- il *riconoscimento* di paragrafi di interesse per il critico, cioé potenzialmente caratterizzati da tipi narrativi (cioé descrizioni narrative proprie dell'ontologia definita)
- la *classificazione* dei paragrafi potenzialmente rilevanti nei tipi narrativi individuali. Nel caso della analisi de Gli Indifferenti, quindi la classificazione dei paragrafi in tipi quali i luoghi esterni, i luoghi interni, le persone maschili o le persone femminili

I due passi definiti sopra vengono quindi eseguiti in cascata. Il riconoscimento automatico dei paragrafi rilevanti puo' essere effettuato attraverso un classificatore binario che da qui in poi verrà chiamato IPC (*Interesting Paragraph Classifier*). Esso produce due opzioni di rilevanza: vero nel caso di accettazione di un paragrafo come di un frammento di interesse narrativo, falso al contrario. Un IPC quindi seleziona solo i frammenti che possono potenzialmente contenere delle descrizioni narrative di interesse. La fase di successiva di classificazione avviene attraverso il classificatore dei tipi narrativi (*Paragraph Type Classifier*, PTC). Il PTC consente la annotazione specifica di nuovi paragrafi mai annotati prima ed è applicata all'insieme dei frammenti rilevanti reso disponibile dal classificatore IPC della fase precedente. Un PTC consiste in una comunità di diversi classificatori binari di sezione, dedicati a riconoscere ciascun tipo narrativo descritto dalla ontologia. Tali classificatori possono essere addestrati attraverso un approccio *uno-a molti* ([12]): un confronto viene effettuato tra le decisioni dei diversi classificatori di tipo ed una sola decisione viene accettata (*voting*). Nel caso in cui più di un tale classificatore accetti il paragrafo, cioé più di una etichetta di tipo narrativo venga proposta per lo stesso paragrafo, quella caratterizzata dalla confidenza più alta del relativo classificatore viene accettata (*majority voting*). Mediante questi due passi, cioé n classificazioni e quindi *majority voting*, il PTC risultante si comporta

come un classificatore multi-classe che decide riguardo ai diversi tipi narrativi descritti nell'ontologia.

L'addestramento di un classificatore è un problema ben noto nella letteratura di Intelligenza Artificiale e sono stati proposti numerosi paradigmi e algoritmi, dai sistemi basati su regole ([13]), ai modelli probabilstici, alle reti neurali ([14]). Tra gli altri le Support Vector Machine (SVM) ([15]) hanno mostrato livelli di prestazione, accuratezza e robustezza, molto soddisfacenti rispetto ai problemi di scala (limitate collezioni di dati di addestramento) e consistenza (errori o incongruenze nelle annotazioni) delle osservazioni. A tale paradigma di apprendimento automatico sarà dedicata la successiva sezione.

3.2 Spazi Geometrici ed Analisi narrativa

Il paradigma delle SVM sfrutta una metafora geometrica del significato nota come *Vector Space Model* ([16]). I concetti che costituiscono l'obbiettivo dell'apprendimento sono infatti rappresentati attraverso vettori di proprietà n-dimensionali, cioe' come punti di uno spazio ad n dimensioni. Essi esprimono in forma vettoriale gli insiemi di proprietà osservabili nei paragrafi annotati. Le proprietà forniscono le dimensioni indipendenti dello spazio e il loro peso (cioé la componente reale del vettore relativa a tale proprietà) ne descrive la relativa importanza nel paragrafo rappresentato. Nel caso più semplice i pesi sono binari e denotano il fatto che tale proprietà narrativa o testuale è presente (1) o no (0). Inoltre i vettori costituiscono esempi o controesempi di un concetto narrativo: il vettore $\boldsymbol{x}$ di un paragrafo caratterizzato da una descrizione di luogo interno sarà un esempio (positivo) di tale concetto; esso sarà un controesempio (o esempio negativo) in caso contrario.

Una SVM apprende l'iperpiano che separa gli esempi positivi da quelli negativi caratterizzato dal margine massimo. Più formalmente, applicando il principio di minimizzazione del rischio strutturale [15], la SVM durante l'addestramento calcola la funzione lineare che definisce l'iperpiano di separazione

$$H(\boldsymbol{x}) = \boldsymbol{w} \times \boldsymbol{x} + b = 0, \tag{1}$$

dove $\boldsymbol{x} \in \Re^n$ è la variabile dell'iperpiano, $\boldsymbol{w} \in \Re^n$ è il suo gradiente (vettore dei coefficienti dell'iperpiano) e $b \in \Re$ una costante. La

funzione $H(\boldsymbol{x})$ è quindi la funzione obbiettivo utilizzata per le predizioni future. Un nuovo oggetto $\boldsymbol{x'}$, mai prima osservato, è un istanza del concetto appreso *se e solo se* $H(\boldsymbol{x'}) > 0$, se cioé si dispone nel sottospazio positivo determinato dall'iperpiano: le nuove istanze vengono quindi respinte (come controesempi) se giacciono sull'iperpiano ($H(\boldsymbol{x'}) = 0$) o nel sottospazio negativo ($H(\boldsymbol{x'}) < 0$).

La fase più critica dell'addestramento di una SVM è la definizione delle proprietà adeguate a codificare i fenomeni necessari per la costruzione della funzione di classificazione $H(\boldsymbol{x})$. La scelta di tali proprietà caratterizza il modello di apprendimento definito. Nella sezione successiva verranno presentate le diverse proprietà linguistiche adottate per caratterizzare le descrizioni narrative.

3.3 Rappresentazione dei fenomeni narrativi

Il riconoscimento e la classificazione dei frammenti narrativi richiede informazioni di tipo differente, ossia propietà individuali, osservabili nei testi oggettivamente ed in modo relativamente efficente. Poiché la complessità delle descrizioni narrative puo' variare da testo a testo, e da critico a critico, essa richiede la fusione di proprietà linguistiche e testuali di tipo diverso, che riguardano i livelli ortografico, morfologico, sintattico e semantico. Di conseguenza, nel nostro studio sono state progettate diverse rappresentazioni per tali proprietà.

Questa distinzione è stata mantenuta esplicita durante l'addestramento ed il test per studiare il contributo dato da livelli linguistici individuali all'accuratezza del classificatore risultante.

Proprietà ortografiche (`orth`)

Queste rappresentano l'insieme delle parole usate nel romanzo (livello ortografico). Ogni parola presente in un paragrafo (in una delle classi chiuse, cioé aggettivi, avverbi, nomi e verbi) è considerata come una proprietà distinta dalle altre (ovevero una dimensione indipendente). Abbiamo trovato 19,274 tali parole uniche nel romanzo. Il valore di ciascuna proprietà ortografica è la frequenza della parola nel paragrafo considerato.

Proprietà morfologiche (`morph`)

Le proprietà morfologiche includono due sottoinsiemi diversi di proprietà.

– **Lemmi**. Ogni lemma del romanzo è una proprietà distinta. Ad esempio *ando'*, *andato* sono rappresentate dallo stesso lemma, l'infinito presente *andare*. Nel romanzo sono stati determinati 8,713 differenti lemmi. Il valore assegnato in un paragrafo a ciascuna tale proprietà è la frequenza dei lemmi individuali del paragrafo.
– **Categorie Sintagmatiche**. Queste proprietà catturano la proprietà morfosintattica di ciascuna parola individuale data dalla sua categoria sintagmatica nei frammenti di esempio. Ogni categoria principale (ad es. nomi e aggettivi maschili e femminili, forme verbali) è quindi rappresentata da una proprietà distinta. Il valore per tali proprietà e' la percentuale di ciascuna di esse in un esempio (ad es. la percentuale di nomi maschili e femminili in un paragrafo). Per i verbi, vengono distinte le percentuali delle diverse forme (ad esempio, gerundive, condizionali congiuntive e finite) e tempi (presente, passato remoto, ...) verbali.
– **Nomi propri**. Questa proprietà individuale rappresenta la percentuale di nomi propri di persona (o luogo) in un paragrafo.

Proprietà sintattiche, `synt`

Le relazioni sintattiche semplici binarie che si stabiliscono nel testo sono rappresentate dalle proprietà sintattiche. Le relazioni tra coppie di categorie sintagmatiche che sono state considerate sono:

– modificazione aggettivo nome (maschile e femminile)
– legame argomentale nome (maschile e femminile) e verbo

Il valore di queste proprietà è fornito dalla percentuale di bigrammi di un certo tipo (ad es. nome maschile e verbo) che si manifestano in ciascun frammento rispetto al totale di tali coppie nell'intera opera. L'estrazione delle proprietà morfologiche e sintattiche dall'opera di Moravia è stata effettuata attraverso CHAOS ([17]), un parser sintattico alle dipendenze modulare sviluppato presso il laboratorio di Intelligenza Artificiale dell'Università di Roma Tor Vergata[3].

Proprietà semantiche, `sem`

Queste proprietà cercano di catturare alcune caratteristiche semantiche del lessico verbale e nominale. Per la definizione delle proprietà semantiche di tali classi parole e' stato utilizzato Multiwordnet

[3] I metodi e le tecniche alla base del sistema CHAOS, sono discusse in un altro lavoro su questa raccolta.

[18], un dizionario semantico organizzato in una tassonomia di sensi nominali: i nomi quindi vengono generalizzati lungo la tassonomia fino a quando alcune classi semantiche di riferimento vengono selezionate. Le classi semantiche di riferimento dipendono dal romanzo poché si assume che esse forniscano informazioni rilevanti agli specifici tipi narrativi individuati. Le classi semantiche utilizzate per i nomi negli esperimenti di seguito descritti sono: *mobilio*, *vestiario*, *parti del corpo*, *luoghi esterni* e *interni*. Nomi che non corrispondono a tali classi vengono semplicemente trascurati dalla analisi di questo livello.

Le classi semantiche dei verbi sono derivate dalle associazioni lessicografiche di Wordnet [19]. Queste includono 11 classi verbali, come ad esempio i verbi di *creazione*, *emozione* o *comunicazione*. Ciascuna classe rappresenta una proprietà individuale. Usando Multiwordnet ogni verbo in un paragrafo è generalizzato in una classe, corrispondente ad uno dei suoi sensi.

Il peso di una generica proprietà semantica è calcolata come il rapporto del numero di nomi (o verbi) in tale classe ed il numero totale dei lemmi costituenti un paragrafo.

4 Risultati sperimentali

In questi esperimenti, abbiamo misurato la accuratezza del riconoscimento di paragrafi di interesse (IPC) e della loro classificazione nei diversi tipi descrittivi (PTC). Inoltre abbiamo anche valutato l'impatto delle diverse proprietà testuali discusse precedentemente rispetto alla accuratezza della classificazione.

4.1 Configurazione sperimentale

Il corpus di riferimento per gli esperienti e' una versione elettronica degli Indifferenti di 91, 000 parole, le cui sezioni narrative fanno riferimento alle descrizioni di luoghi e persone annotate precedentemente. In particolare, 395 paragrafi su 2326 totali (17%) sono stati annotati dai critici, nei termini delle descrizioni e delle loro proprietà (Sezione 2). Da tali 395 paragrafi, 51 sono esempi di luoghi esterni, 113 di luoghi interni, 156 di descrizioni di personaggi femminili e 75 di descrizioni di personaggi di sesso maschile.

Per gli esperimenti di classificazione (Sezione 3.2) è stata usata la piattaforma SVM light [20], addestrata secondo un kernel lineare standard come metrica nello spazio delle proprietà.

Il corpus di esempi annotati e' stato quindi suddiviso in un 30% per la valutazione mantenendo il rimanente 70% per l'addestramento. Quindi il classificatore binario dei paragrafi rilevanti IPC è stato valutato su un insieme di 698 paragrafi diversi mentre il classificatore dei tipi descrittivi PTC è stato sperimentato su 118 paragrafi di esempio.

La accuratezza del classificatore binario IPC, è stata valutata attraverso le misure tradizionali di *precisione* e *copertura* (*recall*). La prima misura la percentuale di risposte esatte fornite dal sistema. La seconda misura la percentuale di esempi (in una certa classe) correttamente classificati dal classificatore. Verranno poi riportati i dati nella misura sintetica nota com $F-measure$ (F_1), cioe' la media armonica tra *precisione* e *recall*. La valutazione del classificatore di tipi descrittivi PTC è stata anche effettuata mediante il fattore di *accuratezza* (che vale *1 - tasso percentuale di errore*). Questo valore indica il numero di decisioni corrette fornite dal classificatore rispetto a tutti i paragrafi di test.

4.2 Risultati della classificazione

La Tabella 1 mostra la *F-measure* del classificatore binario dei paragrafi rilevanti (IPC) in relazione alle diverse proprietà testuali utilizzate. Le colonne 2,3,4,5,6, fanno riferimento ai risultati utilizzando proprietà ortografiche, i soli lemmi, tutte le proprietà morfologiche, sintattiche e semantiche rispettivamente. Le colonne 7, 8, 9, 10 descrivono la qualità della classificazione nei casi di combinazioni delle proprietà individuali.

	orth	*lemmi*	*morph*	*synt*	*sem*	*orth-morph*	*orth-synt*	*morph-synt*	*all*
F_1	84.6	85.7	87.2	18.6	62.2	88.6	83.6	86.1	88.6

Tabella 1. $F-measure$ nel riconoscimento dei paragrafi di interesse (IPC) rispetto alle diverse proprietà linguistiche.

Riguardo alla classificazione dei tipi descrittivi (PTC), la Tabella 2 mostra la accuratezza ottenuta, rispetto ai diversi livelli di infor-

mazioni testuali utilizzate. Le colonne dalla 2 alla 10 forniscono le informazioni rispetto all'uso di proprietà individuali oppure rispetto ad alcune combinazioni tra di esse

	orth	*lemmi*	*morph*	*synt*	*sem*	*ortho-morph*	*ortho-synt*	*morpho-synt*	*all*
Accuratezza	73.7	72.8	73.7	22.2	45.6	78.1	73.7	75.4	79.0

Tabella 2. Accuratezza della classificazione dei tipi descrittivi (*PTC*) rispetto alle diverse proprietà linguistiche.

La valutazione conclusiva della tabella 3 riporta la F-measure dei classificatori dei tipi descrittivi individuali, cioé dei 4 classificatori di *luoghi interni*, *esterni* e di *personaggi femminili* e *maschili*.

Tipo descrittivo	Precisione	Recall	F_1
External Place	61.54	57.14	59.26
Internal Place	77.78	63.64	70.00
Female Person	67.24	86.67	75.73
Male Person	57.14	72.73	64.00
Accuratezza		79.0	

Tabella 3. F_1 dei classificatori individuali delle descrizioni narrative basati su tutte le proprietà linguistiche.

4.3 Discussione dei risultati

I nostri esperimenti preliminari sul riconoscimento dei paragrafi contenenti tipi descrittivi di interesse narrativo forniscono molto materiale utile ad una discussione approfondita. Sintetizziamo di seguito alcune osservazioni prevalenti.

Innanzitutto, il riconoscimento dei paragrafi di interesse sembra esser appreso in modo efficace mediante proprietà testuali molto semplici. Usando infatti solo le proprietà ortografiche, cioé la semplice rappresentazione per parole, senza la applicazione di alcun trattamento morfologico, il sistema raggiunge già una F-measure dell'84.6%. Questa accuratezza è piuttosto alta, simile a quella ottenuta

rispetto a problemi più semplici, quali la classificazione automatica dei documenti in categorie tematiche.

In secondo luogo, il potere di rappresentazione dei lemmi con una F-measure dell'85.7% è più alto di quello delle parole individuali (cioé proprietà puramente ortografiche). Questo conferma come, in una lingua la cui morfologia è ricca, la lemmatizzazione è molto importante. Inoltre, quando sono usate in combinazione le proprietà morfologiche, parole e lemmi, la F-measure cresce ancora di circa 1.5%. Osserviamo che le proprietà ortografiche contengono le proprietà di lemmatizzazione e morfologiche: dato l'insieme esiguo di esempi, è però conveniente mantenere distinte tali informazioni. I lemmi hanno infatti una probabilità media più alta in generale rispetto alle parole individuali. Il sistema quindi accresce la sua *recall* laddove le proprietà morfologiche mantengono alta la precisione complessiva del sistema. Le proprietà sintattiche usate da sole o in congiunzione con altre proprietà non sembrano fornire informazione sufficiente e determinano una penalizzazione della accuratezza. La principale ragione è che la accuratezza raggiunta grazie alle proprietà di base è già abbastanza alta. In questo caso il rapporto tra l'errore medio del parser CHAOS, richiesto per estrarre le proprietà sintattiche, e quello della classificazione è penalizzante. Il parser infatti produce strutture sintagmatiche con una accuratezza del 75-80% circa e quindi il suo contributo non aiuta (nemmeno in combinazione) a migliorare la qualità della classificazione rispetto ad altre proprietà (per es. i lemmi ottengono da soli una accuratezza (*F-measure*) pari a 85.7%).

Infine, l'uso combinato delle proprietà ortografiche e di quelle morfologiche fornisce la accuratezza più alta, 88.6% F- measure, pari al risultato ottenuto con l'insieme di tutte le proprietà. In sintesi, il sistema è in grado di recuperare (e sottoporre alla eventuale analisi del critico) un insieme di paragrafi che nell'88.6 % dei casi corrisponde correttamente ad un qualche tipo descrittivo definito nell'ontologia ed esemplificato durante l'addestramento: tale decisione e' giustificata su una base puramente lessicale (proprietà morfologiche e ortografiche).

I risultati del classificatore di tipi descrittivi PTC in Tabella 2 sono leggermente differenti. Innanzitutto, i lemmi producono una accuratezza più bassa rispetto alle semplici parole (72.8% rispetto al 73.7%). Questo è spiegato dalla maggiore importanza delle parole e

della loro individuale morfologia relativamente a questo tipo di problema. Ad esempio, nelle descrizioni di persone femminili e maschili, è importante rappresentare l'intera parola, dotata delle sue variazioni di genere e numero. I lemmi possono non fornire un'informazione sufficiente poiché normalizzano, perdendola, l'informazione di genere nel lemma di riferimento (singolare e maschile per nomi e aggettivi). In seconda istanza, analogamente al caso del categorizzatore binario, nei paragrafi di interesse la combinazione delle feature ortografiche e morfologiche fornisce una accuratezza molto alta (78.1%) ma, in contrasto con i precedenti risultati, la informazione sintattica sembra fornire una informazione utile rispetto a quella di tipo morfologico (75.4% rispetto al 73.7%). Le ragioni di tale comportamento sono duplici:

- la prestazione del 73.7% è più di 10 punti percentuali al di sotto di quella fornita dal riconoscitore dei paragrafi (categorizzatore IPC). Questo valore è quindi confrontabile con l'accuratezza media del parser CHAOS: quindi probabilmente gli errori dovuti al parser non impattano in modo determinante sulla accuratezza globale
- inoltre il parser sintattico è applicato al sottoinsieme dei paragrafi riconosciuti al passo precedente. I suoi errori quindi sono ben generalizzati dall'algoritmo di addestramento delle support vector machine, che impara utilmente dai dati sintattici utili, e non fa dipendere il suo comportamento da potenziali errori. Questo minimizza l'impatto di questi ultimi sulla accuratezza della classificazione.

Va notato, infine, come la prestazione migliore viene ottenuta quando tutte le proprietà testuali sono usate congiuntamente. Questo suggerisce che le informazioni sintattiche e semantiche hanno un impatto significativo sulla classificazione dei tipi descrittivi, che è un problema di natura concettuale e più complesso, molto più sensibile ai livelli più astratti della informazione linguistica.

In Tabella 3, e' riportatata la prestazione dei classificatori individuali dei quattro tipi narrativi. Osserviamo che questa misura è strettamente proporzionale alla taglia dei dati di addestramento disponibili per le diverse classi. Come atteso, le categorie sono piuttosto simmetriche ed i problemi di classificazione relativi mostrano

complessità simili. Inoltre, le F-measure dei classificatori individuali sono molto più basse di quelle del classificatore globale *PTC* risultante che raggiunge il 79%. Questo non sorprende: poiché la associazione, attraverso il *majority voting*, di un *pool* di classificatori binari in un classificatore *multiclass* complessivo, migliora la accuratezza risultante rispetto a quella dei votanti individuali. Infatti, il *majority voting* è complessivamente più robusto rispetto agli errori locali ai classificatori individuali che vengono compensati dalla loro associazione.

5 Conclusioni

In questo lavoro abbiamo presentato e dimostrato la applicabilità di un nuovo approccio all'analisi dell'opera letteraria basato sui risultati e su tecnologie proprie dell'Intelligenza Artificiale. I risultati sperimentali suggeriscono che il riconoscimento di fenomeni narrativi interessanti nel testo può essere automatizzato con un elevato grado di accuratezza. L'approccio è già utilizzabile in un nuovo scenario tecnologico per lo studio letterario, mirato ad armonizzare l'analisi linguistica del testo e le direzioni critiche dell'esperto. Esso apre quindi prospettive più interessanti per il progetto di ambienti software *letterari* autonomi, adattivi e proattivi, a sostegno del lavoro del critico. Le future attività mirate ad estendere le osservazioni empiriche consentite in questo lavoro si muoveranno su due diverse linee di indagine.

Quantitativamente, la sperimentazione verrà estesa ad opere e ad algoritmi di apprendimento automatico diversi: questo consentirà di consolidare e generalizzare le osservazioni empiriche oggi rese possibili ne Gli Indifferenti. Sarà confrontato il potere espressivo degli spazi geometrici del significato testuale (*Support Vector Machines*) con modelli differenti, ad esempio algoritmi squisitamente statistici ([5,6]) che esprimono ipotesi diverse sugli spazi di riferimento.

Da un punto di vista piu' legato alla analisi narrativa, verrà approfondito lo studio delle implicazioni epistemologiche della fase di apprendimento. Lo studio degli iperpiani di separazione tra esempi e controesempi infatti può suggerire interpretazioni lessico semantiche appropriate delle regole di comportamento dell'agente: un modello testuale complesso (parole, sensi, e relazioni tra di essi) che soggiace

al concetto narrativo appreso. Le correlazioni emergenti automaticamente dagli esempi forniscono non solo uno strumento di classificazione utile alle predizioni nei frammenti testuali nuovi, ma soprattutto livelli interpretativi appropriati non previsti dal critico nella fase di annotazione. Incarnano quindi dimensioni epistemologiche nuove sul testo non postulate a priori. Una analisi accurata della sistematicità di tali evidenze costituirà la più ambiziosa continuazione della linea di ricerca discussa in questo lavoro.

Ringraziamenti

Gli autori riconoscono come inestimabili le discussioni con il prof. Andrea Gareffi che ha ispirato gran parte dei principi e delle scelte progettuali della ricerca qui presentata.

Riferimenti bibliografici

1. Rockwell, G.: What is text analysis, really? Literary and Linguistic Computing **18, n. 2** (2003) 201–219
2. Xenakis, I.: Formalized Music: Thought and Mathematics in Composition. (Pendragon)
3. Chomsky, N.: Lectures on government and binding. Dordrecht: Foris (1981)
4. Chomsky, N.: Language and Problems of Knowledge. Cambridge, MA: MIT Press (1988)
5. Manning, C., Schuetze, H.: Foundations of statistical natural language processing. MIT Press (1999)
6. Jelinek, F.: Statistical methods for speech recognition. MIT Press (1997)
7. Jurafsky, D., Martin, J.: Speech and language processing. Prentice Hall (2000)
8. Jackendoff, R.: Semantic Structures. Current Studies in Linguistics Series. MIT Press (1990)
9. Moravia, A.: Gli Indifferenti. Bompiani (1929)
10. Moravia, A.: Ricordo de Gli Indifferenti. Bompiani (1945)
11. Mitchell, T.: Machine Learning. McGraw Hill (1997)
12. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research **5** (2004) 101–114
13. Muggleton, S.: Scientific knowledge discovery using inductive logic programming. Communications of the ACM **542(11)** (1999)
14. McCulloch, W., Pitts, W.: A logical calculus immanent in nervous activity. Bulletin of Mathematical Biophysics **5** (1943)
15. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1995)
16. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley (1989)
17. Basili, R., Zanzotto, F.M.: Parsing engineering and empirical robustness. Natural Language Engineering **8/2-3** (2002)

18. Bentivogli, L., Pianta, E., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: First International Conference on Global WordNet, Mysore, India (2002)
19. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38** (1995) 39–41
20. Joachims, T.: Making large-scale SVM learning practical. In Schölkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods - Support Vector Learning. (1999)