# Osservatorio (Rubrica del Bollettino della AI*IA, Marzo 1998)

## Incontro dei gruppi di lavoro dell'AI*IA su Apprendimento Automatico e Linguaggio naturale

*Torino, 9-10 dicembre 1997*


*Liliana Ardissono*
*Dipartimento di Informatica*
*Università di Torino*

*Marco Botta*
*Dipartimento di Informatica*
*Università di Torino*

*Anna Goy*
*Dipartimento di Informatica*
*Università di Torino*

*Donato Malerba*
*Dipartimento di Informatica*
*Università di Bari*


I gruppi di lavoro su "Apprendimento Automatico" e su "Elaborazione del Linguaggio Naturale" dell'AI*IA hanno organizzato in modo congiunto le loro giornate di lavoro che si sono tenute presso il Dipartimento di Informatica dell'Università di Torino nei giorni 9 e 10 dicembre 1997.

Scopo delle giornate di lavoro è stato quello di promuovere una disseminazione di informazione e una maggiore collaborazione tra i ricercatori delle due aree. Per questa ragione si è pensato di articolare l'incontro in tre mezze giornate di cui una, quella centrale, è stata dedicata al tema congiunto "Linguaggio Naturale e Apprendimento Automatico", mentre nelle altre si sono svolte sessioni parallele. Alle giornate hanno partecipato più di 60 persone. Gli incontri sono stati organizzati dai responsabili dei due gruppi, Floriana Esposito (Università di Bari) e Leonardo Lesmo (Università di Torino).

APPRENDIMENTO AUTOMATICO. Negli incontri del gruppo di lavoro su "Apprendimento Automatico" sono state presentate quattordici relazioni, di cui sei il pomeriggio del 9 dicembre e otto il pomeriggio del 10 dicembre. Purtroppo altre due relazioni previste per il 9 dicembre ~~sono saltate~~ non sono state presentate per l'impossibilità dei relatori a partecipare all'incontro.

Il primo insieme di sei contributi è stato piuttosto eterogeneo, coprendo tanto le strategie puramente induttive quanto quelle basate su rinforzo ed evolutive. In particolare, Donato Malerba (Università di Bari) ha presentato ATRE, un sistema di apprendimento automatico in grado di indurre teorie logiche mutuamente ricorsive da "oggetti" di addestramento descritti mediante clausole a testa multipla. Ha fatto seguito l'intervento di Nicola Fanizzi (Università di Bari) su una formalizzazione del problema della revisione di teorie logiche mediante il linguaggio della logica lineare del primo ordine, del quale si fornisce una semantica delle fasi. Pierguido Caironi (Politecnico di Milano) ha illustrato due nuovi algoritmi di apprendimento per rinforzo, entrambi in grado di costruire le politiche di controllo per agenti autonomi fisici con variabili di stato e di controllo di tipo continuo. La relazione di Attilio Giordana (Università di Torino) ha riguardato una nuova strategia co-evolutiva per l'apprendimento automatico di formule logiche del primo ordine. Tale strategia, basata sulla teoria delle nicchie e della formazione delle specie, è supportata da un nuovo insieme di operatori genetici che mirano esplicitamente a preservare la diversità nella popolazione. Francesco Amigoni (Politecnico di Milano) ha parlato di uno schema generale in cui inquadrare l'apprendimento in una agenzia dinamica, definita, questa, come un insieme di agenti fissi e mobili che cooperano. I tipi di apprendimento identificati per agenzie dinamiche sono quello esplorativo e quello cooperativo. Infine Marco Dorigo (Université Libre de Bruxelles) ha interessato i presenti con una applicazione di agenti mobili al problema di costruire dinamicamente tabelle di routing per reti di comunicazione.

Più omogeneo il secondo gruppo di contributi, quasi tutti riguardanti l'apprendimento induttivo. In particolare, il sistema presentato da Marco Davite (Elsag Bailey) incorpora un algoritmo di clustering per partizionare in classi quei campi ritenuti rilevanti in alcuni esempi di immagini di documenti. Il partizionamento è ottenuto minimizzando, con algoritmi di simulated annealing, l'informazione sulla posizione spaziale dei campi. Giovanni Semeraro (Università di Bari) ha illustrato l'applicazione del sistema di induzione di alberi di decisione C4.5 al problema di classificare gli utenti di una biblioteca elettronica. La classificazione è effettuata in base ad alcuni attributi estratti automaticamente dal file di log associato ad ogni utente registrato. Nel successivo intervento, Valentina Tamma (Università di Bari, Università di Liverpool) ha discusso potenzialità e limiti dei metodi di apprendimento da esempi di particolari modelli predittivi strutturati ad albero, detti alberi di regressione e alberi di modelli. Tali modelli sono estensioni dei più noti alberi di decisione, pensati per la predizione di attributi di tipo continuo piuttosto che discreto. Sempre sull'apprendimento da esempi, ma questa volta di programmi logici multi-teoria, l'intervento di Fabrizio Riguzzi (Università di Bologna). L'algoritmo illustrato lavora in due passi: apprendimento su istanze (basato su una estensione della definizione di "least general generalization" data da Plotkin) e apprendimento su classi, secondo una direzione bottom-up (dal basso verso la radice della gerarchia di classi). Infine uno strumento per la visualizzazione/manipolazione di dati e generazione di ipotesi è stato presentato da Sergio Caggese (Università di Aberdeen). DaViCCAND, questo il nome del sistema in questione, fu inizialmente progettato come data miner nel campo della metallurgia, ed è stato recentemente generalizzato in modo da poter essere applicato ad altri domini.

Diversi sono stati i contributi che hanno legato l'apprendimento automatico alle scienze cognitive. Incentrato sulla psicologia cognitiva l'intervento di Alessandra Valpiani (Università di Torino), che ha relazionato su un sistema ispirato ad una teoria cognitiva dei processi di ragionamento sul mondo fisico. In questo sistema l'apprendimento consiste in un raffinamento di teorie mediante inferenze che hanno lo scopo di costruire una spiegazione di situazioni complesse. Hanno fatto seguito due interventi sulla costruzione di modelli cognitivi mediante metodi di apprendimento automatico. In particolare, Lorenza Saitta (Università di Torino) ha illustrato gli obiettivi generali del progetto europeo "Learning in Humans and Machines", finanziato dalla European Science Foundation, e quelli specifici dell'unità torinese. Questi riguardano lo sviluppo di un modello computazionale di cambio concettuale, cioè di cambio dei modelli causali usati per interpretare i dati, per spiegare dei fenomeni e per effettuare delle predizioni. Il sistema usato per costruire il modello computazionale è WHY, che apprende e revisiona regole di classificazione usando conoscenza del dominio (causale e fenomenologica) e esempi di addestramento. La presentazione degli obiettivi specifici dell'unità di Bari, che partecipa allo stesso progetto, è stata fatta da Stefano Ferilli. In questo caso si mira a verificare se è possibile usare sistemi di apprendimento automatico per ritrovare gli stessi modelli mentali proposti dagli psicologi cognitivi, riguardanti concetti di fisica naive rilevati in bambini di età compresa fra i sei e gli undici anni. In particolare si è focalizzata l'attenzione sul concetto di forza. L'applicazione del sistema incrementale INTHELEX ha mostrato che gli operatori logici di generalizzazione e specializzazione in esso implementati sono effettivamente in grado di simulare almeno alcuni dei meccanismi fondamentali dell'apprendimento umano.

A conclusione degli interventi della prima giornata si è anche tenuta un'assemblea del gruppo di lavoro di apprendimento automatico. Fra i temi caldi dell'incontro, quello del ruolo attuale del gruppo all'interno dell'AI*IA e le iniziative da intraprendere per il futuro. Per pubblicizzare l'attività del gruppo si è deciso di comporre delle pagine HTML da agganciare a quelle attualmente disponibili per l'intera associazione. Si è anche deciso di arricchire con altri data set utilizzati dalla comunità italiana di apprendimento automatico il repository di dati disponibile all'Università di Cagliari, mirror di quello creato alla University of California, Irvine. Infine si è anche pensato alla organizzazione di una scuola di apprendimento automatico con taglio fortemente applicativo, rivolta soprattutto al personale di grossi centri di ricerca pubblica e industriale, potenziali fruitori di strumenti di apprendimento automatico ritenuti ormai robusti.

APPRENDIMENTO AUTOMATICO E LINGUAGGIO NATURALE. L'incontro congiunto dei due gruppi di lavoro si è tenuto al mattino del 10 dicembre: dieci gli interventi, organizzati in due "timeslot" al temine di ognuno dei quali si è avuto un momento di discussione collettiva.

Le prime relazioni sono state dedicate all'acquisizione di informazione lessicale, considerato uno dei problemi fondamentali nel campo dell'elaborazione del linguaggio naturale. Per risolverlo Simonetta Montemagni (Par.O.La sas, Pisa) ha proposto un approccio all'estrazione di conoscenza lessicale basato sull'analisi sintattica superficiale, intesa come processo di "text chunking", cioè estrazione dal testo libero di "isole di certezza sintattica", dette "chunks". Il risultato del text chunker è analizzato da ABACUS, un sistema che acquisisce automaticamente (cioè apprende) i pattern di sottocategorie di una parola. Walter Castelnovo (Università di Milano) ha guardato al problema dell'acquisizione automatica di nuove componenti lessicali da una prospettiva diversa. Assumendo di disporre di un lessico indipendente dal dominio applicativo, di una teoria generale della grammatica e di un corrispondente algoritmo di parsing, l'idea è quella di acquisire, mediante una strategia di apprendimento basata su spiegazione, la terminologia rilevante per una specifica applicazione. Il contributo presentato da Michele Vindigni (Università di Roma, Tor Vergata) ha considerato specificatamente il problema dell'acquisizione automatica di schemi di sottocategorizzazione verbali. Partendo da vettori di potenziali argomenti di un verbo ottenuti mediante un'analisi sintattica superficiale, si usa un algoritmo di clustering concettuale per organizzare in un reticolo di classi tali vettori. Ogni nodo rappresenta una struttura alternativa di sottocategorizzazione verbale. La scelta delle strutture significative in un preciso dominio applicativo è effettuato mediante tecniche statistiche, realizzando così un vero e proprio processo di tuning lessicale.

Altre due relazioni hanno avuto la disambiguazione come denominatore comune. Nello specifico, il secondo intervento di Simonetta Montemagni è stato dedicato alla disambiguazione del senso delle parole: la similarità semantica di due parole occorrenti in due contesti non è basata su una tassonomia di concetti come quella di WordNet, ma sulla condivisione di tipici contesti d'uso. È questo l'approccio implementato nel sistema SENSE, che usa tecniche statistiche nell'analisi distribuzionale dei contesti d'uso delle parole. Sulla disambiguazione sintattica, questa volta, il secondo contributo, presentato da M. Di Nanni, del gruppo di elaborazione del linguaggio naturale dell'Università di Tor Vergata. L'algoritmo proposto disambigua sulla base di criteri quantitativi di preferenza applicati a soluzioni ottenute dal soddisfacimento di vincoli relativi all'intera sequenza di costituenti ambigui di una frase (di qui il concetto di "località estesa").

Eterogenee le altre relazioni. Graziella Tonfoni (Università di Bologna e Università del Maryland), considerando la "iconizzazione concettuale" come parte del processo di codifica di documentazione testuale, ha proposto forme di ragionamento qualitativo sulla natura dell'informazione. Giorgio Satta (Università di Padova), ha discusso le problematiche legate all’apprendimento automatico di conoscenza linguistica da un corpus, quando si adotta un approccio basato su trasformazioni, ed ha illustrato un algoritmo in grado di contenere la domanda di risorse computazionali per corpus molto grandi. Il lavoro di Vincenzo Lombardo (Università di Torino) si basa su un modello psicolinguistico secondo il quale gli esseri umani elaborano le frasi in modo incrementale, cioè operando da sinistra verso destra, parola per parola. Obiettivo della ricerca è quello di usare tecniche statistiche per determinare un limite superiore al grado di ambiguità di una parola in input, data l’esistenza di un contesto sinistro. Pietro Leo (IBM Semea Sud) ha presentato WebClass, un sistema per la classificazione automatica di pagine HTML, da integrare in un Web assistant proprietario IBM. Il problema fondamentale affrontato è quello della selezione automatica degli attributi che descrivono una pagina Web. Tale selezione si basa su statistiche come la frequenza di un termine occorrente nella pagina HTML, e porta a una selezione di attributi discriminanti fra le classi. Le strategie adottate per la classificazione si basano sull’apprendimento di alberi di decisione obliqui e sulla distanza dal centroide della classe. Infine Fiorella de Rosis (Università di Bari) ha illustrato un metodo abduttivo probabilistico per consentire ad un agente di risolvere un problema di “diagnosi cognitiva”, cioè di inferire lo stato cognitivo di un altro agente che convive ed opera nel suo mondo.

LINGUAGGIO NATURALE. Le due sessioni del gruppo di "Linguaggio Naturale" comprendevano 12 relazioni, ma purtroppo ne sono state presentate solo 11, per impossibilita` di un relatore di partecipare all'incontro.

I contributi del primo incontro si sono incentrati su problemi di parsing, categorizzazioni di parole ed estrazione di informazione da testi. Il primo intervento e` stato di Rodolfo Delmonte (Universita` di Venezia), che ha descritto un parser volto ad un'interpretazione del linguaggio naturale plausibile dal punto di vista cognitivo. In tale parser, si adottano 4 strategie (il Minimal Attachment, la Preferenza Funzionale, la Valutazione Semantica e l'individuazione dei riferimenti) per incorporare intuizioni sui meccanismi psicologici di comprensione delle frasi; gli esperimenti effettuati con tale parser hanno mostrato che le quattro strategie sono fortemente legate tra loro da dipendenze reciproche. Fabio Pianesi (ITC-IRST, Trento) ha descritto il sistema Geppetto: uno strumento per definire applicazioni in linguaggio naturale che offre in un unico ambiente integrato di sviluppo la possibilita` di introdurre, correggere etc. le risorse linguistiche, e come pure di definire e validare l'architettura dell'applicazione stessa. Nel terzo intervento, Sarah Jacqmin-Labat (CERTAL-INALCO, Universita` di Amiens) ha presentato l'analizzatore morfologico di un sistema di analisi di testi in Italiano (SADOTPI). In tale sistema, non si vuole effettuare un'analisi completa delle parole; l'idea e` costruire un dizionario in cui siano riportate tutte le possibili terminazioni, con la rispettiva categorizzazione, e usare tale informazione per classificare le parole di un testo. In seguito, Giacomo Ferrari (Universita` del Piemonte Orientale) ha descritto un progetto di codifica di informazione testuale per il linguaggio scritto e parlato; in tale progetto si sta acquisendo un corpus di testi , che verrà usato per evidenziare le strutture retoriche usate in diversi classi testuali e per definire un sistema di etichettatura utilizzabile sia per il testo scritto che per quello parlato. Luca Dini (CELI, Torino) ha presentato il sistema di estrazione dell'informazione IUTA (Italian Unrestricted Texts Analyzer). IUTA e` basato su una serie di moduli di analisi linguistica che riempiono template a partire dal documento in input. L'attenzione e` stata focalizzata sul modulo di generazione di "Naif Logical Forms" (rappresentazioni semantiche sottospecificate), prodotte tra la fase di shallow parsing e lo riempimento dei template. Infine, Fabio Ciravegna (ITC-IRST, Trento) ha descritto un approccio deterministico all'analisi di testi nell'ambito dell'estrazione di informazioni; l'analisi sintattica e` effettuata da cascate di trasduttori a stati finiti ed e` divisa in due fasi: nella prima, si identificano le strutture predicato/argomento, mentre nella seconda si usano vincoli sintattici e semantici per stabilire gli attacchi dei modificatori.

Durante la seconda sessione, si sono affrontate tematiche piu` vicine alla rappresentazione concettuale e alla pragmatica. In particolare, Anna Goy (Universita` di Torino) ha presentato un approccio alla rappresentazione semantica lessicale che si ispira alla "Case-Frame Semantics", integrandola con idee derivanti dal Lessico generativo e dalle "Prototype Theories". Tale approccio e` stato illustrato mediante un "case study" su un insieme di aggettivi che fanno riferimento a stati emotivi. Dario Bianchi (Universita` di Parma) ha invece affrontato il problema della rappresentazione concettuale nella comprensione di storie, descrivendo un sistema di rappresentazione alla KL-ONE, con una tassonomia lessicale che descrive i predicati ed i loro argomenti, ed una tassonomia che rappresenta gli stati e gli eventi. Il sistema e` usato per analizzare storie in linguaggio naturale. Maurizio Tirassa (Universita` di Torino) ha presentato un approccio psicologico cognitivo alla comunicazione, che e` modellata come interazione tra agenti razionali, indotta dal loro stato mentale. Il suo approccio evidenzia, come requisito principale alla comunicazione, la capacita` di rappresentarsi le dinamiche cognitive dell'interlocutore e la situazione comunicativa stessa. Nell'ambito delle analisi di corpora, Gabriella Braiato (CSELT, Torino) ha descritto i risultati di un'analisi del fenomeno della ripetizione nel dialogo uomo-macchina, distinguendo ruoli e tipi di ripetizione diversi, effettuati dal parlante sulle sue stesse affermazioni o su quelle del sistema. In seguito, Liliana Ardissono (Universita` di Torino) ha descritto un approccio ad agenti alla modellazione del dialogo in linguaggio naturale. In tale approccio, i turni dei partecipanti sono interpretati come coerenti continuazioni del dialogo se contribuiscono al soddisfacimento di obiettivi conversazionali, relativi all'attivita` di dominio o a quella di problem solving dei parlanti durante l'interazione.

Al termine della prima giornata di lavori si e` svolta l'assemblea del gruppo. Dopo un bilancio dell'attivita` del 1997, la discussione si e` focalizzata su due suggerimenti per l'attivita` del 1998. Il primo di essi e` una raccolta di informazioni sulle risorse disponibili presso i centri di ricerca e sviluppo che collaborano all'attivita` del gruppo di lavoro, al fine di permettere una maggiore visibilita` dei risultati ottenuti e di favorire scambi di dati e/o software tra i membri, con conseguente riduzione della duplicazione di sforzi per problemi gia` risolti. Il secondo suggerimento e` quello di stringere maggiori contatti con le aziende. E` stato rilevato come questo implichi la necessita` di una vera e propria attivita` promozionale, attraverso la quale il gruppo riesca a promuovere all'esterno i servizi che e` in grado di al mondo produttivo.