

# BASILI Roberto

## Curriculum vitae

Last Update: December 2017

Nato: [redacted] Agosto 1951

Residenza: [redacted]

**Posizione Corrente:** Idoneo alla Posizione di Professore Ordinario nel Settore Disciplinare **INF/01 - Informatica** dall'Aprile 2017. Svolge attività didattica presso il Dipartimento di Ingegneria dell'Impresa nei corsi di *Basi di Dati*, *Web Mining e Retrieval*, *Intelligenza Artificiale* e *Gestione dei Dati e della Conoscenza*. La attività di ricerca si svolge sui problemi, le metodologie e le tecnologie dell'Intelligenza Artificiale nelle aree dell'*Apprendimento Automatico*, della *Elaborazione del Linguaggio Naturale*, e dell'*Ingegneria dei Sistemi Intelligenti* per il machine Learning su larga scale, il trattamento automatico del linguaggio naturale e per l'accesso all'informazione distribuita sul Web. Tali temi svolgono un ruolo fondamentale nell'ambito della cosiddetta *Data Science*, poichè riguardano i processi di analisi predittiva operanti sui dati non strutturati, che esprimono la maggior parte dei contenuti odierni del Web.

**Posizione precedente:** Professore Associato presso la Facoltà di Ingegneria della Università degli Studi di Roma, Tor Vergata, dall'Ottobre 2004.

**Google Scholar:** <https://scholar.google.it/citations?user=U1A22fYAAAAJ>

**DBLP:** [http://dblp.uni-trier.de/pers/hd/b/Basili\\_0001:Roberto](http://dblp.uni-trier.de/pers/hd/b/Basili_0001:Roberto)

**Studi Universitari:** Laurea in Matematica presso l'Università di Roma "La Sapienza" conseguita nel 1989, con la votazione di 110/110 *summa cum laude*.

**Studi Post Universitari:**

- MIT Summer School in Language and Information (May, 1991)
- Scuola Europea Estiva di Logica, Linguaggio ed Informazione, ESSLI-91, Saarbrücken, Agosto 1991.

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

12-11-1960

- Dottorato di Ricerca in Ingegneria Elettronica, dell'Informazione e dell'Automazione, conseguito presso l'Università di Roma, Tor Vergata, nel 1993, mediante la discussione della Tesi dal titolo: "*Un modello formale per l'apprendimento di concetti linguistici dalla analisi di corpora estesi di testi*"
- Borsa di Studio, presso la Fondazione Ugo Bordoni, dall'Aprile 1992 al Febbraio 1993
- Borsista C.N.R. nel periodo dal Marzo 1993 al Dicembre 1994

# Contents

<b>1</b>	<b>Attività Didattica</b>	<b>4</b>
<b>2</b>	<b>Attività Accademica</b>	<b>6</b>
2.1	Conseguimento di premi e riconoscimenti per l'attività scientifica	6
2.2	Comitati di Programma ed Organizzazione . . . . .	7
<b>3</b>	<b>Collaborazioni Scientifiche</b>	<b>9</b>
<b>4</b>	<b>Attività di Terza Missione: Progetti di Ricerca applicata a finanziamento pubblico o industriale</b>	<b>9</b>
<b>5</b>	<b>Introduzione alla discussione delle Attività di Ricerca</b>	<b>12</b>
<b>6</b>	<b>Descrizione Sintetica delle Attività di Ricerca: anni 2008-oggi</b>	<b>12</b>
6.1	Supervised and Unsupervised Computational Natural Language Learning . . . . .	12
6.1.1	Risultati . . . . .	13
6.2	Semantic Search and Social Media Analytics . . . . .	14
6.2.1	Risultati . . . . .	15
6.3	Natural Language Learning for Human Robot Interaction . . . .	15
6.3.1	Risultati . . . . .	16
<b>7</b>	<b>Descrizione Sintetica delle Attività di Ricerca: anni 2004-2008</b>	<b>16</b>
7.1	Computational Natural Language Learning . . . . .	16
7.1.1	Risultati . . . . .	18
7.2	Ontology Learning . . . . .	18
7.2.1	Risultati . . . . .	19
7.3	Acquisition and Processing of Multimedia Semantics . . . . .	20
7.3.1	Risultati . . . . .	21
<b>8</b>	<b>Descrizione Sintetica delle Attività di Ricerca: anni 1997-2004</b>	<b>21</b>
8.1	Natural Language Based Text Classification and Filtering . . . .	21
8.2	Adaptive Information Retrieval and Information Extraction . . .	22
8.3	Sistemi distribuiti di <i>Information Retrieval ed Extraction</i> . . . .	23
<b>9</b>	<b>Descrizione Sintetica delle Attività di Ricerca, anni 1992-1997</b>	<b>24</b>
9.1	Rappresentazione della Conoscenza . . . . .	25
9.2	Elaborazione del Linguaggio Naturale. . . . .	26
<b>10</b>	<b>Riferimenti Bibliografici</b>	<b>30</b>

# 1 Attività Didattica

Assistente dal 1992 al 1994 del corso di *Strutture Informative*, responsabile della didattica di elementi di logica, dimostrazione automatica dei teoremi, programmazione logica e basi di dati deduttive.

Docente Supplente del corso di *Fondamenti di Informatica 3*, per il Corso di Laurea Breve in Ingegneria Informatica, dell'Università di Ancona, a.a. 1993-94, sulla programmazione orientata agli oggetti e elementi di progettazione dei sistemi operativi.

Dal 1994 assistente alla didattica del corso di *Fondamenti di Informatica 1*, presso la Facoltà di Ingegneria della Università di Roma, Tor Vergata, sulle tecniche di programmazione, strutture dati, algoritmi di base e orientati alla analisi numerica.

Dal 1995 al 1999 assistente nel corso *Basi di Dati* presso il corso di Laurea in Ingegneria Informatica della Università di Roma, Tor Vergata, responsabile della didattica di elementi di logica, dimostrazione automatica dei teoremi, programmazione logica e basi di dati deduttive.

Sin dal 1991 assistente e corelatore di numerose Tesi di Laurea in Ingegneria Informatica presso l'Università di Roma, Tor Vergata e La Sapienza, e in Scienze dell'Informazione presso l'Università, di Roma, La Sapienza.

Invited Speaker su *Unsupervised Machine Learning Methods for Natural Language*, presso la ESSLI School of Logic, Language and Information, Praha, 1996.

Docente Invitato (*Maitre de Conferences*) per il corso di *Statistical Language Processing*, presso l'Université Paris 7, 1999.

Tutorial invitato su *Introduzione alla Estrazione di Informazione*, presso la Conferenza Nazionale della Associazione Italiana di Intelligenza Artificiale, AI\*IA, Bologna, Settembre 1999, in collaborazione con Fabio Ciravegna (Università di Sheffield, UK).

Assistente alla didattica del Corso di *Intelligenza Artificiale*, presso il corso di Laurea in Ingegneria Informatica della Università di Roma, Tor Vergata (a.a. 1999-2000).

Docente Responsabile per il corso di *Basi di Dati* (nuovo ordinamento), presso il corso di Laurea in Ingegneria Informatica della Università di Roma, Tor Vergata, dall'Ottobre 2000, al Settembre 2004 e dall'Ottobre 2007.

Docente Supplente per del corso di *Fondamenti di Informatica 1* (nuovo ordinamento), presso il corso di Laurea in Ingegneria Informatica della Università di Roma, Tor Vergata, dal Dicembre 2000 al Febbraio 2008.

Docente Responsabile per del corso di *Fondamenti di Informatica 1* (Corsi di Laurea On-Line), presso il corso di Laurea in Ingegneria Informatica della Università di Roma, Tor Vergata, dall'Ottobre 2003 al Febbraio 2004.

Docente Responsabile per il corso di *Basi di Dati*, presso il corso di Laurea On-Line in Ingegneria Informatica della Università di Roma, Tor Vergata, dall'Ottobre 2000.

Docente supplente per il corso di *Informatica Generale*, presso il corso di Laurea in Scienze della Comunicazione della Università di Teramo, dal Novembre 1999 al Febbraio 2002.

Tutorial invitato su *Parsing Robusto: Approcci Statistici e Simbolici per il riconoscimento sintattico su larga scala*, presso la Conferenza Nazionale della Associazione Italiana di Intelligenza Artificiale, AI\*IA, Bari, Settembre 2001, in cooperazione con Fabio Massimo Zanzotto (Università di Roma, Tor Vergata).

Invited Speaker: *Methods, representation and linguistic bias in Adaptive IE*, presso il LREC 2002 Workshop on "Event Modelling for Multilingual Document Linking", Las Palmas, Spain, June 2002.

Invited Speaker: *Learning Ontological Models from texts*, presso l'ECAI 2002 Workshop on "Machine Learning and Natural Language Processing for Ontology Engineering", Lyon, France, July 2002.

Docente Responsabile per il corso di *Modelli per la Gestione e la Ricerca di Informazione*, presso il corso di Laurea Specialistica in Ingegneria Informatica della Università di Roma, Tor Vergata, dall'Ottobre 2005 al Settembre 2008.

Docente Responsabile per il corso di *Trattamento Automatico della Lingua*, presso la Facoltà di Lettere e Filosofia della Università di Roma, Tor Vergata, dall'Ottobre 2004.

Invited Speaker: *Natural Language and Ontology Learning: some perspectives for SW applications*, presso la Third Summer School on "Ontological Engineering and the Semantic Web" (SSSW'05), 10th-16th July 2005 - Cercedilla (Spain).

Docente invitato: *Semantica dei Testi, Apprendimento Automatico e Beni Culturali*, presso la scuola invernale "Intelligenza artificiale nei beni culturali", Dipartimento di Informatica, Sistemistica e Comunicazione, Università Milano-Bicocca, 2007.

Keynote Speaker: *From multimedia semantic indexing to cross-lingual retrieval: the Prestospace approach to cultural heritage preservation and dissemination*, presso il Workshop on Natural Language Processing and Knowledge Representation for eLearning environments, September 26th, 2007, Borovets, Bulgaria, In conjunction with RANLP '2007.

Invited Tutorial: *Support Vector Machines and Kernel Methods for Robust Extraction of Textual Knowledge*, presso SWAP 2007, Fourth Italian Workshop on Semantic Web Application and Perspectives, December , 2007, Bari, Italy. In collaborazione con Alessandro Moschitti.

Docente Responsabile per il corso di *Web Mining e Retrieval*, presso il corso di Laurea Specialistica in Ingegneria Informatica della Università di Roma, Tor Vergata, dall'Ottobre 2008.

Docente Responsabile per il corso di *Social Media Analysis and Recommendation Systems*, presso il corso di Master in Big Data della Università di Roma, Tor Vergata, dal Gennaio 2016.  
URL del Corso: <http://ai-nlp.info.uniroma2.it/basili/didattica/BigData/>.

## 2 Attività Accademica

**Posizione di Insegnamento Corrente:** Professore Associato presso la Facoltà di Ingegneria della Università degli Studi di Roma, Tor Vergata, dall'Ottobre 2004, dove svolge attività didattica nei corsi di *Basi di Dati*, *Web Mining e Retrieval* ed *Intelligenza Artificiale, Gestione dei Dati e della Conoscenza*.

Professore Associato confermato dal 30 Ottobre del 2007, presso la Facoltà di Ingegneria della Università degli Studi di Roma, Tor Vergata.

Professore Associato non confermato dal Maggio 2003 all'Ottobre 2007, presso la Facoltà di Ingegneria della Università degli Studi di Roma, Tor Vergata.

Ricercatore presso la Facoltà di Ingegneria della Università di Roma, Tor Vergata, dal Dicembre 1994 al Maggio 2003.

Editor in Chief della rivista "*Italian Journal of Computational Linguistics*", dal Settembre 2015.

Membro del Comitato Editoriale della rivista "*Intelligenza Artificiale*", IOS Press, dal Settembre 2016.

Membro del Consiglio Direttivo della Associazione Italiana per l'Intelligenza Artificiale, AI\*IA dal Settembre 2005.

Membro del Consiglio Direttivo della Associazione Italiana per la Linguistica Computazionale, AI-LC dal Settembre 2015.

Coordinatore del Gruppo di Semantic Analytics (SAG) dell'Università degli studi di Roma, Tor Vergata, dal Gennaio 2008.

Membro del Centro Interuniversitario di ricerca sull'Elaborazione Cognitiva in Sistemi Naturali ed Artificiali (ECONA) dal 1996.

Membro del Gruppo *ART*, *Artificial Intelligence at Tor Vergata*, del Dipartimento di Informatica, Sistemi e Produzione della Università di Roma, Tor Vergata dal 1990.

Responsabile del Laboratorio di Intelligenza Artificiale del Dipartimento di Informatica, Sistemi e Produzione della Università di Roma, Tor Vergata dal 1991 al 2007.

Responsabile del Programma Erasmus per il Corso di Studio di Ingegneria Informatica, Università di Roma, Tor Vergata, dal Febbraio 2005 al Marzo 2012.

### 2.1 Conseguimento di premi e riconoscimenti per l'attività scientifica

- IBM Best paper award COLING Conference Dublin, August 2014.
- Second Best paper award CICLING Conference, Iasi, Romania, March 2010.
- First Best paper award, CICLING Conference, Mexico City, Mexico, March 2009.

- Best Paper Award at 2009 Conference on Analysis of Noisy Documents (AND2009), Barcelona, 2009.
- Second Best system at the MIREX 2005 (Music Information Retrieval Evaluation Campaign) for the automatic classification of musical genres over symbolic (i.e. MIDI) representations (see the *Symbolic Genre Classification Results* at [http://www.music-ir.org/mirex/wiki/2005:Symbolic\\_Genre\\_Classification\\_Results](http://www.music-ir.org/mirex/wiki/2005:Symbolic_Genre_Classification_Results)).
- Among the best systems (12-13 among 86 competing systems) at the last Semeval 2012 Campaign in the task of Semantic Text Similarity (<http://www.cs.york.ac.uk/semeval-2012/task6/>)
- Best system at the ACL SemEval 2016 competition, task "Community Question Answering" at <http://alt.qcri.org/semeval2016/task3/>.

## 2.2 Comitati di Programma ed Organizzazione

Co-chair of the CliC-It 2017 Conference, Roma, Italy, 11-13 December 2017.

Co-chair of the AI\*IA 2017 Conference, Bari, Italy, November 2017.

Area co-chair of the *Semantics* track of the EMNLP-2017 (*Empirical Natural Language Processing*) Conference, Copenhagen, Denmark, September 2017.

Co-Chair of the First Italian Conference on Computational Linguistics, held in Pisa December 2014.

General chair of the Italian Workshop on Information Retrieval, 2014.

Program co-chair (with G. Semeraro) of the Italian Workshop on Information Retrieval, 2013.

Area co-chair (with Suzanne Stevenson) of the *Semantics* track of the NAACL-2012 (*North American Chapter of ACL*) Conference, Montreal, July 2012.

Member of the Program Committee of the CoNLL-2008, The Twelfth Conference on *Computational Natural Language Learning*, Manchester, UK, August 16-17th 2008.

Program chair del 10° Congresso Internazionale AI\*IA, *Artificial Intelligence and Human-oriented Computing*, Roma, 10-13 Settembre 2007.

Membro del Comitato di Valutazione della "International Conference on Computational Linguistics, COLING", dalla edizione del 2002.

Membro del Comitato di Valutazione della "International Conference of the Association for Computational Linguistics, ACL", dalla edizione del 2002.

Membro del Comitato di Valutazione della "International ACM Conference of the Special Interest Group on Information Retrieval, SIGIR", dalla edizione del 2002.

Member of the Program Committee in the Workshop on *Information Extraction Beyond The Document*, jointly held in ACL 2006, Sydney, Australia, July, 2006.



Member of the Program Committee in the ACL/COLING Workshop on *Ontology Learning and Population (OLP2)*, jointly held with ACL 2006, Sydney, Australia, July, 2006.

Program chair del Workshop on "*Learning Structured Information in Natural Language Applications*", presso lo European Chapter of the Association for Computational Linguistics, Trento, Aprile 2006.

Program chair del Workshop on "*Toward Computational Models of Literary Analysis workshop*", presso la International Conference On Language, Resources and Evaluation (LREC 2006), Genova, Maggio 2006.

Program chair of the Workshop on *Beyond Named Entity Recognition: Semantic labelling for NLP tasks*, in association with the 4th International Conference On Language Resources And Evaluation, LREC2004, LISBON, Portugal, 25th may 2004.

Program Chair del Workshop ROMAND 2002, "*ROMAND 2002, 2nd Workshop on RObust Methods in Analysis of Natural language Data*", European Space Agency, Esrin, Frascati, 17 July, 2002.

Membro del Comitato di Programma del "*ECAI 2002 WS on Machine Learning and Natural Language Processing for Ontology Engineering*", Lyon, France, July 2002.

Organizzatore della SCIE *International Summer School on Information Extraction*, Frascati (Italy), nelle edizioni 1997, 1999 e 2002.

Membro del Comitato di Programma della "*2nd Learning Language in Logic (LLL) Workshop*", 13th - 14th September 2000, Lisbon - Portugal, Co-located with the *International Conference on Grammar Inference (ICGI)* and the *Conference on Natural Language Learning (CoNLL)*.

Membro del Comitato di Programma della "*EKAW'2000, 12th International Conference on Knowledge Engineering and Knowledge Management*", Juan-les-Pins, French Riviera, October 2, 2000.

Membro del Comitato di Programma, "*ROMAND 2000, 1st workshop on RObust Methods in Analysis of Natural language Data*", Lausanne, Switzerland, October 19-20 2000.

Membro del Comitato di Programma del Workshop "*TANLPS - Towards adaptive NLP-driven systems: linguistic information, learning methods and applications*" presso la European Conference on Machine Learning (ECML), Chemnitz University, Germany, April 24, 1998.

Membro del Comitato di Programma del Workshop "*Adapting Lexical and Corpus Resources to Sublanguages and Applications*", presso la prima LREC Conference, Granada (Spain), May, 1998.

Membro del Comitato di Programma del Workshop "*The Evaluation of Parsing Systems*", presso la prima LREC Conference, Granada (Spain), May, 1998.

Membro del Comitato di Programma della "*Second International Conference on Natural Language Processing and Industrial Applications*", Université de Moncton, New Brunswick (Canada), August 1998.

### 3 Collaborazioni Scientifiche

- Qatar Computing Research Institute (QCRI), Doha, Qatar: Sviluppo di framework ad oggetti per la progettazione di algoritmi di Machine Learning strutturato agenti su dati linguistici di larga scala. Il gruppo SAG di Tor Vergata e il QCRI sono i responsabili della piattaforma open-source KeLP per metodi kernel applicati alla elaborazione linguistica dei dati testuali (URL: <https://github.com/SAG-KeLP>). Dal Novembre 2015 ad oggi.
- Centro Ricerche RAI (Torino): ricerca e sviluppo sugli strumenti per la annotazione semantica automatica di notiziari radiotelevisivi. Dal Novembre 2003 al Dicembre 2007.
- Center for Language and Speech Processing, John Hopkins University, Baltimore, 2003. Co-Coordinator of the 2003 Summer Workshop on "*Semantic Analysis Over Sparse Data*".
- Free University of Brussels, gruppo di Software and Knowledge Engineering, coordinato dal prof. Robert Meersman, su tecniche di rappresentazione della conoscenza per l'accesso e l'integrazione di risorse (basi di conoscenza) eterogenee, dal 1999.
- University of Sheffield, Natural Language Processing group, coordinato dal prof. Yorick Wilks, dal 1996, per i modelli di acquisizione lessicale per l'adattamento automatico della semantica lessicale dei verbi, e la loro sperimentazione congiunta sulle lingue Inglese ed Italiano. Tale collaborazione iniziata all'interno del progetto ECRAN, Esprit Project n. 2110, nel periodo Dicembre 1996 è tuttora stabilmente in corso.
- Università de Paris 7, gruppo TALANA per la Elaborazione del Linguaggio Naturale, coordinato dal Prof. Laclois, dal 1996, per la sperimentazione congiunta sul francese e l'italiano delle tecniche di acquisizione lessicale progettate dal gruppo della Università di Roma Tor Vergata.
- Fondazione Ugo Bordoni dal 1996, per la applicazione di tecniche di elaborazione del linguaggio naturale a problemi di reperimento automatico di informazione in ambienti distribuiti.
- IBM T.J. Watson Research Center di YorkTown Heights, per il periodo Agosto 1992 - Agosto 1993, per lo studio cross-linguistico (italiano/inglese) di tecniche di estrazione di conoscenza lessico-semantica da testi liberi.
- Istituto di Studi sulla Ricerca e sulla Documentazione Scientifica del C.N.R., durante il periodo 1991-1992, per la ricerca sulla applicazione di metodi di *rappresentazione della conoscenza e reasoning* alla gestione ed alle interfacce di basi di dati statistiche.
- Università di Ancona, dal 1990 per la ricerca congiunta su modelli robusti di prestazione linguistica ed algoritmi statistico-simbolici per l'acquisizione di conoscenza lessicale da testi liberi.
- ESA, Agenzia Spaziale Europea durante il periodo 1989-1991, per lo studio, la progettazione e lo sviluppo di un sistema esperto per l'accesso intelligente ai dati telerilevati

### 4 Attività di Terza Missione: Progetti di Ricerca applicata a finanziamento pubblico o industriale

Le attività di ricerca di seguito presentate sono state condotte in un regime di finanziamento spesso garantito da risorse private, leate a progetti applicati a

problemi o settori industriali o da fondi europei per l'innovazione.

- *Socio fondatore di Reveal s.r.l.* spin-off accademico, con sede a Roma, attivo dal 2012 nel settore delle tecnologie di *Machine Learning* ed *Intelligent Information Retrieval* applicate a dati documentali, organizzativi e normativi delle imprese (ad es. banche) ed al *sentiment analysis* applicato su large scala ai dati dei Social Media (ad es. Twitter). Principali clienti e use case di Reveal: UniCredit, Monte Paschi Siena, AkerSolution (Norway), Comune di Roma, XeroX, PricewaterhouseCoopers .
- *Responsabile Scientifico* per il Dipartimento di Ingegneria dell'Impresa dell'Università di Roma Tor Vergata, nel progetto "*Semantic Search Engine: Enterprise Search and Process Management*" finanziato da UniCredit, Milano, Dicembre 2013-Marzo 2015.
- *Responsabile Scientifico* per il Dipartimento di Ingegneria dell'Impresa dell'Università di Roma Tor Vergata, nel tutoraggio della Tesi di Dottorato "*Social Media Analytics for Brand Reputation*" finanziata da Telecom Italia, Roma, Dicembre 2013-oggi.
- *Main investigator* per l'unità di Roma Tor Vergata, nel progetto europeo INSEARCH (Dicembre 2010-Dicembre 2012)
- *Main investigator* per l'unità di Roma Tor Vergata, nel progetto denominato DIVINO, coordinato dalla azienda Mastroberardino, nell'ambito delle attività di market intelligence nel dominio enogastronomico, attraverso la automazione dei processi di Web Mining ed Opinion Analysis come studiati e sviluppati nell'ambito di Industria 2015. Dicembre 2009 - Marzo 2014.
- *Main investigator* per l'unità di Roma Tor Vergata, nel progetto finanziato da parte della agenzia FILAS, Progress-It (dal Luglio 2012 al Febbraio 2014)
- *Responsabile di Unità* della Università degli Studi di Roma Tor Vergata, nel progetto PRIN 2008, PARLI "Portale per l'Accesso alle Risorse Linguistiche per l'Italiano" (2009-2012)
- *Main investigator* nel progetto *MIND - Companion Front End and Information Broker*. Dialogo Uomo macchina in ambienti di servizio multifunzionali, mobili e Web. Progetto finanziato dal Ministero delle Attività Produttive, Marzo 2006-Marzo 2007.
- *Main investigator* nel progetto "*RSRnT*". *Riconoscimento di Strutture informative Relazionali da Testi annotati*. Applicazione di modelli statistici per la estrazione di relazioni concettuali dai testi, in domini di business and security intelligence. In collaborazione con la CM Sistemi S.p.A. e la Direzione Nazionale Antimafia. Dal Febbraio 2007.
- *Main ART investigator* in *Prestospace* (Progetto europeo 6<sup>th</sup> Framework, IST-FP6-2003-507336) sulla analisi semantica di dati multimediali per la archiviazione e la pubblicazione degli archivi digitali radiotelevisivi delle principali televisioni e centri audiovisivi europei (BBC (UK), RAI (Ita), INA (Fr)), dal Febbraio 2004 al Gennaio 2008.
- *Main investigator* del gruppo di Intelligenza Artificiale della Università di Roma Tor Vergata, sub-contractor del progetto *ff-Poirot, financial fraud Prevention Oriented Information Resources using Ontology Technology* (Progetto europeo 6<sup>th</sup> Framework, IST 2001-38248) sulla analisi di dati Web basata su ontologie per la lotta alla frode su rete, dal 2004 al 2006.

- *Senior investigator* del team di ricerca "Semantic Analysis Over Sparse Data" del John Hopkins Summer Workshop, finanziato dal National Science Foundation (NSF), Luglio-Agosto 2003. Lo scopo è verificare la applicabilità di metodi di apprendimento automatico per un approccio semantico ai problemi di *data sparseness* che affliggono molte aree del NLP.(URL: <http://www.clsp.jhu.edu/ws03/>)
- Progetto Nazionale Industriale dell'Istituto delle Poste e Telecomunicazioni, TAL (*Trattamento Automatico della Lingua Italiana*), dal 1999 al 2001. Coordinatore tecnico per l'Università di Tor Vergata, presso il Consorzio di Ricerca Industriale ed Universitaria CERTIA, membro del gruppo TAL.
- NAMIC, IST-HLT (IST-12392) project, dal Febbraio 2000 al Giugno 2004. Responsabile tecnico dell'unità di Roma, Tor Vergata. Definizione e sperimentazione di modelli di *authoring* automatico in collezioni di testi multilinguali (notizie di agenzia). Automazione in ambiente distribuito (Web) dei processi di riconoscimento e classificazione di documenti, estrazione di informazione finanziaria e sportiva e delivery dei metadati generati.
- MURST 1998-2000 "Agenti Intelligenti per l'Estrazione di Informazione Distribuita". La ricerca ha avuto come obiettivo lo studio della applicazione di tecniche di elaborazione robusta dei testi per la classificazione automatica dei testi e la estrazione di informazione semantica.
- *Web Learning* (Progetto cofinanziato CNR) . Responsabile dell'unità di Roma Tor Vergata nella linea AEIOU. Il tema del progetto è lo studio, la progettazione e lo sviluppo prototipale di metodi e strumenti per il Web Learning, orientati alla diffusione delle conoscenze, sia di base che specialistiche, nel settore delle tecnologie avanzate. La linea AEIOU si concentra sugli aspetti di Innovazione Tecnologica, Organizzazione e Usabilità per gli Ambienti Educativi Interattivi.
- TREVI, Esprit Telematics project n. 21130, dal Novembre 1997 al Dicembre 1999. Responsabile tecnico della unità di Roma, Tor Vergata, nella ricerca su tecniche di larga scala per l'elaborazione di testi, e la progettazione di un sistema per la categorizzazione, l'"enrichment" e il *delivery* di notizie di agenzia, secondo una esplicita rappresentazione di profili d'utenza. Il prototipo industriale risultante è uno dei primi esempi di sistema distribuito e orientato agli oggetti per la analisi dei testi basata su tecniche avanzate di elaborazione del linguaggio naturale.
- ECRAN, Esprit Project n. 2110, Dicembre 1995, Dicembre 1998. Responsabile tecnico dell'Unità di Roma, Tor Vergata, nella ricerca sulle tecniche di acquisizione semantico-lessicale in scenari di estrazione di informazione da testi. Il risultato del progetto è un prototipo di estrazione di informazione in ambito finanziario basato sull'architettura GATE, per tre lingue (Francese, Inglese e Italiano), e dotato di capacità di adattamento al dominio conoscitivo. Tale funzionalità è stata ampiamente utilizzata per la costruzione semiautomatica delle estese basi di conoscenza richieste dal sistema.
- Nomos, ESPRIT Project n. 5330, per il periodo Gennaio 1991-Ottobre 1992, in collaborazione con SOGEI e Tecsiel, per la ricerca sui metodi di acquisizione lessicale applicati al linguaggio legale e la progettazione dei relativi sistemi prototipali.

## 5 Introduzione alla discussione delle Attività di Ricerca

Le attività di ricerca svolte coprono il periodo dal Settembre 1992 sino ad oggi e meritano quindi una ampia discussione analitica. Per una più chiara esposizione sono state definite aree ed attività all'interno di quattro fasce temporali, individuate funzionalmente ai temi ed alle diverse fasi professionali. Saranno quindi dedicate alcune sezioni distinte alle descrizioni sintetiche delle attività relative ai periodi:

- 1992-1997: Studi sulla rappresentazione della conoscenza e sulle tecnologie per il trattamento automatico delle lingue naturali.  
Principali obbiettivi della ricerca: Forme ibride di rappresentazione della conoscenza (che integrano logica e metodi quantitativi), fondamenti di Elaborazione del Linguaggio Naturale (NLP) e sviluppo di strumenti abilitanti per sistemi di NLP.
- 1997-2004: Metodologie di Progettazione di applicazioni Web distribuite per l'Information Retrieval e l'Information Extraction.  
Principali obbiettivi della ricerca: Ingegnerizzazione di sistemi Web per l'accesso intelligente ai dati distribuiti (nel Web)
- 2004-2008: Modelli statistici di apprendimento automatico per problemi complessi di *pattern matching* e per le applicazioni intelligenti in ambito linguistico e multimediale.  
Principali obbiettivi della ricerca: Fondamenti teorici e applicazioni di modelli quantitativi ed induttivi del significato, nei dati, testi e documenti multimediali
- 2008-oggi: Modelli vettoriali della semantica lessicale, Apprendimento basato su kernel e fenomeni linguistici complessi tipici delle reti sociali: apprendimento di lessici computazionali su larga scala, Question Answering, Web Semantic Search, Social Data & Big Data analytics.  
Principali obbiettivi della ricerca: Sviluppo di metodologie di apprendimento automatico accurate e scalabili per task semantici complessi, mediante l'integrazione efficiente di metodi di *supervised* ed *unsupervised learning*.

Per una più semplice fruizione di questo documento si propone qui una discussione in ordine temporale inverso.

## 6 Descrizione Sintetica delle Attività di Ricerca: anni 2008-oggi

Le attività di ricerca principali di questo periodo si sono sviluppate lungo tre aree principali:

- *Supervised and Unsupervised Computational Natural Language Learning*
- *Semantic Search and Social Media Analytics*
- *Natural Language Learning for Human Robot Interaction*

### 6.1 Supervised and Unsupervised Computational Natural Language Learning

#### Motivazioni e Attività principali

In continuità con la ricerca degli anni precedenti i problemi di apprendimento da esempi basati su metodi *supervised*, e soprattutto il successo nella acqui-



sizione di modelli lessicali dai corpora di grandi dimensioni attraverso i cosiddetti *wordspace model*, che si caratterizzano come modelli pienamente *unsupervised*, già esplorati in IC68, sono stati più recentemente estesi attraverso la progettazione di kernel complessi di tipo sintattico-semantic. Questi integrano informazioni a diversi livelli linguistici: dal livello lessicale espresso da forme vettoriali di rappresentazione a quello sintattico, impliciti nel matching ricorsivo sfruttato dai *tree kernel*. Per lo sviluppo di tali kernel sono stati integrati metodi distributional per la acquisizione automatica di lessici molto espressivi da corpora testuali, studiati ampiamente in diversi lavori (e.g. IC105, IC94, IC96, IC113) combinati in diversi modi in *tree kernel* esistenti.

Il modello di kernel introdotto in IC105 è noto come *Smoothed Partial Tree Kernel* (SPTK) integra informazione lessicale (espressa dalle foglie nelle rappresentazioni dei *parse tree* derivati da una frase) all'interno delle funzioni di similarità già usate nei *tree kernel*, totalmente dipendenti dalla struttura sintagmatica. Il modello SPTK estende quindi un *tree kernel* iniettando semantica lessicale di tipo vettoriale alle foglie. Ampiamente sperimentato SPTK si è dimostrato superiore ai diversi kernel precedentemente proposti dalla letteratura in diversi task quali *Question Classification* (IC105, IC106), *human-robot interaction* (IC133) e *verb classification* (IC115). Infine, il recente lavoro IC136, estende ulteriormente il kernel SPTK esprimendo un modello compositivo delle semantica lessicale basato sul *parse tree* di una frase. Il modello propone un'algebra ai nodi di un *parse tree* tra i vettori dei costituenti lessicali, in modo da associare informazioni semantiche sia ai nodi terminali (corrispondenti ad entità lessicali semplici) che ai nodi non terminali (come composizione matematicamente significativa della informazione lessicale relativa ai costituenti): questo risulta in una nuova famiglia di kernel, detta *Compositionally Smoothed Tree Kernel* (CSTK), che fornisce funzioni più espressive come mostrato ampiamente in diverse sperimentazioni (IC136, IC121) tutt'ora in corso.

Una linea di ricerca interessante che affronta la efficienza delle fasi di addestramento per kernel complessi è quella dello studio di algoritmi *on-line*. Questi pur producendo soluzioni (cioè modelli) ottimi solo localmente all'insieme delle osservazioni rese disponibili sino ad un certo istante di tempo, sono interessanti per la loro scalabilità rispetto a grandi volumi di dati in input e per la loro adeguatezza cognitiva come paradigmi incrementali di apprendimento. I modelli di kernel linguistici complessi (per esempio l'SPTK visto prima) forniscono diversi spunti alle tecniche di manutenzione dei modelli che caratterizzano gli algoritmi *on-line* basati su kernel, come il *Passive Aggressive learning*. In particolare, sono state proposte durante questa ricerca alcune strategie originali atte a gestire intelligentemente gli esempi in ingresso, riservando modelli più complessi solo ai casi più difficili. Questi risultano in metodi molto efficienti che producono una stratificazione naturale del modello ottenibile (IC137) di facile manutenzione e distribuzione del carico in reti di calcolatori.

E' da osservare che studi estesi sono stati fatti anche nella applicazione ibrida di schemi supervised (ad esempio basati su *tree kernel*) e schemi unsupervised basati su modelli vettoriali del lessico per l'apprendimento di basi di conoscenza linguistiche specifiche (ad esempio, classi concettuali di verbi o *frame* linguistici), nella prospettiva di modelli di apprendimento della lingua verosimili dal punto di vista cognitivo (Bk26, IC115).

### 6.1.1 Risultati

Lo sviluppo dei *tree kernel* cosiddetti semantic, che integrano informazioni lessicali (in forma di algebre vettoriali agenti sulle parole) e le informazioni grammaticali proprie dei *parse tree* in input è discusso in [IC105, IC106, IC113, IC114, IC115, IC136]. Un primo insieme di lavori si occupa della applicazione di kernel semantic alla acquisizione di informazioni linguistiche come i *frame* del paradigma della *Frame Semantics* [Bk26, IC86, IC97] o i modelli ontologici del lessico [IC80, IC82, IC101, IC115] ed della loro applicazione a task di analisi

semantica, ad es. il *semantic role labeling* [IC92, IC100]. I lavoro che estendono i paradigmi di on-line learning in forme utili alla analisi di fenomeni semantici sono discussi in [IC118, IC123, Bk27, IC137]

## 6.2 Semantic Search and Social Media Analytics

### Motivazioni e Attività Principali

Negli scenari del Web e delle reti sociali su Internet, l'ultima decade ha visto crescere l'interesse sui temi della ricerca di informazione distribuita per i processi tradizionali di *information retrieval* (*search engines*) ma anche per fenomeni nuovi come la analisi dei comportamenti utente (profilazione e personalizzazione) ed i processi collaborativi (blogs, forum e *recommending systems*). Queste attività sono intimamente legate alle lingue poichè insistono quasi esclusivamente su dati non strutturati (come le interrogazioni eseguite verso i motori di ricerca che caratterizzano un utente o gli interventi degli utenti nei social network).

E' stata avviata una specifica linea di ricerca sulla progettazione di strumenti di ricerca semantica basati su lessici vettoriali acquisiti in base a paradigmi di *wordspace*. Tali lessici basati sulla analisi alle co-occorrenze di corpus estesi di testi e integrando tecniche algebriche di fattorizzazione matriciale consentono di derivare autonomamente lessici di dominio di grande scala e sono molto importanti nei processi di *enterprise search*. Questi infatti sono utili alle ricerche basate sulla expertise di domini ristretti ma hanno bisogno di confrontarsi su scala Web. Quindi richiedono la precisione di modelli fortemente concettuali ma anche robustezza e flessibilità da esibire nel mondo dei contenuti aperti del Web. In diversi progetti di ricerca queste attività hanno consentito lo sviluppo di piattaforme industriali di *search*, che tengono conto della atura concettuale del dominio della impresa e, al contempo, di lavorare con lessici e modalità di interrogazione molto flessibili che vanno dalla interrogazione per concetto, tramite frasi o frammenti documentali o direttamente per keyword. Tali attività sono documentate ad esempio in IC102, IC110, IC117.

La analisi dei testi è anche uno dei fattori abilitanti di forme molto specializzate di personalizzazione ed analisi delle comunità nelle reti sociali. Infatti la opinione espressa dagli utenti del Social Web si esprime attraverso frasi o discorsi di natura soggettiva, la cui caratterizzazione emozionale (cioè il riconoscimento e la estrazione di preferenze, gusti o opinioni) non può prescindere dalla analisi linguistica del testo. La ricerca condotta a Tor Vergata nel laboratorio SAG si è concentrata nello sfruttamento di dati annotati di facile reperimento per la induzione di modelli di opinione (intesa come la polarità delle espressioni soggettive dell'utente). In tale ambito sono state trattate lingue (Italiano ed Inglese) e media diversi (forum, review in linea ma anche tweet) attraverso modelli di *machine learning* basati su *kernel* complessi (lessicali, grammaticali e semantici) già precedentemente discussi 6.1. Uno dei lavori recenti [IC136], che ha meritato il premio IBM per il best paper nella conferenza COLING 2014, tenutasi a Dublino nello scorso Agosto, è basato sulla applicazione di metodi di *Structured Learning*, noti come SVMHMM, che combinando le proprietà dei modelli markoviani (HMM) con la flessibilità dei metodi basati su *Support Vector Machines* consente il trattamento di intere sequenze di Tweet (sequenze dialogiche rispetto ad un tema) e fornisce strumenti predittivi di particolare efficacia.

Tutte le sperimentazioni sui metodi di Opinion Mining sono state anche usate per la partecipazione a campagne valutative internazionali (Semeval organizzata dalla American Association of Computational Linguistics) o nazionali (Evalita 2011, 2014) a partecipazione accademica ed industriale con ottime prestazioni:

- Evalita 2011 Task: "*Frame Labeling over Italian Texts*" (Best system)
- SemEval 2012 Task: "*Semantic Text Similarity*"

- StarSem 2013 Task: "*Semantic Text Similarity*" (Best System on "Semi-structured Text Similarity" recognition sub-task)
- SemEval 2013 Task: "*Sentiment Analysis in Twitter*"
- SemEval 2013 Task: "*Spatial Role Labeling*"
- Evalita 2014 Task: "*Sentiment Polarity Classification in Twitter*" (Best System on the "Irony Detection sub-task")
- SemEval 2014 Task: "*Aspect Based Opinion Mining*" (Best System on the Topic Recognition sub-task)

### 6.2.1 Risultati

I risultati nell'ambito del Semantic search sono stati recentemente discussi in [IC110, IC109, IC112, IC117, IC125]. Sono stato approfonditi i risultati sui metodi di Opinion Mining e Sentiment Analysis nelle pubblicazioni [IC135, IC131, IC123, IC119]

## 6.3 Natural Language Learning for Human Robot Interaction

### Motivazioni e Attività Principali

L'apprendimento linguistico è certamente un processo sfidante per sistemi che accedono ai materiali testuali disponibili on-line ed in forma elettronica, a causa della complessità delle dipendenze che nei livelli linguistici diversi (morfologia, sintassi, semantica e pragmatica) caratterizzano i dati e che determinano spazi di ricerca per gli algoritmi di Machine Learning di formidabile dimensionalità. E' stato molto discusso nell'ambito dell'IA come l'apprendimento linguistico sia difficilmente disgiunto dai progressi cognitivi compiuti anche in altri sottosistemi, come quello visivo, motorio e percettivo, in generale. L'*embodiment* cioè dei processi di sviluppo concettuale che riguardano il linguaggio all'interno di un sistema, che fisicamente è alimentato da tutti i livelli cognitivi, è un obiettivo rilevante di ricerca. Questa ricerca cerca infatti di colmare il gap tra gli algoritmi di apprendimento del linguaggio esistenti e gli sviluppi possibili di essi ove integrati in automi o dispositivi robotici più complessi, che realisticamente siano cioè più vicini all'esperienza che caratterizza lo sviluppo dell'intelligenza umana. Le interfacce linguistiche robotiche (*Human Robotic Interfaces, HRI*) in questo senso costituiscono un ambito molto promettente per la sperimentazione di modelli di apprendimento linguistico più efficienti e, d'altro canto, per lo sviluppo di robot avanzati in grado di comunicare fluidamente in linguaggio naturale nei loro ambienti operativi.

In collaborazione con i Laboratori di Robotica Cognitiva della Università della Sapienza, le ricerche sono state sviluppate lungo due linee. La prima mira a definire **workflow linguistici generali per le piattaforme robotiche** indipendenti dal sistema sottostante e dal task operativo del robot. Questo consente da un lato la generalizzazione di molti metodi di trattamento del linguaggio naturale sfruttati in robotica cognitiva sino ad oggi, ma dall'altro valorizza molte ricerche generali in linguistica computazionale, che in ambito robotico possono essere riutilizzate abbattendo la complessità nello sviluppo dei sistemi futuri. Ad esempio diverse teorie lessicali, come ad esempio la *frame semantics* di Fillmore, spiegano il lessico della lingua ma consentono anche la automazione di diversi aspetti (ad esempio il trattamento delle strutture tematiche verbali e dei ruoli ad esse associate) che sono molto rilevanti del processo di interpretazione semantica dei testi, ma che caratterizzano anche i linguaggi di comandi robotici o il dialogo uomo-macchina. A questa linea sono dedicate le sperimentazioni discusse in IC133 o IC130, dove un workflow basato sui processi *re-ranking* e *semantic role labeling* è discusso ed ampiamente sperimentato. Il sistema discusso



in IC120 ha partecipato con successo alla competizione SemEval 2013 nel Task di "Spatial Role Labeling" ed applica una teoria linguistica dell'interpretazione spaziale dei testi nel processo di interpretazione interfacce robotiche.

La seconda linea tende ad investigare gli aspetti dinamici dello sviluppo linguistico, **integrando fortemente l'apprendimento e la interpretazione semantica nei robot**. In questi lavori si dimostra che interfacce robotiche basate su algoritmi di apprendimento automatico da esempi (ed in particolare orientate a sfruttare *tree kernel* e *kernel* semantici) sono particolarmente accurate e possono esibire proprietà significative di flessibilità preservando la robustezza dell'interpretazione in condizioni operative e task molto diversi tra loro. E' stata quindi realizzata tale architettura adattativa che attualmente è integrata in diverse architetture robotiche come discusso in IC130, IC133 ed IC124. In particolare, in questa ricerca, è stato messo a punto un ambiente per lo sviluppo di esempi per l'apprendimento in ambito HRI che ha consentito lo sviluppo di un corpus vocale (HuRIC, vedi IC132), completamente annotato dei livelli morfologici, sintattici, semantici ed operativi. Tale corpus ha, da un lato, consentito di sperimentare il workflow adattativo in fase di sviluppo, ma ha, d'altro canto, l'ambizione di determinare uno standard per lo sviluppo e la valutazione comparativa sistematica di interfacce robotiche, nell'ambito della ricerca in domotica.

### 6.3.1 Risultati

I principali risultati di questa ricerca sono discussi in conferenze internazionali in ambito robotico (e.g. Best paper in IC130 o IC133) ma anche in ambito di linguistica computazionale (e.g. IC132, IC120 o IC122). La valutazione nella campagna internazionale di valutazione Semeval 2013 è discussa in IC120.

## 7 Descrizione Sintetica delle Attività di Ricerca: anni 2004-2008

Le attività di ricerca principali di questo periodo si sono sviluppate lungo tre aree principali:

- *Computational Natural Language Learning*
- *Ontology Learning*
- *Acquisition and Processing of Multimedia Semantics*

di cui di seguito viene fornita una breve sintesi.

### 7.1 Computational Natural Language Learning

#### Motivazioni

Le tradizionali applicazioni dei metodi di apprendimento automatico al trattamento delle lingue, ispirate dai modelli di indagine lessicografici, sono generalmente basate su modelli bayesiani dei fenomeni classici delle occorrenze, cioè delle distribuzioni delle parole nei testi. Tali modelli ispirati alla statistica bayesiana (per esempio ai modelli a massima entropia) ipotizzano un modello strutturato del fenomeno aleatorio da indagare e supportano inferenze complesse nei problemi della ambiguità linguistica (per esempio Part-Of-Speech (POS) tagging, word sense disambiguation o parsing statistico). I problemi principali risiedono quindi nella definizione di un modello efficace del fenomeno e stimatori adeguati delle probabilità in gioco. Tali approcci vengono generalmente detti *generativi* poichè la distribuzione osservabile è generata da un modello stocastico del fenomeno sottostante, come nel caso degli automi a stati finiti probabilistici che caratterizzano i modelli markoviani del riconoscimento della lingua parlata.

Il recente sviluppo della teoria statistica dell'apprendimento (Vapnik, 1998) riconduce una vasta classe di problemi induttivi al caso della *classificazione*, ed in particolare alla classificazione lineare. Gli algoritmi induttivi di questa classe, detti *Support Vector Machines*, seguono il principio di minimizzazione del rischio (Vapnik, 1998) e derivano la migliore funzione ipotesi massimizzando il margine delle distribuzioni dei dati noti. Essi sono quindi modelli *discriminativi* dell'apprendimento e non richiedono una generalizzazione sistematica delle osservazioni. In questo senso, il vantaggio è che possono essere applicati ad una più ampia gamma di problemi, in cui i modelli generativi sottostanti non siano noti e per i quali non esistono stimatori efficaci. Sono stati quindi ampiamente utilizzati in problemi di riconoscimento linguistico. In particolare i processi complessi della interpretazione linguistica (per esempio la analisi semantica) sono stati qui ricondotti a sequenze di passi di classificazione. In particolare la teoria statistica dell'apprendimento fornisce una serie di risultati teorici che rendono disponibili (1) algoritmi ottimi di induzione da esempi (quali le *Support Vector Machines*) e (2) un approccio matematicamente rigoroso alla modellazione del problema, attraverso le funzioni *kernel*, che costituiscono le metriche di base per la stima della similitudine tra istanze negli spazi complessi sottostanti.

### Attività Principali

In questa area le attività principali della ricerca svolta sono state incentrate su:

- la caratterizzazione di alcuni task tradizionali del TAL, quali la analisi semantica, il riconoscimento dei nomi propri o la analisi delle interrogazioni nel *Question Answering*, nei termini di problemi di classificazione.
- Lo studio di funzioni kernel specializzate per il trattamento di fenomeni linguistici e la loro combinazione nell'apprendimento di task linguistico-testuali complessi, in particolare la classificazione automatica dei documenti ed il Semantic Role Labeling
- Lo studio di modelli unsupervised di apprendimento come strumenti per lo sviluppo incrementale di applicazioni intelligenti (*semi-supervised learning* o *boosting*)

La definizione di un sistema completamente basato su kernel per il Semantic Role Labeling e per la classificazione delle interrogazioni nel Question Answering rientra nelle attività dei primi due punti.

In particolare, il sistema (RTV-SRL) di Semantic Role Labeling sviluppato presso il laboratorio di AI del DISP è il primo sistema europeo di questo tipo: nelle recenti competizioni lanciate dalle iniziative SensEval 2004, CoNLL 2005 e Semeval 2007 le sue prestazioni si attestano tra le migliori del mondo. Nel task di Framenet labeling task in SensEval 2004 il sistema si è collocato in prima posizione. Inoltre, è risultato tra i primi 8 sistemi (tra 21 partecipanti) nel *semantic role labeling task* del CoNLL 2005 e secondo nella competizione SemEval del 2007 (Task 17, *semantic role labeling task* su testi in inglese).

L'indagine sulla nozione di kernel ed il suo impiego come modello di principi e strutture linguistiche ha prodotto un insieme di definizioni e strumenti che migliorano le qualità della classificazione ottenibile e consentono una ampia applicabilità anche in scenari dove la disponibilità di risorse (esempi etichettati) è minima. Questo caratterizza una migliore portabilità a nuovi domini, una robustezza maggiore al rumore in ingresso ed una maggiore applicabilità a task diversi. Un kernel originale basato sulla risorsa Wordnet è stato sperimentato su task di classificazione e rappresenta il primo esempio di kernel che esprime formalmente la conoscenza semantico-lessicale di una risorsa di larga scala.

La terza linea di ricerca invece ha indagato l'uso di modelli geometrici di similarità semantica per la analisi dei testi. In particolare sono state applicate

metodologie della Principal Component Analysis, nei termini della *Latent Semantic Analysis*, come paradigma per la acquisizione automatica di lessici di dominio, classi semantico-lessicali e concetti ontologici. Kernel di questo tipo sono stati applicati a diversi task quali la scoperta di concetti ontologici nei testi, il riconoscimento di fenomeni narrativi complessi nei testi letterari ed il tracciamento di notizie nei flussi di stampa. L'interesse di questi kernel è che essi costituiscono paradigmi rappresentazionali rilevanti per diverse inferenze statistiche in campo linguistico, nei termini di funzioni kernel efficaci per il *clustering* (*unsupervised discovery*) e la classificazione automatica (*supervised example-based learning*).

### 7.1.1 Risultati

Le pubblicazioni [IJ17, IJ15, Bk16, Bk20, IC71, IC70, IC68, IC64, IC63, IC62] sono il principale frutto della ricerca sui sistemi *kernel-based* di *Semantic Role Labeling* e di *Question Answering*. La piattaforma di RTV-SRL è stata usata nella competizione SensEval 2004, CoNLL 2005 e SemEval del 2007 per lo sviluppo del primo sistema di Semantic Role Labeling del mondo completamente basato su kernel. Attualmente tale piattaforma è ampiamente utilizzata all'interno di due progetti nazionali per la estrazione ed il riconoscimento automatico di relazioni semantiche nei testi. Una applicazione di kernel sintagmatici come modelli delle strutture proteiche nel riconoscimento di siti attivi è discusso in [IC69].

Kernel semantici basati su risorse e basi di conoscenze esterne alle distribuzioni dei dati (in particolare su Wordnet) sono discussi in [IC65] e [IJ15], nell'ambito del problema della classificazione automatica di documenti.

La linea di ricerca sui modelli di LSA applicati alla semantica lessicale ha prodotto modelli *unsupervised* di *annotazione semantica* [IC68, IC75] e modelli robusti di *Cross-language Information Retrieval* [IJ16, Bk17]. Quest'ultimo in particolare è lo strumento usato per la gestione di flussi di notizie multimediali e multilinguali (cioè notiziari televisivi RAI) trattate nel progetto Prestospace [Bk17].

## 7.2 Ontology Learning

### Motivazioni

I problemi recenti della interoperabilità semantica negli scenari delle applicazioni Web e nei processi automatici di ricerca di informazione sono basati in modo cruciale su forme complesse di rappresentazione della conoscenza. I tre elementi essenziali per una ingegneria di tali sistemi sono quindi le *ontologie*, i *metadati* ed i *linguaggi standard per la rappresentazione della conoscenza*.

I metadati ontologici consentono di accedere ai dati attraverso delle spiegazioni esplicite riguardo ai loro contenuti che favoriscono interpretazioni semantiche valide, localmente alle diverse, e in potenza divergenti, applicazioni. Le ontologie sono risorse problematiche poichè rappresentano modelli complessi dei domini applicativi.

Anzitutto la loro scala è molto estesa poichè la quantità di conoscenza necessaria alla interoperabilità in generale non è trascurabile. La ingegneria delle risorse ontologiche richiede quindi un ampio sforzo da parte di profili professionali altamente specializzati, che includono almeno ingegneri della conoscenza ed esperti del dominio. Un secondo problema è che i modelli sono in generale soggettivi, al punto che ontologie riferite a domini "simili", o totalmente sovrapposti, possono essere reciprocamente incompatibili a cause di divergenze o sovrapposizioni denotazionali che si manifestano in modo inconsistente ai livelli terminologico e strutturale.

La armonizzazione di schemi ontologici distinti è detta *mappatura* ed è in generale necessaria tra due, o più, ontologie. Quando il *mapping* avviene a

*run-time*, il processo di mappatura è in generale definito come *negoiazione del significato*.

La modellazione di processi di negoziazione ha recentemente attratto una ricerca massiccia nell'area della elaborazione del linguaggio naturale. La mappatura ontologica è stata inquadrata come un nuovo e forse più complesso processo di inferenza lessicale: la disambiguazione dei concetti di una ontologia sulla base delle evidenze linguistiche disponibili riguardo ad essi (come ad es., le definizioni testuali, i loro nomi o alias, la struttura interna delle denominazioni). Questa impostazione del problema della negoziazione del significato ha un fondamento nella tradizione che in linguistica computazionale si è occupata di acquisizione di conoscenza lessicale, ed in particolare nelle teorie e metodi della disambiguazione del senso delle parole (*word sense disambiguation*).

### Attività Principali

La ricerca condotta ha sfruttato i vantaggi che la adozione di una prospettiva linguistica porta riguardo all'*ontology mapping* (e alla *negoiazione del significato*). Anzitutto la acquisizione automatica da testi liberi (così come studiata sin dagli anni '90), costituisce uno dei passi di base della ingegnerizzazione di una ontologia di dominio. Al contrario della soggettività che caratterizza i metodi manuali di creazione di ontologie, i metodi di *Machine Learning* sostengono processi sistematici che forniscono in generale risultati oggettivi e riusabili. In secondo luogo, gli approcci orientati al linguaggio sono semanticamente ben fondati ed aumentano le accuratezza (in particolare la precisione) raggiungibile nella interpretazione delle primitive linguistiche. Infine, la scalabilità di approcci linguistici all'*ontology engineering* è più alta grazie alla disponibilità di immensi patrimoni testuali in forma elettronica.

Sono stati largamente applicati metodi di induzione automatica di dizionari terminologici di dominio per il popolamento automatico di ontologie basati su processi di induzione automatica dai testi [IC63, IC68, IC74, IC75]. Tali metodologie combinano la disponibilità di risorse lessicali già esistenti (ad esempio Wordnet) a modelli geometrici e statistici migliorando la precisione raggiungibile agli stessi livelli di copertura dei fenomeni. I problemi aperti sull'impiego di tali strumenti, sono stati indagati mediante indagini sperimentali su domini reali (e.g. la frode finanziaria su Web). I principali approcci studiati sono basati su modelli di Latent Semantic Analysis ed applicati alla disambiguazione del senso ed alla costruzione dinamica di strutture ontologiche. Il contenimento dei costi di sviluppo raggiungibili mediante tali metodi sono un elemento chiave per il successo delle tecnologie semantiche, rispetto ad uno scenario industriale dove i tempi di realizzazione sono sempre più percepiti come un fattore critico di successo.

#### 7.2.1 Risultati

I lavori in [IC68, IC74, IC75] sono stati i risultati principali di questa linea di ricerca. Nello studio [IC74] viene discusso il ruolo di modelli geometrici del significato<sup>1</sup> nella caratterizzazione semantica di pattern linguistici di parafrasi o implicazione testuale. In [Bk19, Bk21] viene discusso un modello ontologico generato automaticamente usato come risorsa semantica di base per un sistema di dialogo per l'*Interactive Question Answering*. Approcci basati sulla semantica latente del lessico nei testi letterari e la loro applicazione al riconoscimento di strutture narrative di interesse critico sono discussi in [IC61].

<sup>1</sup>Una discussione generale sull'approccio "geometrico" ai modelli di semantica lessicale e testuale è riportata nel lavoro *Teorie Geometriche del Significato*, nella rivista, a carattere filosofico, *Res Cogitans*, URL: <http://www.rescogitans.it/main.php?articleid=228>



## 7.3 Acquisition and Processing of Multimedia Semantics

### Motivazioni

Nell'ambito dell'elaborazione di video ed immagini, viene spesso enfatizzato il cosiddetto "semantic gap" tra la semantica di alto livello necessaria all'indicizzazione di materiale audio-visivo e le proprietà di basso livello offerte dall'analisi dei contenuti multimediali. Si presenta quindi la necessità di arricchire la semantica disponibile attraverso la fusione di quest'ultima con contenuti provenienti dalle altre forme espressive, associate all'audio ed al video: il testo ed il parlato qui svolgono un ruolo cruciale. Il valore degli archivi audiovisivi e delle crescenti risorse multimediali disponibili sul Web può essere appieno sfruttato solo mediante la automazione (almeno parziale) dei processi di riconoscimento di informazione semantica attraverso la interazione esplicita tra dati di livelli (e media) diversi.

In tale contesto i metadati giocano un ruolo centrale. Infatti nello scenario degli archivi radio-televisivi, ad esempio, caratterizzare informazioni e strumenti di ricerca ad alto livello è necessario per permettere agli utenti interessati di ritrovare efficacemente il materiale audio-visivo desiderato con significativi livelli di accuratezza.

### Attività Principali

La ricerca qui condotta si è incentrata sulla definizione di modelli vettoriali di rappresentazione dei metadati, in particolare ispirati a modelli geometrici della semantica dei dati testuali, come ad esempio LSA. Questi sono stati applicati nel processo di "discovery" automatico di correlazioni tra livelli multimediali diversi. In questi casi la interazione tra testi (che descrivono o convivono con le immagini nelle pagine Web) e le proprietà visuali di basso livello può essere oggetto di misura ed il riconoscimento di associazioni latenti può essere in gran parte automatizzato. Le associazioni derivabili sono rilevanti per diversi motivi. Primo, essi forniscono spiegazioni dichiarative dei fenomeni visuali osservati a basso livello, tramite le parole che a tali proprietà vengono assegnate. Questo ha implicazioni nella integrazione dei dati visuali propri di molte componenti della multimedialità con risorse semantiche pre-esistenti, come nel caso delle ontologie di dominio per la multimedialità. E' da sottolineare anche che tali informazioni sono utili (in questo sono fenomeni emergenti dalle analisi unsupervised dei dati) in tutte le fasi di sviluppo di risorse ontologiche in questi ambiti (*multimedia ontology learning*). Secondo, le parole associate alle proprietà visuali ed ai dati multimediali forniscono astrazioni molto rilevanti nei processi di retrieval: vengono infatti abilitati processi di indicizzazione delle immagini (o dei video) basati sulle stesse parole usate per le interrogazione, definendo il lessico derivato come linguaggio di interrogazione. La sperimentazione condotta dimostra che le correlazioni così indotte dai dati in modo automatico migliorano task ad alto livello, quali, la classificazione tematica di immagini ([Bk18,Bk22]) o la ricerca mediante linguaggi testuali di interrogazione ([Bk17]).

Una linea di ricerca specifica è stata invece definita sulla classificazione automatica dei dati musicali secondo schemi di classificazione in generi musicali. E' stata sperimentato l'impatto della adozione di algoritmi di apprendimento automatico supervisionati, agenti sulle proprietà a basso livello osservabili nei formati musicali diversi (cioè MIDI, audio e testi descrittivi, come le review critiche di dischi o opere). I risultati sono stati discussi nelle conferenze ISMIR del 2004 e 2005 [IC59], e RIAO 2007 [IC65]. In particolare, nella competizione MIREX 2005, associata alla conferenza (ISMIR 2005), il sistema sviluppato dal laboratorio di Intelligenza Artificiale della nostra Università si è classificato come secondo miglior sistema tra i partecipanti (Task: Symbolic music classification, URL: <http://www.music-ir.org/evaluation/mirex-results/sym-genre/index.html>).

### 7.3.1 Risultati

I lavori [Bk17,Bk18,Bk22] sono i risultati principali di questa linea. Il sistema Prestospace per la indicizzazione dei notiziari RAI è discusso in [IJ16,Bk17,IC73]. I modelli di classificazione automatica dei dati audio in generi musicali sono invece discussi in [IC59, IC65].

## 8 Descrizione Sintetica delle Attività di Ricerca: anni 1997-2004

In questo periodo i risultati ottenuti nella ingegnerizzazione di strumenti di elaborazione delle lingue hanno trovato una naturale sintesi con i problemi di accesso alle informazioni enfatizzati dallo sviluppo delle reti di calcolatori e del Web. Le attività di ricerca principali di questo periodo si sono sviluppate lungo tre aree principali:

- *NLP-based Text Retrieval and Classification*
- *Adaptive Information Extraction*
- Sistemi distribuiti di *Information Retrieval ed Extraction*

di cui di seguito viene fornita una breve sintesi.

### 8.1 Natural Language Based Text Classification and Filtering

Lo sviluppo delle tecniche di elaborazione del linguaggio naturale, a cui la ricerca qui discussa (in sezione 9) ha contribuito significativamente, e la esplosione della quantità di informazione testuale disponibile grazie alle reti di comunicazione, e principalmente al Web, ha prodotto in questi anni la forte esigenza di strumenti per la intermediazione intelligente su larga scala, di cui i motori di ricerca rappresentano solo un esempio. I requisiti principali di sistemi in grado di sfruttare le informazioni distribuite nel Web sono quindi una maggiore flessibilità computazionale rispetto ai modelli proposti dalla ricerca dei primi anni '90, e la possibilità di modellare in modo più efficace la comunicazione uomo macchina e la condivisione di informazione all'interno di sistemi multicomponente. Il superamento degli schemi di ricerca per parole chiave tipica dei motori di ricerca ha condotto in quegli anni all'utilizzo di strumenti di NLP per la estrazione di informazione dai testi utile a processi di indicizzazione e recupero automatici. L'Information Extraction (IE) si sviluppa in questi anni come la tecnologia che consente di riconoscere nei testi scritti informazioni di tipo concettuale e di rappresentarla in forma esplicita (*data record* o *templates*) al fine di sfruttarne le possibilità all'interno di tecnologie di ricerca tradizionali, quali quelle delle basi di dati. Come affermato nel tutorial presentato nel AIIA 99:

*"Information Extraction is the process by which an automatic system is able to process documents in a linguistically motivated way and derive a structured representation of (part of) their content. It is to be seen as a process through which a structure is derived from unstructured and noisy texts."*

La esplicitazione dei concetti di interesse impliciti in un testo pone quindi un insieme di problemi su cui la ricerca di questi anni si è concentrata. Tali problemi emersi in quegli anni godono oggi di una considerazione definitiva anche in ambito industriale nel Semantic Web che di quella fase rappresenta una consolidata e consapevole continuazione.

Un primo problema è dato dalla flessibilità richiesta alle componenti di NLP, dalla tokenizzazione dei testi sino al riconoscimento di fenomeni semantici, che

debbono poter essere adattati a diversi domini applicativi mantenendo le caratteristiche di accuratezza ed efficienza richieste dalla scala dei problemi applicativi (ad esempio i flussi di notizie di agenzie giornalistiche). Le architetture studiate in questi anni hanno sfruttato i paradigmi *client-server* al fine di decomporre il processo complessivo di IE in moduli riutilizzabili. La progettazione di sistemi distribuiti di IE quindi è stata consentita dalla definizione di moduli di processi linguistici, dalla definizione di stritture dati condivise da essi e dedicate, e dallo sviluppo di algoritmi di NLP largamente indipendenti dalla conoscenza linguistica di base, definita separatamente in vaste basi di conoscenza indipendenti. Esempi di tali basi di conoscenza sono le grammatiche ed i lessici computazionali necessari a trattare i fenomeni dei diversi domini.

Un secondo problema sollevato quindi è dato dalla scala richiesta alle risorse, in grado di garantire la necessaria adeguatezza ai fenomeni linguistici tipici del dominio applicativo. Lo sviluppo di tali risorse quindi richiede processi di ingegneria della conoscenza robusti e semi-automatizzati indagati attraverso gli studi sui modelli di acquisizione lessicale degli anni precedenti il 1997 (vedi sezione 9). Quei modelli hanno consentito di adattare il lessico e le grammatiche di un sistema di NLP al dominio applicativo abbattendo i costi di ingegnerizzazione di sistemi IE di vasta scala. A questa linea di ricerca è dedicata la successiva Sezione 8.2

Sulla base di sistemi di IE di questo tipo sono stati quindi in questi anni sviluppate diverse applicazioni (correlate a sistematiche sperimentazione su casi di studio industriali). L'uso di tecniche e strumenti avanzati di Elaborazione del Linguaggio ha consentito la progettazione di sistemi in cui la componente di IE alimenta i processi successivi di Information Retrieval. In questi ultimi sia il processo di indicizzazione che quello di reperimento (misura della attinenza dei documenti alla interrogazione) sono fondati su una rappresentazione esplicita del significato.

Il paradigma qui indagato è stato quello di spazi vettoriali basati su caratteristiche derivate tramite processi di IE, per esempio il riconoscimento nei testi di entità, quali persone o luoghi, o gli eventi a cui queste partecipano. Il modello di spazio vettoriale risultante ha una base semantica ed esprime più coerentemente il sottostante dominio conoscitivo, risultando più espressivo (per la descrizione dei contenuti e dei requisiti dell'utente). Infine esso garantisce una maggiore leggibilità (trasparenza semantica) dei risultati per l'utente finale.

Una esempio di tale paradigma è stato perseguito nella ricerca sulla *classificazione automatica dei testi*, tecnologia abilitante per ogni sistema che si muova su una scala comparabile a quella del Web. Modelli basati sulla elaborazione linguistica dei dati testuali di riferimento sono stati progettati e sperimentati su benchmark di ricerca e su casi industriali, come ad esempio discusso in [IJ11,IJ14,Bk13,IC41,IC43, IC46,IC51]. Alla base di tali modelli sono approcci induttivi di tipo quantitativo, cioè algoritmi lineari di classificazione, basati su esempi già classificati, fondati su caratteristiche (cioè proprietà osservabili) di tipo linguistico: lemmi e classificazioni sintagmatiche, occorrenza di nomi propri. Alcuni sistemi industriali di IE ed Information Retrieval sono ulteriormente discussi in Sezione 8.3.

## 8.2 Adaptive Information Retrieval and Information Extraction

La possibilità di sviluppare sistemi di IE in domini diversi dipende strettamente dalla possibilità di adattare tutte le risorse semantiche necessarie, qui intendendo i lessici del sottolinguaggio relativo alla applicazione, i dizionari terminologici di dominio, le grammatiche e le regole semantiche di interpretazione (interfaccia sintassi-semantica).

Algoritmi di acquisizione lessicale statistici e simbolici (discussi in Sezione 9) sono stati in questi anni applicati alla ottimizzazione di tali risorse diverse

per sistemi di IE, dando luogo a quelli che sono stati più tardi definiti *processi adattativi di IE* (*adaptive IE*).

L'obiettivo dell'adattamento è qui l'eliminazione di informazione improprie da lessici general purpose, l'individuazione di nuovi sensi delle parole, il riconoscimento di forme terminologiche (per esempio espressioni frasali, o *multi-word expressions*, non composizionali) e di regole di interpretazione per l'IE (interfaccia sintassi semantica dei fenomeni obiettivo della attività di estrazione).

**Adattamento del Lessico e della Grammatica** In quest'area sono stati proposti un insieme di algoritmi quantitativi per la disambiguazione lessicale e per la specializzazione di lessici per l'IE, come ampiamente discusso in [IJ9, Bk8, Bk9, IC27, IC30, IC35, IC40]. In particolare, in [IC52, IC54] è stata definita una misura originale per la stima di similarità semantica rispetto ad una tassonomia concettuale (in particolare, Wordnet): la sua applicazione alla determinazione dei sensi intesi in una risorsa ontologica è discussa in [IC50].

L'adattamento dei processi di riconoscimento grammaticale in Italiano ed Inglese è invece discusso in [IJ10, Bk14, IC27, IC30, IC31], sulla base dei risultati della ricerca sul parsing discussa in 9.

**Acquisizione di dizionari terminologici di dominio** Il problema della costruzione di dizionari terminologici di dominio è ampiamente discussa in [IC37, IC39, IC42, IC45, IC47].

**Estrazione di regole di IE da corpora** Lo sviluppo di sistemi di regole per il riconoscimento di entità ed eventi è stato a lungo riconosciuto come il paradigma dominante nell'IE. La natura complessa e costosa di questo processi di ingegneria della conoscenza è stato oggetto di ricerche volte alla definizione di metodi per la estrazione automatica di tali regole, atti a minimizzare i costi di sviluppo ed aumentare l'impatto e la diffusione della tecnologia di IE. I metodi utilizzati, risultato di algoritmi ibridi quantitativo-statistici e basati su induzione logica sono discussi ampiamente in [IJ10, IJ12, Bk9, IC27, IC30, IC34, IC36, IC40, IC44, IC48].

### 8.3 Sistemi distribuiti di *Information Retrieval ed Extraction*

I principali risultati delle ricerche relative a processi di *Adaptive IE* e *Text Classification* discussi sin qui sono stati ampiamente applicati nella progettazione di sistemi industriali all'interno di progetti nazionali (TAL) ed internazionali (TREVI, [IC16, IC7, Bk3, Bk2], NAMIC [IJ12, IC40, IC44]). Tali sistemi sono dotati di una maggiore flessibilità computazionale e consentono, come prime esperienze internazionali in quest'area, di modellare in modo trasparente la comunicazione all'interno di una architettura *client-server*. In particolare in TREVI è stato sviluppato un sistema distribuito per la estrazione di informazione e la classificazione di news giornalistiche in Inglese e Spagnolo. Il sistema NAMIC sviluppato in collaborazione con la agenzia ANSA ed il Financial Times ha esplorato le tecniche di adaptive Information Extraction nei domini finanziario e sportivo, sulla lingua Inglese ed Italiana. I risultati di questa ricerca hanno condotto alla progettazione ed implementazione di un prototipo industriale che ha costituito il nucleo di un prodotto commercializzato. Tali sistemi sono discussi in [IJ12, Bk10, Bk11, Bk12, IC34, IC40, IC44].

Le sperimentazioni estensive dei modelli di classificazione dei testi effettuate su collezioni industriali (all'interno dei progetti europei, TREVI e NAMIC) e su collezioni di valutazione tradizionalmente disponibili in ambito accademico (ACM Reuters 3 collection) hanno consentito la valutazione delle prestazioni in scenari realistici e dimostrato la superiorità del modello rispetto ai precedenti classificatori dello stesso tipo proposti in letteratura (e.g. classificatori bayesiani), come discusso in [Bk13, IC32, IC41, IC43, IC46, IC51].

L'utilizzo di processi di IE nello sviluppo di un sistema di Question Answering è stato oggetto di una ricerca significativa nell'ambito del progetto Moses. In questo scenario i metodi adattativi di IE sono stati applicati allo



sviluppo di un processo deduttivo di *question answering* (*automatic FAQ reply*) diretto agli studenti in ambito universitario sulle informazioni riguardo alla didattica accademica. I risultati architetturali e modellistici sono discussi in [IC53, IC56, IC57, IC58].

Una architettura completamente basata su agenti per l'IE e il Question Answering in ambito multilinguale, sviluppata presso il Laboratorio di Intelligenza Artificiale dell'Università di Roma Tor Vergata, è discussa in [IC55].

## 9 Descrizione Sintetica delle Attività di Ricerca, anni 1992-1997

La ricerca che ha caratterizzato questo periodo è concentrata su tre ambiti principali:

- La definizione e la sperimentazione di *modelli originali di rappresentazione della conoscenza*. Gli ambiti complessi di applicazione hanno riguardato lo sviluppo di sistemi esperti, database statistici, e problematiche specifiche legate al linguaggio naturale ed alla componente semantico-lessicale della conoscenza linguistica.
- *Algoritmi per l'apprendimento e l'estrazione automatica di conoscenza*, con una particolare enfasi sui problemi legati al trattamento automatico della lingua. Paradigmi diversi (logico-simbolici, neurali e probabilistici) sono stati esplorati relativamente a problemi complessi (dalla acquisizione di regole *neurofuzzy* per il riconoscimento automatico di forme, alla induzione di regole da esempi fino al trattamento statistico di dati testuali). In particolare modelli induttivi integrati logico-simbolici e probabilistici sono stati definiti per l'apprendimento di conoscenza linguistica, o la generazione automatica di forme tassonomiche di conoscenza (reticoli concettuali).
- L'*ingegnerizzazione* (definizione, realizzazione e valutazione) di *sistemi robusti su larga scala per la elaborazione di testi*. La ricerca in questa dimensione si è realizzata lungo tre aree principali.

La *ingegnerizzazione di sistemi di elaborazione linguistica* che, in base allo sviluppo degli studi contemporanei sulle lingue naturali, presenta problemi inerenti la sottospecificità dei modelli di rappresentazione e delle architetture coinvolte. La ricerca si è incentrata sulla definizione di architetture linguistiche per sistemi di larga scala. I principali risultati sono quindi legati ad una serie di algoritmi per i diversi livelli linguistici (analisi sintattica, disambiguazione sintattica e semantica) ed a modelli avanzati per la rappresentazione e lo scambio di dati linguistici.

Una particolare enfasi ha ricevuto la ricerca su *metodi di analisi sintattica robusta*, in grado di sostituire all'approccio monolitico (a grammatica sintagmatica eventualmente lessicalizzata) modelli a cascata di riconoscitori specializzati (basati su forme più efficienti e meno espressive, e.g. grammatiche discontinue specifiche e lessicalizzate). I principali risultati in quest'area hanno riguardato la definizione di parser robusti per l'estrazione di conoscenza da testi per la lingua Italiana e l'Inglese.

Infine la natura complessa e dinamica del linguaggio naturale ha richiesto l'approfondimento di *modelli di valutazione delle prestazioni* dei sistemi di trattamento della lingua. Questi ultimi che non si adattano ad approcci simulativi analitici tradizionalmente adottati dall'ingegneria del software sono stati affrontati nell'ambito delle ricerche svolte mediante l'introduzione di metriche specifiche e la migliore caratterizzazione di risorse di riferimento (oracoli).

Verranno di seguito approfondite le principali linee di ricerca ed i risultati secondo due aree di indagine classicamente riconosciute nell'Intelligenza Artificiale: la *Rappresentazione della Conoscenza* e la *Elaborazione del Linguaggio Naturale* (o *Natural Language Processing*).

## 9.1 Rappresentazione della Conoscenza

Nelle seguenti sezioni si tratteranno i principali studi e risultati ottenuti nell'area della rappresentazione della conoscenza e relativi a particolari aspetti del problema e a domini applicativi complessi. La ricerca in quest'area è relativa all'elaborazione automatica del linguaggio naturale verrà invece descritta nell'ambito dei capitoli successivi.

### Modelli di pianificazione in Sistemi Esperti Diagnostici

I modelli principali per la progettazione dei sistemi esperti sono tradizionalmente legati ai sistemi di regole e alle logiche terminologiche per la gestione di (ristrette proprietà di) ereditarietà. Il contributo principale della ricerca svolta in quest'area si riferisce a:

- i) l'estensione dei formalismi tradizionali disponibili (sistemi di regole (alla MYCIN), frames e reti semantiche), i cui limiti in termini di complessità ed espressività sono poi stati ampiamente sottolineati in letteratura, mediante l'adozione di un formalismo ibrido (rete semantica ed frames ad ereditarietà) e il progetto di un metodo di *reasoning* a più livelli basato sulla pianificazione della inferenze;
- ii) la modellazione degli aspetti intrinseci di incertezza (*subjectivity*, *uncertainty* e *vagueness*) nella soluzione di problemi diagnostici e della loro propagazione nel *reasoning*.

Gli studi conseguenti hanno suggerito una rappresentazione della conoscenza ibrida basata su una rete semantica specializzata, a due livelli di generalizzazione, dotata di meccanismi di ereditarietà e basata sull'uso di strutture di frames. Tale framework consente la modellazione del *reasoning* mediante l'uso di una componente strategica (pianificazione dei meccanismi inferenziali da adottare) ed una componente di inferenza vera e propria, dotata di meccanismi di valutazione della qualità delle soluzioni proposte (algoritmi di attraversamento in reti semantiche complesse guidati da funzioni di costo/utilità e operatori di propagazione, entrambe basate su rappresentazioni fuzzy dei concetti). Gli aspetti legati al modello originale di rappresentazione della conoscenza e *reasoning* e la loro applicazione nel sistema esperto sono discussi in [IC2, IC3].

### Rappresentazione della conoscenza e gestione dell'incertezza

L'incertezza e l'imprecisione che caratterizzano pervasivamente il ragionamento umano sono stati oggetto di studi estensivi in Intelligenza Artificiale. Nonostante il tradizionale utilizzo di modelli probabilistici, un approccio che è stato studiato, in particolare, è quello fondato sulle logiche di tipo *fuzzy* (o approssimate). I *fuzzy set* [Zadeh, 1965], sono stati tradizionalmente utilizzati come modelli di descrizione di: concetti vaghi ed imprecisi (ad es. *alte temperature*, *misure molto accurate*, ...), relazioni di natura semantica sul dominio che determinano regole imprecise e variabili linguistiche usate frequentemente nella descrizione della conoscenza (ad es. *molto*, *quasi*, *scarsamente*, ...) ricondotte in algebre *fuzzy* ad operatori sulle funzioni caratteristiche.

Insieme *fuzzy* per la rappresentazione di concetti come *accuratezza*, *risoluzione* e *copertura geografica* nel telerilevamento e come modelli per la selezione dei risolvibili (conseguenze logiche) dei passi elementari di inferenza sono stati definiti

in [IC2, IC3]. In tale framework sono state integrate anche tecniche per la composizione di criteri di ottimalità (*utility functions*) di tipo *fuzzy* in grado di garantire le proprietà opportune di monotonicità e predittività del *reasoning*.

#### **Rappresentazione della conoscenza: modelli sub-simbolici, reti neurali**

Una area specifica della ricerca è stata condotta per l'approfondimento del trattamento numerico della imprecisione, in relazione alla integrazione delle algebre *fuzzy* con i modelli neuronali (Rumelhart, 1987). Il rilievo di questo studio nasce dalla forte analogia tra i due paradigmi di modellazione della conoscenza, laddove sia i *fuzzy set* che le reti neuronali rappresentano una descrizione quantitativa di modelli di classificazione. Un modello integrato *neurofuzzy* per la supervisione, mediante basi di regole *fuzzy*, dell'apprendimento di un riconoscitore neurale di caratteri è stato proposto e sperimentato in [IC4].

#### **Rappresentazione della conoscenza e basi di dati statistiche**

Un linguaggio per la rappresentazione della conoscenza relativa ai dati ed agli operatori di una base di dati statistica è stato oggetto di studio in collaborazione con l'Istituto I.S.R.D.S. del C.N.R. [Bk2]. In tale caso è stato definito un modello di interfaccia intelligente tra un database statistico ed un utente non esperto. Sono stati affrontati i problemi logici di verifica della correttezza della query e di composizione delle operazioni da effettuare sui dati richiesti per svincolare l'utente non esperto da tali compiti.

#### **Rappresentazione di conoscenza linguistica in sistemi di NLP**

La specifica natura complessa della informazione linguistica e la relativamente scarsa attenzione da essa ricevuta nella aree di tradizionale indagine della ingegneria del software è alla base di una scarsa penetrazione degli strumenti di trattamento della lingua nel panorama delle applicazioni industriali. La crescente disponibilità di metodi e strumenti di elaborazione linguistica ha posto questi temi alla attenzione della comunità scientifica. La ricerca in questa area si è incentrata sulla modellazione di rappresentazioni che, pur adeguate espressivamente rispetto alla varietà di livelli linguistici necessari alla comprensione dei testi, rispondano a precisi criteri di trattabilità computazionale. Efficienza, portabilità tra domini conoscitivi e riutilizzabilità sono stati perseguiti nella progettazione di diversi sistemi su larga scala. ARIOSTO ([IJ4, IJ3, IJ2, IC5, IC6]), un sistema di estrazione di conoscenze lessicali da corpus estesi di testi, per la sua natura complessa (diverse componenti per il processamento e diversi sottosistemi indipendenti per l'apprendimento) è uno dei primi risultati in quest'area: una rappresentazione logica di conoscenza linguistica ha qui realizzato gli aspetti di trasparenza e trattabilità necessari alla elaborazione di diverse lingue (Inglese e Italiano) e diversi domini (telerilevamento, finanziario e scientifico-tecnico). Collegata con questi studi è la ricerca riguardo alla progettazione e lo sviluppo orientati agli oggetti di sistemi NLP di larga scala discussa in Sezione 8.

### **9.2 Elaborazione del Linguaggio Naturale.**

L'indagine sui *modelli di competenza linguistica* (Chomsky, 1965) ha caratterizzato gli studi sulla elaborazione del linguaggio naturale almeno fino ai primi anni '80. Contributi multidisciplinari hanno caratterizzato questa fase da definire *metodologica* in debito con la Linguistica, l'Informatica (in particolare l'Intelligenza Artificiale), la Filosofia (del linguaggio e Semantica) e la Psicologia (cioè Psicolinguistica e Psicologia Cognitiva).

I problemi legati alla gestione della varietà degli aspetti di pervasiva ambiguità dimostrati dal linguaggio nei diversi livelli di rappresentazione ed elaborazione, come pure alla intrinseca idiosincrasia che le parole di una lingua possiedono rispetto a regole generali di "legalità" dei comportamenti morfologici, sintattici e semantici sono stati infatti messi in evidenza dallo sforzo realizzativo seguito alla riflessione formale e metodologica degli anni '60 e '70.

Una nuova fase, che può definirsi *ingegneristica*, ha enfatizzato piuttosto gli aspetti della *prestazione* legata a fenomeni quali comprensione del linguaggio, inferenza e generatività linguistica a partire dalla fine degli anni '80. Un vero e proprio rinascimento delle tecniche numeriche è testimoniato dallo spazio crescente dedicato a metodi ed algoritmi quantitativi nelle riviste scientifiche e dalla nascita di alcune conferenze o gruppi di lavoro divenuti in breve forum di riferimento: l'enfasi di tale ricerca è su principi fondanti di sviluppo ed acquisizione di *modelli d'uso* del linguaggio ritenuti via principale (e praticabile) per sistemi robusti ed a vasta copertura richiesti da compiti linguistici reali. Metodi quantitativi ispirati al modello del 'canale rumoroso' (Shannon & Weaver, 1965) sono stati proposti in diversi problemi legati all'ambiguità linguistica.

E seppure in una prospettiva di imminente riconciliazione le due tendenze si sono mantenute durante questo periodo nettamente distinte per strumenti e fonti di ispirazione. E' nella integrazione tra i due approcci che si inserisce la ricerca condotta e di seguito dettagliata.

### Parsing Robusto dei linguaggi naturali

L'enfasi sui modelli d'uso della lingua e la necessità di strumenti di riconoscimento grammaticale accurati ed efficienti pongono tra i problemi cruciali della ricerca in NLP di questi anni i modelli di *parsing* e il miglioramento delle loro prestazioni. La nozione di *robust parsing* è stata introdotta in questi anni per descrivere una classe di riconoscitori grammaticali che ammettono almeno le seguenti proprietà caratterizzanti:

- la tolleranza per gli aspetti di agrammaticalità spesso esibita da testi ed enunciati linguistici propri della comunicazione reale
- la efficienza computazionale, generalmente ottenuta a spese della espressività grammaticale (cioè la specializzazione dei modelli a sottoinsiemi dei fenomeni grammaticali, e.g. nomi propri)
- la limitata dipendenza da conoscenze lessicali, largamente utilizzate al contrario per la modellazione delle irregolarità grammaticali in approcci tradizionali (ad es. grammatiche sintagmatiche monolitiche, HPSG, LFG)

Tra gli obiettivi di questo modelli, oltre al riconoscimento di strutture grammaticali specifiche (ad es. i nomi propri, o i sintagmi nominali complessi), va sottolineata la loro applicabilità nelle fasi iniziali di progetto di una applicazione di NLP. I parser robusti infatti garantiscono la possibilità di "esplorare" i testi che dovranno essere oggetto della elaborazione finale del sistema. Tali strumenti grammaticali robusti consentono la estrazione "iniziale" di esempi di uso delle parole che, a loro volta, alimentano una serie di modelli induttivi atti a produrre nuova conoscenza (lessicale) e quindi migliorare la qualità della analisi grammaticale successiva.

In questo quadro sono stati studiati modelli robusti per la elaborazione di corpora estesi di testi che garantiscono le caratteristiche proprie dei parser robusti, minimizzando al contempo le dipendenze dall'informazione lessicale disponibile. Un modello basato su grammatiche logiche discontinue è stato inizialmente proposto in [IJ1] per la estrazione di sintagmi verbali, nominali e preposizionali dai testi con una complessità quadratica nell'insieme delle parole. Tale modello è stato poi largamente sperimentato su diverse tipologie di testi

ed infine su lingue diverse (Italiano ed Inglese) nell'ambito di collaborazioni Internazionali (Università di Sheffield, Università di Bruxelles). Le prospettive aperte da questa modellazione all'induzione di conoscenza linguistica da testi scritti sono descritte nella sezione corrispondente.

Un modello più recentemente proposto in [IC9, IC22, IC26, IC28, IC30, IJ10], basato invece sulla conoscenza lessicale (in particolare schemi di sottocategorizzazione dei verbi), combina i principi della stratificazione grammaticale con quelli delle grammatiche logiche discontinue, già esplorate. Il modello risultante è un modello robusto con diverse forme di utilizzazione: fortemente lessicalizzate (quando la conoscenza lessicale è disponibile) e indipendenti dal lessico (nelle prime fasi di esplorazione di un corpus).

Il risultato più recente della ricerca in quest'area è la definizione di un *framework* per il *parsing* che modella l'analisi grammaticale in termini di composizione funzionale (cascata) di algoritmi dedicati, integrando sistematicamente principi grammaticali generali con informazione lessicale specializzata. La sistematicità di questo approccio consente la valutazione indipendente degli algoritmi base della composizione sui propri fenomeni grammaticali specifici. Questo apre una prospettiva nuova nella progettazione di applicazioni NLP, cioè la determinazione per vie analitiche e sperimentali, della migliore configurazione (e.g. composizione funzionale) del *parser* rispetto alle collezioni ed al dominio linguistico di interesse. I risultati principali di questa linea di ricerca sono documentati in [IC26, IC28, IC30, IJ10].

### **Induzione di conoscenza lessicale da corpora di testi**

Una componente rilevante dei modelli di uso della lingua è relativa certamente alla conoscenza lessicale che modella l'uso delle parole di una lingua e che ne determina le interazioni (o i vincoli) a livello sintattico ma soprattutto semantico. La complessità nella progettazione di tale conoscenza linguistica, la sua dimensione (lessici significativi variano da 20,000 a 200,000 entrate), e la sua inerente difficoltà di gestione la rendono una risorsa tra le più costose nell'ambito dei sistemi basati su conoscenza.

Le ricerche in quest'area si sono sviluppate attorno a modelli e metodologie per la generazione semi-automatica di conoscenza lessico-semantica a partire da collezioni estese di testi. I principali risultati hanno condotto alla definizione di algoritmi induttivi originali basati sulla combinazione di approcci simbolico generativi alla analisi morfologica e sintattica dei testi (*parsing* robusto), e di modelli statistici delle regolarità osservate nei testi. Le pubblicazioni più esaustive sulle ricerche in esame sono [IJ2, IJ3, IJ4, IJ5].

Una analisi più dettagliata dei temi relativi a questa linea di ricerca verranno di seguito sintetizzati.

### **Modelli quantitativi per la descrizione linguistica**

In linea con alcuni approcci quantitativi (modelli Bayesiani o catene di Markov) sono stati affrontati fenomeni di ambiguità sintattica e semantica, mediante modelli di tipo probabilistico. Collezioni estese di testi (corpora) come realizzazione di esempi di uso dei lemmi vengono qui suggerite come sorgente di informazione. I metodi di *parsing* robusto risultanti da precedenti ricerche hanno qui dimostrato le caratteristiche di efficienza computazionale e portabilità (tra domini linguistici diversi) garantendo l'estrazione di informazione morfosintattica su larga scala.

Le regolarità strutturali di una lingua che si manifestino a livello lessicale (ad es. le strutture argomentali dei verbi o le proiezioni grammaticali dei nomi, cioè le relazioni sintattiche "imposte" da essi) sono in seguito ottenuti mediante una modellazione statistica delle forme di rappresentazione estratte a livello grammaticale [Bk3, IJ2].



Una indagine ulteriore è stata condotta relativamente alla induzione di regolarità di livello semantico. A tale scopo sono stati definiti formalismi di rappresentazione superficiale della semantica di una lingua (gerarchie di tipi semantici, forme relazionali di rappresentazione del significato di frasi ((Evens,1989) (Miller,1990)). Essi hanno garantito la definizione di modelli specifici per l'apprendimento di componenti semantico-lessicali per il popolamento di basi di conoscenza linguistiche e reti di concetti specifici ad un dominio. Allo studio di questi ultimi aspetti sono dedicate le pubblicazioni [IJ3,IJ4,Bk4,Bk5,Bk8,IC5]. In particolare un lavoro estensivo relativo allo studio della semantica dei verbi dell'italiano è stato condotto in [IJ8].

Tali linee di ricerca hanno portato allo sviluppo di un Sistema per la Acquisizione automatica, denominato ARIOSTO e sviluppato all'interno del gruppo di Intelligenza Artificiale dell'Università di Roma, Tor Vergata, presso il Laboratorio Software del Dipartimento di Informatica, Sistemi e Produzione [IC5,IC6]. Il sistema è in grado di produrre il lessico di un sottolinguaggio applicativo mediante la elaborazione di un corpus di testi afferenti a tale sottolinguaggio (la cui taglia si aggira intorno all'500.000-1.000.000 di parole) nelle due lingue italiano ed inglese. La progettazione del sistema è stata seguita da fasi di sperimentazione e validazione su diversi corpora italiani e uno inglese.

### **Induzione di conoscenza linguistica mediante algoritmi simbolici**

Nel corso di dottorato è stata particolarmente approfondita la problematica dell'apprendimento induttivo da corpora di informazione lessico-semantica relativa ai verbi dell'italiano. Gli approcci dell'apprendimento automatico alle conoscenze di tipo linguistico rappresentano infatti una sfida di particolare importanza per il carattere di grande complessità che quest'ultime presentano. La complessità formale dell'apprendimento grammaticale ed il particolare linguaggio (la lingua naturale) utilizzate per la sorgente dell'informazione sul campione rendono il compito induttivo un problema legato a due fasi:

- la scelta del livello di rappresentazione sia delle osservazioni linguistiche che del risultato dell'apprendimento;
- la scelta per i diversi livelli di rappresentazione dell'algoritmo induttivo migliore, in grado di operare su uno spazio di ricerca molto ampio, come quello che caratterizza le generalizzazioni linguistiche, e pure in presenza di un rumore sorgente non banale (dovuto alla essenziale sottospecificità della conoscenza linguistica che ha generato il campione utilizzato).

Una rassegna dei metodi induttivi utilizzabili ed uno studio sulle loro caratteristiche e limitazioni è discussa nella Tesi di Dottorato, [PhD]. L'approfondimento sperimentale ha condotto ad alcune scelte metodologiche ed alla definizione di un modello simbolico di apprendimento, noto in letteratura come *clustering concettuale* o *unsupervised instance-based learning*, applicato alla derivazione di raggruppamenti lessico-semantici di verbi simili in una lingua. Tali tassonomie verbali sono quindi considerate anche come fondamento della schematizzazione concettuale di un dominio, configurandosi così come conoscenza estratta automaticamente dai testi. Approfondimenti e dettagli sono stati discussi nelle seguenti pubblicazioni [IJ8, Bk4, Bk7, Bk8, IC5, IC6, IC7, IC12, IC13].

### **Induzione di conoscenza linguistica mediante algoritmi statistici**

La disponibilità di corpora sempre più estesi di testi ha permesso la sperimentazione di modelli di induzione su larga scala, di carattere meno complesso dal punto di vista del livello linguistico di rappresentazione richiesto. Algoritmi per la costruzione automatica di schemi di classificazione delle parole basati sull'uso di contesti puri (cioè semplici sequenze di parole di una lingua) sono stati definiti e sperimentati su corpora diversi. L'essenziale livello di rappresentazione

richiesto dai metodi garantisce la loro estrema portabilità ed utilizzabilità di una varietà di domini linguistici, senza un diretto intervento manuale dell'uomo. I metodi e gli esperimenti condotti sono stati discussi in [IJ2,Bk3,Bk6,IC6].

## 10 Riferimenti Bibliografici

- (Chomsky,1965) N. Chomsky, *Aspects of the Theory of Syntax*, MIT Press, 1965.
- (Charniak,1993) E. Charniak, *Statistical Approaches to Natural Language Processing*, MIT Press, 1993.
- (Evens,1989) M. Evens, *Relational Approaches to the Lexicon*, Cambridge University Press, 1989.
- (Miller,1990) George Miller, *Introduction to WordNet an On-line Lexical Database*, International Journal of Lexicography, n. 4, vol. 13, pp. 235-312, 1990.
- (Rumelhart,1986) Rumelhart, McClelland, *Parallel distributed processing*, MIT Press, 1986.
- (Shannon & Weaver,1949) C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, University of Illinois, Press, 1949.
- (Vapnik,1998) Vladimir Vapnik, *Statistical Learning Theory*, John Wiley and Sons, 1998.
- (Zadeh,1965) L.A. Zadeh, *Fuzzy sets*, Information and Control, 8, 338-353, 1965.
- (Zadeh,1975) Zadeh, L.A., *The calculus of fuzzy restrictions*, in "Fuzzy sets and their application to decision process", New York, Academic Press, 1975.