



Outline

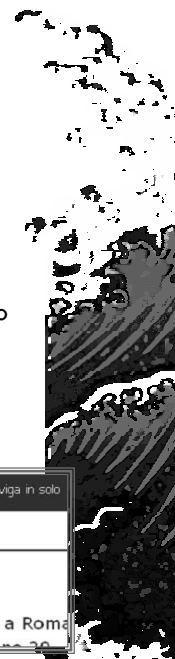
- Metadata Extraction and HLT
- Challenges
- The RTV Experience
- Conclusion

Computer, Lingue ed Applicazioni

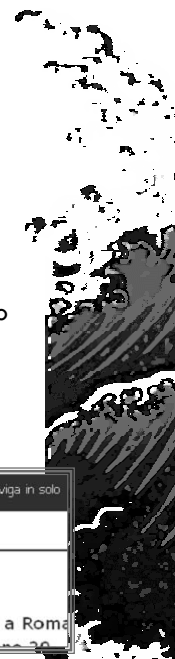
- Perché rendere gli elaboratori capaci di comprendere i contenuti testuali?
 - I testi sono i veicoli principali di significato per molte altre tipologie di dati (e.g. multimediali) ...
 - e per la definizione, trasmissione e condivisione di conoscenze
 - Ricerca di informazione

Elaborazione come processo di *interpretazione*

- Elaborare un testo in tale ambito significa interpretarne aspetti *rilevanti* del significato
 - Area tematica (e.g. cronaca/politica)
 - Obbiettivi (e.g. virus/spam nell'e-mail)
 - Personaggi e luoghi coinvolti
 - Eventi dichiarati (e.g. news)
 - Obbiettivi comunicativi (e.g. dialogo e pianificazione)
- Risultato: *rappresentazione esplicita del significato testuale* ...
 - che attiva alcune *inferenze* (e.g. *rilevanza*)



-



Notizie di Agenzia (3): Multilingualità

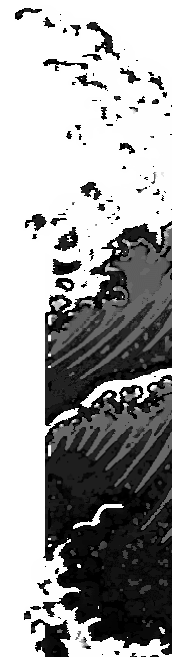


Alcune Riflessioni

- 📖 La *comprensione* di informazione linguistica richiede conoscenza riguardo:
 - 📖 La lingua (e.g. *sintassi*)
 - 📖 Il mondo (e.g. *rugby, squadre e nazioni*)
 - 📖 Come la prima fa riferimento al secondo
- 📖 L'accesso e la pubblicazione (elettronica) "*intelligente*" implica conoscenze riguardo:
 - 📖 L'obiettivo, i.e. ricerca
 - 📖 Il mondo in cui la comunicazione e' immersa
 - 📖 I produttori vs. gli utenti del testo

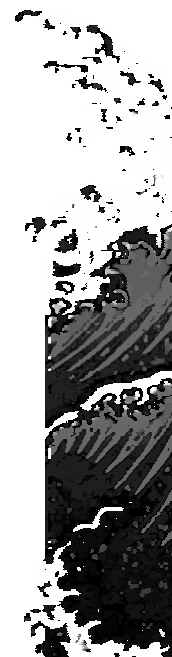
Sfide e Ricerca

- ▣ *Accuratezza* del riconoscimento/produzione
- ▣ *Robustezza* (errori/rumore/incompletezza)
- ▣ *Scala*
 - ▣ Copertura di Lessici e Grammatiche
- ▣ *Espressività*
 - ▣ Dizionari, Lessici e *Thesaura*
 - ▣ Modelli del mondo ed inferenza
- ▣ *Flessibilità*
 - ▣ Lingua del produttore vs. consumatore
- ▣ *Naturalzza*



HLT: Ricerca e Risultati

- ▣ *Risorse e Standard*
 - ▣ Dizionari e Lessici
 - ▣ Corpora
- ▣ *Riconoscimento:*
 - ▣ Morfologia ed Analisi a Stati Finiti
 - ▣ Riconoscimento Sintattico e Disambiguazione
 - ▣ Riconoscimento e Classificazione dei Nomi Propri (e.g. luoghi o persone, >93% acc.)
 - ▣ Disambiguazione del Senso (≈90% acc.)
 - ▣ Riconoscimento Eventi (MUC Conferences, 87-98)



IE in MUC

MUC Named Entity Task: Tagging of Person, Location, Organization

<CORP>General Electric Co.</CORP> of the
</GEO> and <GEO>Taiwan </GEO>'s <CORP>MITAC
International Corp. </CORP> announced a joint venture to
produce advanced electronics and information-processing
equipment for defense use in<GEO>Taiwan </GEO>.

The partners will have equal stakes in the new company,
dubbed <CORP> GETAC </CORP>. <CORP> GE </CORP> will
take charge of technology transfer and foreign sales, while
<CORP> MITAC </CORP> will be responsible for local
management and sales, said <PERSON> Matthew Miao
</PERSON>,<CORP> MITAC Group </CORP>chairman.

Identifies basic building
blocks of the text

IE in MUC

Coreference

General Electric Co. of the U.S and Taiwan's MITAC
International Corp. announced a joint venture to produce
advanced electronics and information-processing equipment
for defense use in

The partners will have equal stakes in the new company,
dubbed GETAC. GE will take charge of technology transfer
and foreign sales, while MITAC will be responsible for local
management and sales, said Matthew Miao, MITAC Group
chairman.

MUC-6 coreferring entities
connected by solid lines;
includes name coreference,
definite reference,
pronouns

HLT: Ricerca e Risultati (2)

■ Apprendimento di conoscenza linguistica

- Analisi sintattica per esempi
(e.g. *parsing* statistico, HMM)
- Acquisizione automatica di terminologia
(e.g. *broadband communication* o *imposte dirette*)
- Regole di disambiguazione del senso (WSD) rispetto a dizionari semantici
 - *battere/sport* vs. *battere/colpire*
 - Wordnet (Miller et al, 91) ed i corpora
- Information Extraction Rules
 - Apprendimento di schemi frasali
(`[companies] acquire [companies]`)

HLT @ DISP, Tor Vergata

The AI-NLP group at Computer Science Dept

- Prof. Maria Teresa PAZIENZA
- Prof. Roberto BASILI
- 2-4 Researchers
- 1-2 PhD students
- 1-2 Visiting scientists

- URL: `ai-nlp.info.uniroma2.it`

HLT @ Rome, Tor Vergata

Aree di Ricerca:

Ingegneria delle Lingue

-  Metodi robusti per il TAL
-  Modelli Computazionali del Lessico e Disambiguazione del Senso

Apprendimento Automatico per il TAL su larga scala

Applicazioni del TAL:

-  Information Extraction e Retrieval (Categorizzazione)
-  Question Answering

Ingegneria della Conoscenza Ontologica mediata linguisticamente

HLT @ DISP, Tor Vergata

Sistemi

-  **ARIOSTO ('92):** Acquisizione Automatica di conoscenza lessicale
-  **CHAOS ('98.02):** Analizzatore sintattico (Italiano, Inglese)
-  **RGL ('97):** Analisi formale dei concetti per l'acquisizione di schemi di sottocategorizzazione verbale (Reticoli di GALOIS)
-  **GoDoT ('98):** Disambiguazione Semantica
-  **SATOR ('00):** Apprendimento Automatico di schemi per IE
-  **ONTOLOAD ('01):** Acquisizione di ontologie di dominio a partire dai testi

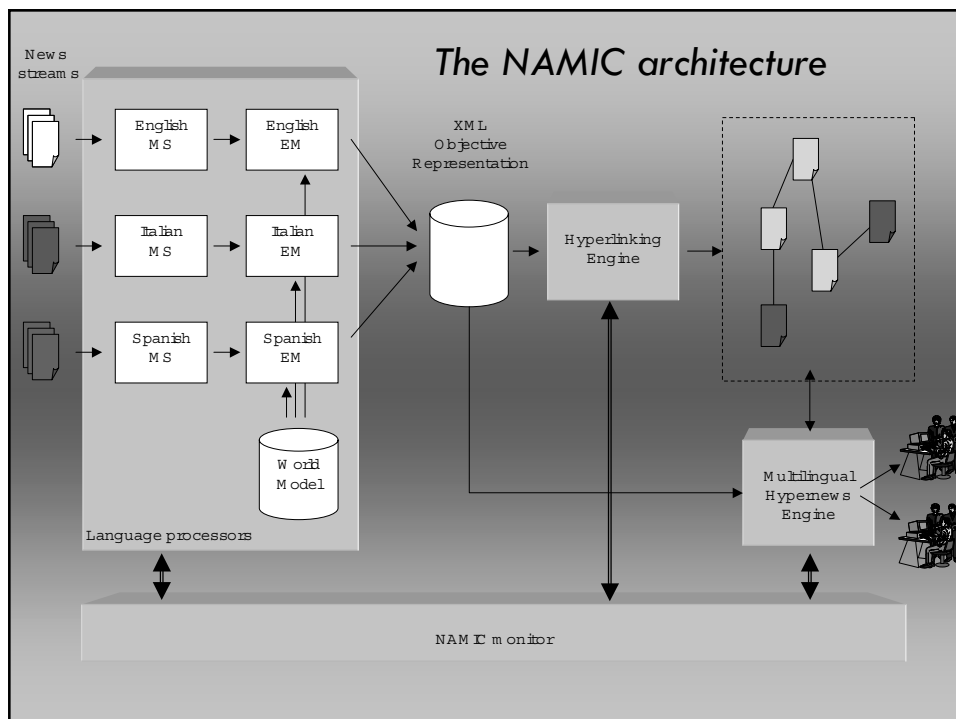
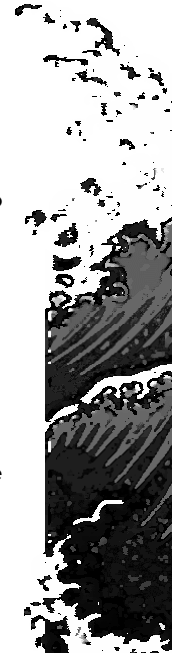
-



HLT @ Rome, Tor Vergata

Progetti Internazionali (EU Esprit, 5-6 FW)

- **ECRAN (97-98)** Apprendimento lessicale per IE adattivo
- **TREVI (99-00)** Categorizzazione e personalizzazione basata sul testo
- **NAMIC (00-02)** Hyperlinking multilinguale su flussi di notizie di agenzia
- **MOSES (02-04)** Question Answering basato su ontologie
- **PrestoSpace (04-)** Indicizzazione ed Interrogazione Semantica di dati multimediali (RAI)



Un es. *Rassegna Stampa Automatica*

■ Obbiettivi

- Ricerca autonoma di notizie in tempo reale
- Classificazione delle notizie
- Estrazione di informazioni puntuali (*dichiarazioni*)

■ Integrazione di due tecnologie diverse

- Ricerca Autonoma su Web (*Web spider*)
- TAL per il trattamento automatico dei Testi

■ Cooperazione tra gruppi HLT e DIST del Dipartimento Informatica, Sistemi e Produzione

Ricerca Autonoma su Web (*Web spider*)

■ Gruppo DIST (*Distributed Information System*), prof. M. Angelaccio, B. Buttarazzi

■ Metodologie:

- *Web Searching & Visualization*
- *Web-based Workflow Engines*

■ Risultati

- VSEARCH, Web local searching tool
- WISH, Wireless VSEARCH con Caching Semantico

Rassegna Stampa Automatica (3)

Ruolo delle due tecnologie diverse

Ricerca Autonoma su Web (Web spider)

-  Ricerca siti interessanti (per es. giornali su Web)
-  Gestisce accesso e richiamo di pagine remote (ad es. cronaca, politica, sport)

Trattamento automatico dei Testi

-  Riconosce fenomeni di base nei testi
-  Classifica *estratti* dei testi giornalistici
-  Riconosce persone/agenti e loro dichiarazioni
-  Estrae tutte le informazioni “scoperte” aggiornando un archivio locale

Rassegna Stampa Automatica (3)

Demo (in coll. con F. Perrazzoni, F. Di Cola)

Funzionalità:

-  Ricerca e *Download* di Pagine Web (*Corriere della Sera*)
-  Selezione delle pagine e dei personaggi di interesse
-  Estrazione linguistica delle informazioni (GATE+CHAOS)
-  Navigazione attraverso le citazioni/dichiarazioni estratte