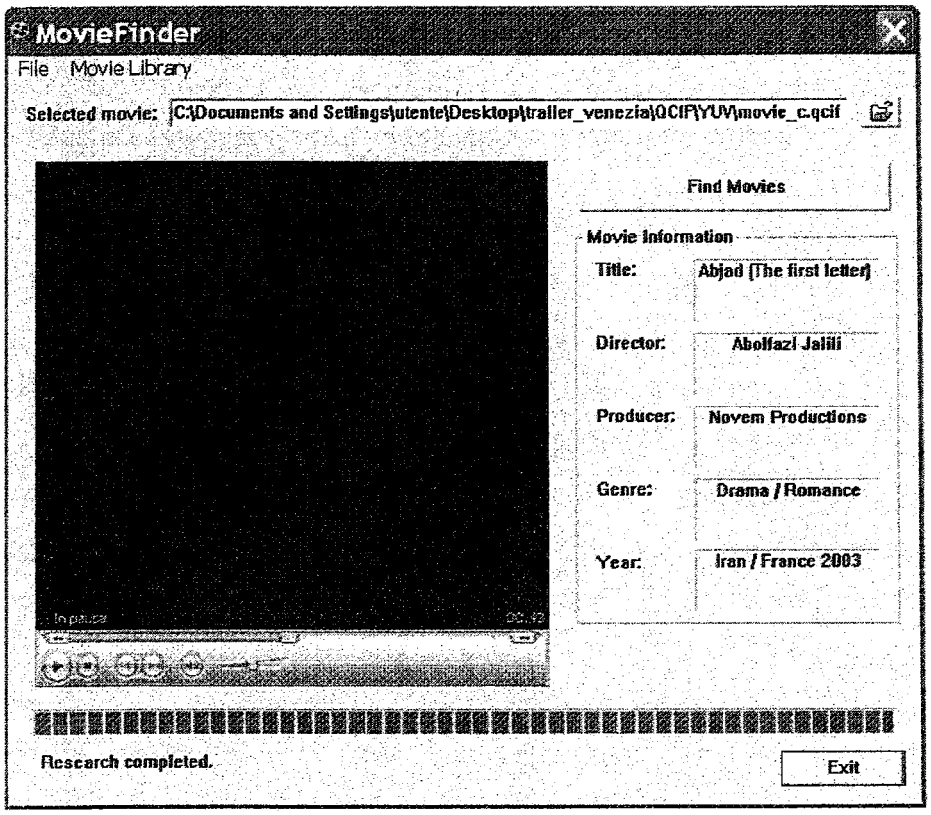

Figura 2 – Interfaccia grafica per il riconoscimento di film

3.5.2 Sistemi avanzati di recupero delle informazioni per media e domini eterogenei (X-Search)

Negli ultimi anni si è accentuato il divario fra la quantità di informazioni disponibili in forma digitale e l'efficacia degli strumenti per il loro reperimento. Da una parte abbiamo mezzi sempre più estesi per la raccolta, la pubblicazione e la diffusione delle informazioni (come wiki, blogs, news, archives); dall'altra abbiamo strumenti ancora insoddisfacenti per selezionare automaticamente le informazioni effettivamente utili e analizzarle. Questo divario ha importanti conseguenze sociali ed economiche perchè non consente di scoprire informazioni critiche che sono teoricamente disponibili.

I motori di ricerca per internet e intranet rappresentano fedelmente la situazione. Oggi, con decine di migliaia di computer in parallelo, si riescono a censire miliardi di pagine web, a seguirne gli aggiornamenti in modo sempre più puntuale e a rispondere a migliaia di interrogazioni simultaneamente. All'accrescimento delle prestazioni hardware e della larghezza di banda non è pero' corrisposto un sostanziale miglioramento delle tecnologie di elaborazione dell'informazione, se è vero che spesso non troviamo quello che cerchiamo e che anche per interrogazioni relativamente semplici il sistema restituisce molti risultati ridondanti o palesemente sbagliati.

Un problema intrinseco dei sistemi convenzionali per il reperimento delle informazioni è che essi si basano ancora sul meccanismo delle parole chiave, per cui un documento viene recuperato soltanto quando contiene esattamente i termini adoperati nell'interrogazione. Altri aspetti insoddisfacenti sono legati alla modalita' prevalente di presentazione dei risultati, che richiedono la scansione di una lista testuale analitica, e all'incapacita' di utilizzare eventuali informazioni sulla semantica e la struttura dei documenti per migliorare la precisione dei risultati. Negli ultimi anni, in particolare dopo l'avvento di Google, non ci sono state da questo punto di vista innovazioni decisive nella tecnologia dei sistemi per il reperimento delle informazioni. Al contempo però, si sono affermati nuovi media e modelli inediti di comunicazione e fruizione delle informazioni, sono state sviluppate tecniche per la descrizione delle informazioni finalmente semantiche ed e' fiorito il mercato dei dispositivi mobili, caratterizzati da interfacce utente assolutamente specifiche. Queste trasformazioni non hanno avuto un adeguato corrispettivo nell'evoluzione degli strumenti di ricerca. Alle sfide tecnologiche non risolte si somma un preoccupante fenomeno di concentrazione industriale, con pochissime aziende (Google, Yahoo! e Microsoft) che hanno monopolizzato il mercato mondiale dell'accesso alle informazioni di rete. Quello a cui stiamo assistendo è un fenomeno di *digital divide* di secondo livello: non esiste più solo una distanza tra i cittadini che sanno fare un uso regolare ed efficace degli strumenti avanzati per l'accesso alle informazioni e quelli che ne sono esclusi, ma anche un divario industriale tra i paesi, quali gli Stati Uniti, che investono nella eleborazione della conoscenza, e gli altri paesi che corrono il rischio di essere esclusi da questo fondamentale sviluppo tecnologico.

L'obiettivo generale di questa attività è lo studio, progettazione e realizzazione sperimentale di sistemi avanzati per l'accesso alle informazioni, in particolare per media e domini eterogenei. Verranno considerate tre specifici temi che verranno poi di seguito descritti:

Web semantico: documenti in formato RDF e XML

Dispositivi mobili

Motori di ricerca operanti in modalità intranet e enterprise search, o internet per porzioni più limitate del web (ad esempio domini specifici, quali i blogs)

3.5.2.1 Motori di ricerca per il web semantico

Nella visione del Web Semantico, le pagine vengono arricchite con annotazioni interpretabili dal computer che catturano il significato del contenuto delle pagine stesse. I vantaggi per l'accesso alle informazioni sono evidenti, in linea di principio. Utilizzando linguaggi come RDF per descrivere il contenuto dei dati e ontologie formali per specificare concetti e regole di derivazione, è possibile trovare risposte che sono collegate *logicamente* alle interrogazioni, superando i limiti sintattici dei

motori di ricerca tradizionali.

Attualmente i marcatori semantici non vengono utilizzati dai motori di ricerca nel calcolo dei risultati. Poiché è probabile che vedremo un numero crescente di pagine descritte in modo misto (testo + annotazioni semantiche), il trattamento esplicito di entrambi gli aspetti, con l'integrazione di reperimento testuale ed inferenza logica, rappresenta una sfida ma anche una opportunità per gli attuali motori di ricerca.

Una situazione di complessità intermedia è costituita dalla gestione delle pagine descritte in XML, in cui lo schema, la tipizzazione dei dati e la prossimità strutturale possono essere adoperate a fini semantici. La strutturazione può consentire ad esempio di reperire l'elemento informativo di granularità appropriata, che risponda all'interrogazione in modo esaustivo ma che sia anche sufficientemente specifico. Il vantaggio di considerare pagine XML è anche che si tratta di uno standard di descrizione estremamente diffuso per la pubblicazione e lo scambio dei dati, anche multimediali.

Il reperimento di informazioni da dati XML è l'oggetto di INEX (Initiative for the Evaluation of XML retrieval), un programma pluriennale di ricerca finanziato in parte dalla Comunità Europea. L'obiettivo di INEX è lo sviluppo di tecniche di reperimento delle informazioni che integrino contenuto e struttura, cercando di fondere i metodi di "information retrieval" coi linguaggi di interrogazione per basi di dati strutturate. INEX è diventato il principale forum mondiale del settore; alla edizione 2006 (http: //inex.is.informatik.uni-duisburg.de/2006/) hanno partecipato una settantina di università e laboratori di ricerca industriali. Per la sperimentazione e valutazione dei prototipi, da quest'anno viene utilizzata un sottoinsieme della collezione Wikipedia (http: //en.wikipedia.org) descritto in linguaggio XML.

Questa attività intende quindi contribuire allo sviluppo di sistemi per il reperimento intelligente di informazioni quando nei dati sono presenti marcatori semantici e, in particolare, quando sono descritti in XML. L'attività, che verrà svolta anche partecipando direttamente ad INEX, prevede tre obiettivi.

A.1. Definizione e implementazione di un prototipo innovativo per "XML retrieval".

A.2. Sperimentazione del prototipo nel contesto di INEX 2006 e INEX 2007.

A.3. Estensione del prototipo per interrogare in tempo reale la collezione Wikipedia.

Si noti che la fase.3. rappresenta un obiettivo applicativo che potrebbe avere ricadute significative, considerata l'importanza planetaria che ha assunto detta collezione e la relativa efficacia dei sistemi che sono attualmente disponibili per cercare le informazioni che essa contiene.

È stata definita e implementata una prima versione di un modello innovativo per "structured information retrieval". Il modello consente di utilizzare sia interrogazioni strutturate che non

strutturate e di reperire specifici elementi pertinenti all'interrogazione all'interno di una base documentaria descritta in linguaggio XML.

Si prevede quindi che sia l'interrogazione dell'utente sia i documenti oggetto di ricerca possano essere logicamente strutturati in elementi con differente grado di specificità e il modello è basato sull'idea di stimare in modo differente le probabilità di occorrenza delle parole (nella interrogazione e nei documenti) a seconda del grado di specificità degli elementi in cui esse compaiono.

Mentre gli elementi teorici del modello sono stati definiti, il problema della sua complessità di calcolo non è stato ancora completamente risolto. Una versione approssimata del modello è stata implementata e sperimentata all'interno di INEX 2006 (http: //inex.is.informatik.uni-duisburg.de/2006) utilizzando la collezione Wikipedia.

Questa attività proseguirà con la definizione di un metodo di calcolo più efficiente e con la implementazione /sperimentazione del modello non approssimato.

L'attività viene svolta in collaborazione con FAO.

3.5.2.2 Ricerca delle informazioni da dispositivo mobile

Una serie di studi e analisi di mercato molto recenti indicano che siamo alla vigilia dell'esplosione di un nuovo mercato, quello della ricerca delle informazioni da dispositivo mobile. Questo fenomeno si sta affermando a dispetto delle oggettive difficoltà riscontrate dagli utenti che adoperano palmari o cellulari di ultima generazione. Su questi dispositivi, infatti, la ricerca di informazioni può rivelarsi particolarmente difficoltosa, noiosa e dispendiosa a causa delle caratteristiche intrinseche del mezzo; schermi piccoli, capacità di immissione dati limitata, banda di connessione stretta.

Un approccio che è stato studiato in FUB per superare questo problema consiste nell'utilizzare un motore di ricerca a categorie, che cioè raggruppa i risultati in una gerarchia di categorie specifica a ciascuna interrogazione e consente all'utente di accedere direttamente ai risultati associati a ciascuna categoria. Il vantaggio è che non c'è bisogno di scorrere lunghe pagine di risultati e di digitare nuovi termini per raffinare l'interrogazione, che specialmente per un dispositivo mobile sono due operazioni critiche.

I motori di ricerca a categorie sono una realtà relativamente nota per le ricerche effettuate mediante computer da tavolo – un esempio è CREDO (http: //credo.fub.it), il sistema sviluppato in FUB - ma non sono ancora stati impiegati sui dispositivi mobili. Il solo sistema di questo tipo è Credino (http: //credino.dimi.uniud.it), nato da una collaborazione tra FUB e Università di Udine, che utilizza i risultati prodotti da CREDO e li trasforma per essere visualizzati su un palmare.

L'attività in oggetto intende proseguire questo promettente filone di ricerca con tre obiettivi

principali.

- Miglioramento delle prestazioni di *Credino*, relativamente a tempi di risposta, efficacia delle categorie e usabilità dell'interfaccia.
- Valutazione delle prestazioni di *Credino* rispetto a motori di ricerca a categorie alternativi e a motori di ricerca convenzionali.
- Sviluppo e valutazione di una versione di CREDO/ *Credino* per telefoni cellulari.

È stata sviluppata una versione più efficiente e portabile di *Credino*, che ora potrà essere utilizzato da un maggior numero di dispositivi mobili con le caratteristiche di un palmare (Personal Digital Assistant).

In secondo luogo è stata condotta una sperimentazione pilota con utenti per valutare se nella ricerca di informazioni Internet tramite palmare sia più conveniente utilizzare un motore di ricerca a categorie (*Credino*) oppure la versione mobile dei normali motori di ricerca (Google Mobile). La ricerca con Credino si è rivelata effettivamente più precisa e veloce. Gli esperimenti hanno inoltre evidenziato che utilizzando invece un normale computer da tavolo il motore di ricerca a categorie non è necessariamente migliore del motore di ricerca convenzionale.

Questa attività proseguirà sviluppando una versione per cellulare di CREDO e valutandone l'efficacia rispetto a sistemi convenzionali.

L'attività viene svolta in collaborazione con la Università di Udine e con il Poznan Supercomputing and Networking Center. Contemporaneamente, sono stati avviati contatti per stabilire una partnership con una azienda del settore per la fornitura di un servizio commerciale basato sulla metodologia sviluppata.

3.5.2.3 Motori di ricerca operanti in modalità intranet, enterprise e internet search

Oggi, imprese di dimensione medie e grandi investono nella tecnologia *Digital asset management (DAM)*, anche chiamata "desktop search", "enterprise search" o "digital asset warehousing, " DAM si riferisce alla costituzione di un archivio centralizzato per tutti i tipi di contenuto digitale. I sistemi DAM contengono tipicamente brevi descrizioni del contenuto informativo del documento, in formato XML, mediante i quali i dati vengono indicizzati in formato eventualmente compresso utile ad un recupero veloce dell'informazione. La ricerca ha stabilito che il ritorno di investimento in queste nuove tecnologie va dalle 8 alle 14 volte il capitale investito.

L'informazione di ogni organizzazione è oggi distribuita su diversi servers e desktop computers. L'attività in oggetto intende fornire le funzionalità di base per un servizio avanzato di ricerca con

archivio centralizzato. L'obiettivo è quello di indicizzare, ricercare, estrarre e presentare in forma avanzata l'informazione in modalità ``desktop search''. Le due fasi principali sono le seguenti.

Costruzione di un prototipo desktop search basato su l'open source Terrier (http: //ir.dcs.gla.ac.uk/terrier/about.html) per indicizzazione e recupero di documenti eterogenei (excel, word, pdf, testo, codici programma, ecc.)

Costruzione di un servlet/API per l'uso in rete (Intranet search) e integrazione con il prototipo basato su tecnologia open source mysql, php già utilizzato in FUB nei progetti Agire Digitale, Octofub e Hermes (per esempio http: //logistica.fub.it/homepage.php).

Il weblog, o più brevemente il blog, è una tecnologia che permette la pubblicazione online di opinioni e articoli. La facilità di produzione e gestione degli articoli in un weblog ha reso estremamente facile la possibilità di pubblicazione di opinioni da parte di un vasto pubblico, inclusa la parte meno istruita della popolazione. Gli articoli non sono in generale filtrati da parte di moderatori, come invece accade nei forum di discussione o nei wiki.

Come ogni altro media i blogs hanno un argomento particolare di discussione, dalla politica alle notizie, e sono diventati importanti luoghi virtuali di espressione e formazione dell'opinione pubblica.

L'avvento dei blogs ha sollevato una serie di problemi di carattere legale, dalla diffamazione alla diffusione di informazioni personali e confidenziali, delle quali gli Internet Service Providers non sono in generale responsabili in quanto informazione pubblicata da terze parti (direttiva UE 2000/31/EC).

Per l'importanza dei blogs, il NIST (National Institute for Standard and Technology) ha istituito quest'anno una sessione di valutazione dei blog alla TEXT RETRIEVAL CONFERENCE (TREC) (http://trec.nist.gov/call06.html). La sessione TREC prevede la valutazione dei sistemi relativamente ad un insieme di argomenti, notizie e eventi significativi dell'ultimo anno contenuti in una collezione di 135GB di documenti dei blogs.

L'attività consiste nell'acquisizione, analisi e valutazione delle tecniche per il recupero dell'informazione dai blogs.

Sviluppo di un motore di ricerca distribuito

Lo sviluppo del sistema di recupero distribuito è stato sviluppato in collaborazione con l'università di Tor Vergata e la prima versione è ancora in modalità batch. È basato sul motore di ricerca open source Terrier. L'architettura è costituita da un broker centrale che riceve una lista di interrogazioni e la trasmette ad un cluster di macchine. Ciascuna macchina gestisce una sottocollezione di circa 20GB di documenti e a fronte di ogni interrogazione recupera sugli indici locali tutti i documenti

che soddisfano l'interrogazione. I documenti recuperati vengono forniti al broker centrale che restituisce all'utente l'insieme ordinato dei documenti rilevanti.

Il sistema è incrementale rispetto al numero di macchine, ovvero il sistema è stato ideato per lavorare con un numero qualunque di macchine. Il tempo ottimale di recupero viene stabilito in base alla dimensione degli indici locali e il numero di macchine su cui viene distribuita la collezione.

Tecniche di estrazione di informazione

Oltre a presentare i documenti in cui l'informazione è presente, i sistemi avanzati di estrazione dell'informazione devono anche fornire direttamente le porzioni di documento in cui l'informazione cercata è presente. Ad esempio all'interrogazione "Quali sono gli aspetti negativi della tecnologia wi-fi" il sistema di recupero può fornire la risposta "...aspetti negativi come una portata limitata, e una intrinseca debolezza sul versante della sicurezza".

È stata definita una tecnica di pesatura che non dipende dal contenuto e dalla dimensione della collezione. È un sistema che non fa uso uso di parametri che richiederebbero altrimenti un processo costoso ed elaborato di "personalizzazione" che richiede il supporto di chi rilascia il servizio. La Fondazione Ugo Bordoni ha fornito il primo metodo di recupero di informazione che non richiede una fase preliminare di apprendimento e le cui prestazioni sono competitive con i metodi che si basano sull'ottimizzazione dei valori dei parametri.

3.5.3 Ventinovesima Conferenza Europea di Information Retrieval ECIR 2007

ECIR è la maggiore conferenza europea e fra le più importanti al mondo nel campo del reperimento delle informazioni. Vengono trattati tutti i principali metodi di ricerca delle informazioni (interrogazione, filtraggio, classificazione, navigazione) su vari tipi di dati (testi, pagine Web, audio e video, posta elettronica), sia per Internet che per Intranet.

La Fondazione Ugo Bordoni si è aggiudicata la presidenza scientifica e l'organizzazione di 29th European Conference on Information Retrieval (ECIR 2007), Roma, 2-5 Aprile 2007, quale riconoscimento dell'eccellenza raggiunta in questo settore.

Ai fini della preparazione di questa Conferenza nel 2006 sono state svolte le seguenti attività:

- Costituzione del Comitato di Programma, al quale hanno aderito circa 150 esperti di tutto il mondo.
- Installazione e gestione del programma per l'acquisizione degli articoli scientifici e successivamente delle loro recensioni da parte dei membri del Comitato di Programma

- Sono stati sottoposti alla conferenza 292 articoli, mentre l'attività di recensione è consistita di circa 900 recensioni totali.
- Acquisizione degli sponsor e dei patrocini, in particolare il Ministero delle Comunicazioni e il Comune di Roma

Il livello scientifico della conferenza sarà molto elevato perché dei quasi 300 articoli che sono stati sottoposti a ECIR 2007 ne sono stati selezionati in tutto 82, e solo 42 per la presentazione orale alla conferenza (gli atti saranno pubblicati da Springer-Verlag). Gli autori provengono dalle principali università, aziende e centri di ricerca attivi nel settore, con una forte presenza extra-europea.
Le attività svolte si possono consultare sul sito http: //ecir2007.fub.it

3.5.4 Prosecuzione dell'iniziativa sul Trattamento Automatico della Lingua Italiana

Il Trattamento Automatico della Lingua (TAL) è un tema di ricerca di particolare interesse per il Ministero delle Comunicazioni e per la Fondazione Ugo Bordoni sia per l'importanza che la lingua riveste nell'ambito della cultura, sia per la stretta connessione che il trattamento del linguaggio parlato e scritto ha con i servizi fruibili attraverso i sistemi informatici di nuova generazione, quali la larga banda e i diversi media digitali. Le tecnologie TAL contribuiscono a diffondere la lingua italiana e sono elemento chiave per le iniziative di e-democracy.
Il Ministero delle Comunicazioni ha istituito nel 2003 un Forum Permanente sul TAL allo scopo di promuovere iniziative di ricerca e sviluppo nell'ambito di questa tematica.
Oltre all'organizzazione di numerose occasioni di incontro, è stato realizzato un Libro Bianco sul TAL ed è stato realizzato un sito web (www.forumtal.it).
La prosecuzione delle attività avviate dal Forum vanno nella direzione di un sempre maggiore coinvolgimento di enti e persone nell'attività del forum, nella diffusione delle tecnologie TAL attraverso mezzi tradizionali quali il bollettino e le conferenze, e più innovative utilizzando opportunamente la rete. Anche la formazione nell'area del TAL e la diffusione via rete di strumenti TAL di cui si possa provare tangibilmente il beneficio sono oggetto dell'attività.
Infine si ritiene opportuno valutare la necessità di allargare all'ambito europeo l'attività del TAL e la possibilità di promuovere nuove iniziative progettuali per la lingua italiana.
Un primo obiettivo è stato ed è il coordinamento delle riunioni del Forum, nell'ambito del Ministero delle Comunicazioni, assicurando il necessario supporto allo svolgimento dei lavori. Le attività del Forum verranno potenziate, con il coinvolgimento di nuove realtà e la definizione di nuove iniziative.

Tra gli obiettivi del ForumTAL vi è certamente la formazione, e conseguentemente verrà promossa, in coordinamento con gli istituti universitari, l'offerta formativa, anche con l'attivazione di borse di studio e con attività di informazione circa le prospettive occupazionali.

Per assicurare una migliore diffusione delle iniziative del Forum sarà potenziato il portale del forum e creato un bollettino informativo di periodicità bimestrale.

Sarà anche valutata la opportunità di organizzare in modo più efficiente ed efficace lo scambio di dati e di programmi di elaborazione nell'ambito del TAL con l'individuazione di un Istituto a ciò preposto.

Verrà altresì coordinato ed integrato il sottoprogetto "l'italiano nelle comunicazioni digitali" da svolgere in collaborazione con il Ministero delle Comunicazioni.

L'attività svolta nel periodo in oggetto riguarda una seconda fase per la quale gli obiettivi primari sono la prosecuzione delle attività informative, attraverso le riunioni del forum, l'aggiornamento del sito Web, la creazione di un bollettino informativo distribuito per via internet con cadenza bimestrale.

Altro obiettivo rilevante è rappresentato dalla organizzazione della terza **conferenza nazionale** sul tema del TAL caratterizzata da una spiccata presenza degli utilizzatori della tecnologia. La formula proposta prevede la presenza di tutte le realtà operative nel settore privilegiando l'esposizione per i costruttori di TAL e gli interventi alla conferenza per gli utenti in modo da evidenziare la concretezza e la maturità dei prodotti TAL.

Il prossimo obiettivo del ForumTAL sarà quindi diffondere, anche attraverso il libro bianco le tecnologie della lingua, evidenziando i contributi di queste tecnologie possono fornire agli uffici, al legislatore, ai ministeri, a tutti coloro che utilizzano la lingua come principale mezzo di lavoro.

Nel periodo gennaio – dicembre 2006 sono proseguite le riunioni del forum (23 gennaio, 29 marzo, 7 giugno, 18 ottobre, 7 dicembre) con i seguenti obiettivi organizzativi:

organizzazione della Conferenza TAL 2006 Uomini e macchine, un colloquio possibile;

predisposizione degli atti della Conferenza;

potenziamento del sito con l'introduzione di aree di maggior richiamo come l'area news, la sezione dedicata al bollettino, la sezione dedicata al mantenimento delle informazioni su soggetti e prodotti nazionali; sono in preparazione la raccolta delle statistiche di visita, un'area riservata per servizi ai membri del ForumTAL e un'area per la diffusione delle Risorse Linguistiche Nazionali;

aggiornamento del libro Bianco sul TAL nella versione in rete;

collaborazione con la REI (Rete di Eccellenza dell' Italiano Istituzionale) nella realizzazione del loro sito al fine di creare sinergia con il sito del TAL;

partecipazione alla Conferenza TALeP e all'iniziativa ESFRI;

organizzazione della Conferenza internazionale LT forum.

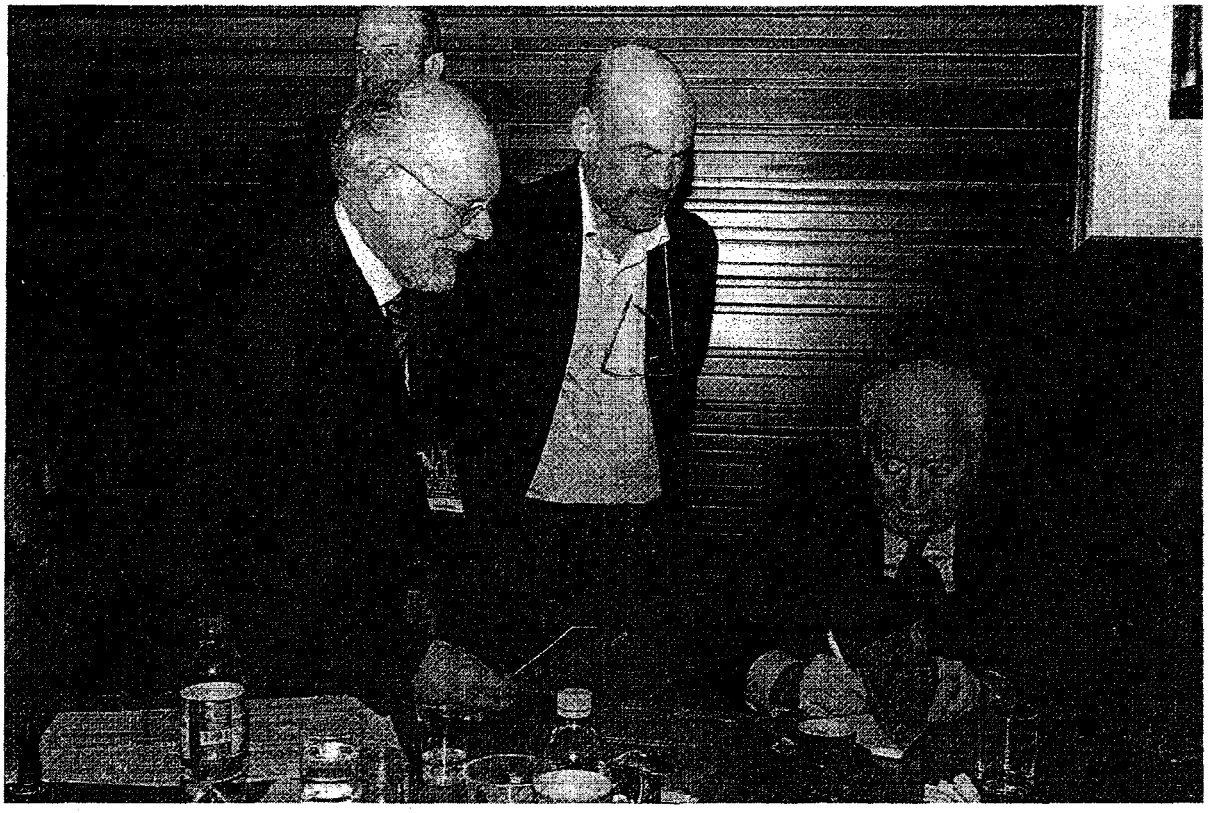

Fig. 3: Un momento della Conferenza TAL 2006 Uomini e macchine, un colloquio possibile

La conferenza *TAL 2006 Uomini e macchine, un colloquio possibile,* svoltasi sotto l'alto patronato del Presidente della Repubblica, ha visto la partecipazione di oltre 800 persone, 20 espositori e 45 conferenzieri. Si ritiene importante far notare che la presenza tra gli espositori di enti di ricerca, industrie e associazioni, ha favorito uno scambio di conoscenze tecnologiche di particolare valore. I conferenzieri i erano tutti di elevato livello: è tuttavia opportuno citare la presenza tra essi del Presidente Emerito della Repubblica Francesco Cossiga.

L'obiettivo principale della Conferenza era estendere la consapevolezza che il Trattamento Automatico della Lingua è una tecnologia matura: gli oggetti visibili negli stand, gli esempi forniti dai conferenzieri, il dibattito nelle due Tavole Rotonde riteniamo abbiano fornito una chiara conferma.

3.5.5 IDEM_2006

Nell' ambito del tema del Trattamento Automatico del Linguaggio (TAL) la Fondazione svolge da tempo, in collaborazione con l'arma dei Carabinieri, attività di ricerca sul tema dell'identificazione del parlante.

Il SW IDEM per il riconoscimento del parlante a scopo forense è stato oggetto di una fornitura al RaCIS (Raggruppamento Carabinieri Investigazioni Scientifiche) nel 1990 ed è attualmente oggetto di un contratto annuale di manutenzione che ne prevede l'aggiornamento e l'adeguamento al sistema operativo disponibile. Offerta analoga è stata inoltrata al RUD (Raggruppamento Unità Difesa). Lo stesso SW viene richiesto annualmente da alcune Università e Istituti CNR.

Obiettivi del progetto sono l'aggiornamento del SW ai nuovi sistemi operativi resi disponibili e lo studio di nuovi algoritmi atti a migliorare il sistema nell'ottica applicativa di tipo forense.

Un secondo importante obiettivo è la valutazione delle prestazioni del sistema su lingue diverse dall'italiano.

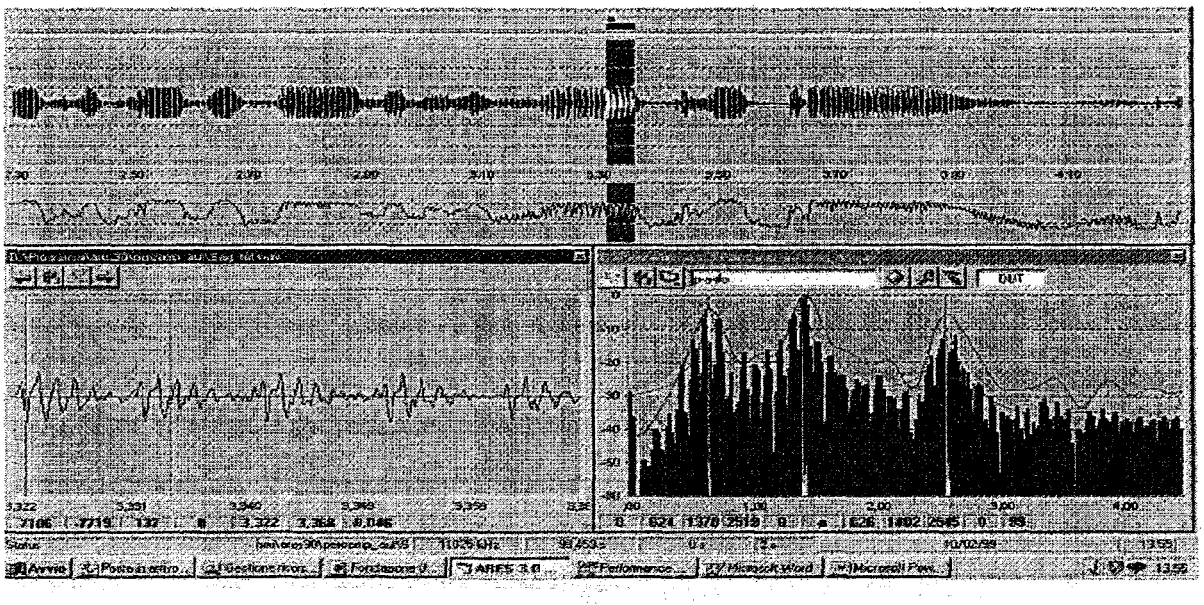

Fig. 4: diagrammi segnali vocali

Nell'ambito della fonetica forense si è anche approfondito il tema della trascrizione di segnali scarsamente comprensibili.

Relativamente al tema del riconoscimento del parlante nel corso dell'anno si è provveduto ad eseguire alcuni test sul corretto funzionamento del sistema: un primo test in lingua brasiliana (portoghese parlato in Brasile) ed un secondo test in lingua araba. Entrambi i test hanno dimostrato il corretto funzionamento del sistema con le altre lingue, fermo restando la necessità, per avere dati statisticamente significativi, di disporre di un campione di riferimento sufficientemente vasto (almeno 30 diversi soggetti) della popolazione di riferimento.

Successivamente è stato fatto un test su segnale di qualità nell'ambito della collaborazione con il RUD (Raggruppamento Unità Difesa).

Relativamente agli aspetti della trascrizione, a seguito di riunioni informali e di accadimenti processuali, che hanno evidenziato le numerose difficoltà legate alla trascrizione di un segnale audio, in ispecie quando si tratti di parlato spontaneo come nel caso delle intercettazioni predisposte dalla Magistratura, si è organizzato un gruppo di studio presso la Polizia Scientifica allo scopo di predisporre Linee Guida sulla trascrizione e proporre una soglia di intelligibilità del segnale al disotto della quale la trascrizione non debba essere effettuata. Lo studio su questa problematica è stato oggetto della relazione al convegno AISV a Trento (Andrea Paoloni "Limiti alla trascrizione giudiziaria" Trento 28.11.2006).

3.5.6 3I - Passaporto per l'Italia

Il progetto **"3I – Passaporto per l'Italia"** si propone di realizzare un corso di Italiano via internet del tutto innovativo, basato su contributi audiovisivi forniti dalla RAI.

Il corso può essere fruito a partire da qualsiasi piattaforma di distribuzione, in particolare è stato implementato sulla piattaforma "docent" attualmente attiva presso la Fondazione (www.fub.it/voice). La realizzazione di un corso di italiano vuole contribuire alla diffusione della nostra lingua come lingua di cultura per alimentare quell'interesse per l'italiano che sembra caratterizzare l'attuale panorama mondiale.

Il corso è espressamente dedicato ai cultori della nostra lingua residenti all'estero o in Italia e agli italiani all'estero di generazioni successive.

E' noto che sul mercato sono disponibili varie offerte per l'insegnamento della lingua italiana a parlanti altra madre lingua, il presente corso tuttavia si differenzia dagli altri per numerose caratteristiche di particolare rilievo che lo renderanno senza possibilità di dubbio, sia in Italia sia all'estero, il corso più avanzato ed efficace.

Si tratta di un corso autosufficiente (non integrativo di lezioni in presenza) fruibile in modo asincrono (in qualsiasi momento) ma dotato di un tutor (qualora lo si richieda) in grado di seguire, valutare e incoraggiare lo studente.

Il corso sarà tenuto da un docente virtuale che, attraverso la voce e con l'ausilio di schemi che appariranno sullo schermo, guiderà l'utente verso la comprensione delle strutture della lingua (grammatica) e alle attività connesse alla comprensione e alla produzione in italiano; al termine di ciascun segmento, in relazione alla percentuale di esercizi svolti in modo corretto, il docente virtuale fornirà informazioni sul livello di apprendimento dell'utente e lo guiderà verso il nuovo segmento di corso o, diversamente, verso una revisione delle strutture linguistiche già presentate ma non ancora apprese.

Altre caratteristiche del corso sono:

è conforme allo standard SCORM 1.2 e quindi multipiattaforma;

fa uso della tecnologia del TAL (trattamento automatico della lingua);

guida lo studente alla competenza nell'uso delle strutture linguistiche secondo una prospettiva funzionale, lasciando in secondo piano le competenze metalinguistiche;

fa riferimento agli standard europei per l'apprendimento della lingua ed è conforme agli standard sull'e-learning;

fa uso di materiali multimediali di elevata qualità grazie alla disponibilità del materiale delle Teche RAI;

al termine del corso l'utente è in grado di misurare la propria idoneità a superare con successo l'esame, con conseguente rilascio della certificazione coerente con il sillabus europeo da parte dell'Università della Tuscia.

Il corso è organizzato in tre livelli, il livello di base (principiante assoluto, principiante con conoscenze elementari), il livello intermedio e il livello avanzato. Il numero di ore per ciascun livello è di oltre 80 ore formative calcolate secondo gli standard europei.

Obiettivo del progetto è realizzare il primo livello del corso di e-learning in oggetto; al progetto partecipano l'Università della Tuscia, come fornitore dei contenuti e degli standard di certificazione, la RAI i come fornitore dei contenuti iconografici e multimediali e la ditta Infobyte specilizzata nello sviluppo di software multimediale, che si occuperà della riprogettazione dei contenuti in funzione del modello di didattica a distanza definito (instructional design), della personalizzazione della piattaforma Docent e dell'architettura e della realizzazione software, la Fondazione Bordoni fornirà, oltre alla direzione del progetto, la piattaforma Docent e le necessarie competenze nel trattamento del linguaggio naturale.

Una prima fase del progetto è stata dedicata a definirne dettagliatamente la struttura e a valutarne la fattibilità. In particolare è stata definita la tipologia di utenza, decidendo di rivolgersi al "target" dei cultori della lingua italiana, ovvero a persone di cultura medio alta interessati ad apprendere il nostro idioma, e sono state precisate le specifiche tecniche e gli standard nonché la piattaforma da impiegare, orientandosi sull'applicativo "docent, ".

Al termine di questa fase sono state rese disponibili una relazione sulla validità dl corso in confronto con le altre prospettive disponibili (Benmarking1.2), una scheda riassuntiva delle sue caratteristiche e un prototipo costituito da una singola lezione programmata sotto ambiente Docent.

Successivamente è stato messo a punto un documento (Doc.001/05) con la descrizione della struttura del corso che è articolato in 3 livelli:

1. livello di base, durata 60 ore complessive;

2. livello intermedio, durata 60 ore complessive;
3. livello avanzato, durata 60 ore complessive.

Per ciascun livello, in base ai moltiplicatori normalmente adottati, si avrà una erogazione pari a circa 180 ore di lezione frontale.

Le 60 ore di erogazione previste da ciascun livello saranno divise in 30 ore di contenuti didattici e test di valutazione, 10 ore di filmati, 10 ore di testi, 10 ore di materiale audio che può corrispondere alla lettura dei testi sopra citati.

Le 30 ore di contenuti didattici sono a loro volta suddivise in 90 Unità Didattiche (Learning Objects) di cui nel documento citato è dettagliata la struttura che segue la tipologia degli esercizi contenuta in un ulteriore documento di lavoro (Doc.00205).

Nel mese di settembre è stata completata la realizzazione di un nuovo prototipo sulla base del quale si è avviata, in collaborazione con l'Università della Tuscia e la RAI, la realizzazione del livello di base del corso (livelli A1 e A2), il cui completamento è previsto entro i primi mesi del 2007. Sarà poi necessaria una fase di verifica da effettuare con un gruppo di allievi di cui monitorare i problemi e la velocità di apprendimento.

E' stata realizzata la versione definitiva della grafica dei moduli sulla base di specifiche dettagliate degli storybord, la cui struttura è stata oggetto di un particolare approfondimento.

In giugno sono state completati gli storybord e sono stati selzionati, presso le Teche RAI, due terzi dei filmati di interesse per la realizzazione del corso.

In collaborazione con il gruppo di informatica è stato installato l'ambiente Software Docent su un server della Fondazione ed è stata individuata la procedura di sviluppo remoto che verrà utilizzata da infobyte per procedere nella realizzazione del corso.

Il corso è in fase di completamento (limitatamente ad alcuni LO di A2) per la mancanza di alcuni filmati e parallelamente si sta procedendo alla revisione dei refusi e degli errori per arrivare alla fase dei test di usabilità.

3.5.7 Corpora

Nell' ambito del tema del Trattamento Automatico del Linguaggio (TAL) si è ritenuto opportuno avviare un' attività di riorganizzazione e rivalutazione dei materiali audio che nel corso degli anni è stato raccolto per iniziativa dell'ISCOM e della Fondazione ed eventualmente anche di altro materiale audio reso disponibile da altri centri di ricerca.

Si tratta di corpora vocali di diversa natura, originariamente destinati ad applicazioni di sviluppo tecnologico. Con il termine corpora vengono designate raccolte di segnali vocali o di testi organizzate secondo particolari criteri. Per le ricerche e lo sviluppo nel campo del trattamento

automatico la lingua è spesso necessario disporre di corpora di grandi dimensioni sia di segnali vocali sia di testo.

L'attività verrà svolta coordinandosi con le realtà nazionali e, ove necessario, con la struttura europea preposta a questo compito (ELRA).

Realizzare un archivio del materiale audio organizzato in una piattaforma di facile uso al fine di rendere utilizzabili, per l'intera comunità scientifica, alcuni corpora oggi virtualmente inesistenti in quanto non conosciuti e non immediatamente disponibili.

La base di dati avrà come interfaccia un sito web appositamente progettato per essere consultato al fine di individuare, tra i corpora disponibili, quelli aventi caratteristiche idonee ad un particolare esperimento.

Il D.B.M.S. (DataBase Management System) dovrà consentire l'accesso a un particolare corpus sulla base di caratteristiche linguistiche ed acustiche. Tra i parametri evidenziati il numero e le caratteristiche dei parlanti, il tipo di parlato (spontaneo, letto, parole isolate), le caratteristiche di registrazione (frequenza di campionamento, livelli di quantizzazione).

Un sito web è il più immediato ed efficace strumento per cercare i corpora di interesse e attraverso di esso verranno messi a disposizione esempi di voci e/o intere porzioni di base di dati.

Inizialmente l'attività svolta è stata relativa all'analisi del materiale disponibile, allo studio dei formati con il quale è stato memorizzato ed alla realizzazione di una copia di sicurezza del segnale.

Si è successivamente proceduto a realizzare una base di dati dotata di interfaccia grafica per la consultazione rapida del data base, in vista di un suo successivo inserimento del sito del Ministero delle Comunicazioni.

La struttura della base di dati, che costituisce il motore del sito, è articolata in tre principali tabelle, la prima relativa ai corpora, che contiene i dati necessari a caratterizzarli, la seconda relativa ai parlanti e la terza relativa ai singoli file audio.

Successivamente si è realizzata l'interfaccia grafica, curandone la leggibilità e la possibilità di ottenere l'accesso a un corpus disponendo di una descrizione sommaria dello stesso.

Infine si è proceduto all'inserimento della descrizione di alcuni corpora al fine di eseguire un esperimento di consultazione della base di dati.

A base di dati e l'interfaccia grafica sono attualmente disponibili anche in un apposito CD rom realizzato nell'ambito del progetto.

3.5.8 Usabilità e accessibilità dei siti web

Sono state svolte delle attività sulla **Usabilità e Accessibilità dei siti web** che si sono occupate della valutazione dell'usabilità e dell'accessibilità di siti pubblici e privati, finanziato da Italgas,

Spes e la società Laziomatica della Regione Lazio.

ITALGAS: Valutazione dell'accessibilità del portale ItalgasPiù

Al fine di valutare l'accessibilità del portale ItalgasPiù, utilizzando la metodologia di valutazione dell'accessibilità sviluppata dalla Fondazione Ugo Bordoni, sono stati svolti tre tipi di valutazione: (1) valutazione automatica e semi automatica; (2) valutazione svolta da un esperto; (3) valutazione svolta da un esperto non vedente.

La valutazione automatica è stata svolta utilizzando i software, W3C Html Validator e W3C Css Validator, che analizzano la conformità del codice alle specifiche grammaticali.

La valutazione semi-automatica è stata svolta utilizzando il software Torquemada, sviluppato dalla FUB, e i software Cynthia e Bobby per l'analisi della conformità alle WCAG (Web Content Accessibility Guidelines).

La valutazione svolta da un esperto è stato una valutazione euristica; quella svolta dall'esperto non vedente è stata sia una valutazione euristica sia un test *task based.*

È stato quindi prodotto un documento di proposte per la correzione degli errori riscontrati dalle valutazioni.

Una volta che i tecnici dell'Italgas hanno realizzato le correzioni e le modifiche da noi proposte, sono stati ripetuti i tre test.

SPES (Centro di Servizio Volontariato del Lazio): Valutazione dell'accessibilità del sito www.volontariato.lazio.it

È stata svolta un'analisi dello stato di accessibilità del sito, utilizzando la metodologia di valutazione dell'accessibilità sviluppata dalla Fondazione Ugo Bordoni, che ha compreso un'analisi attraverso i valutatori automatici composta di 29 verifiche costituite a loro volta da vari sottopunti in cui si indica al "verificatore" cosa controllare e come.

È stato quindi steso un rapporto sullo stato del sito.

LAZIOMATICA – REGIONE LAZIO

Il lavoro si è articolato su tre tematiche:

Monitoraggio e supervisione della manutenzione del portale www.regione.lazio.it dal punto di vista dell'accessibilità.

Utilizzando la metodologia di valutazione dell'accessibilità sviluppata dalla Fondazione Ugo Bordoni, sono stati svolti tre tipi di valutazione: (1) valutazione automatica; (2) valutazione semi automatica; (3) valutazione con utenti.