Allegato A

Descrizione Tecnica del Progetto

Introduzione

Il inguaggio è una delle caratteristiche distintive della mente umana e rimane la forma più efficace di trasmissione delle conoscenze. Il linguaggio è usato non soltanto per esprimere, ma influenza la forma stessa del pensiero. Nonostante una parte sempre più vasta delle conoscenze e del pensiero umano sia disponibile in forma testuale e digitale, gli strumenti tuttora principalmente utilizzati per il suo trattamento si limitano ad analisi di tipo superficiale, ad esempio con tecniche di estrazione di parole chiave e di ricerca per chiavi, tipiche dell'Information Retrieval.
Per estrarre conoscenze ed effettuare una analisi non solo superficiale dei contenuti, occorre utilizzare strumenti di analisi del linguaggio di tipo sintattico/semantico. Le tecniche di estrazione di conoscenze da testi sono fondamentali in un ampio spettro di applicazioni, dal Question Answering, al Knowledge Management, al Semantic Web.
Un sistema di analisi dei testi deve essere in grado di assimilare dati testuali di qualunque dimensione e struttura, di estrarne i termini principali, assegnandoli a categorie di significati (tassonomia o ontologia) e individuarne relazioni semantiche.
Il risultato dell'analisi sono specifiche strutture organizzative della conoscenza, in grado di fornire adeguato supporto alle varie funzioni, dalla ricerca (Information Retrieval), alla risposta a domande (Question Answering).
L'Information Retrieval classico si occupa del recupero di documenti da una collezione tipicamente omogenea e ben organizzata come i testi in una biblioteca o di un archivio aziendale, generalmente di dimensioni relativamente limitate (qualche migliaia di documenenti).
Negli anni recenti, nuove problematiche sono emerse nell'ambito dell'IR dall'emergere del Web. I documenti sul Web hanno caratteristiche molto diverse dalle collezioni controllate, in particolare:

- disomogeneità: di contenuti, articolazione e struttura
- dimensioni: decine di milioni, aniché migliaia di documenti
- lingua e dizionario: svariate lingue che portano ad un dizionario complessivo di 10-100 milioni di termini
- variabilità: oltre il 20% del materiale cambia giornalmente.

In compenso la strutturazione delle pagine Web in HTML con la presenza di iperlink, offre specifiche opportunità per un'analisi dei contenuti e delle loro relazioni, quali ad esempio la classificazione per contesto [1].
Il trattamento di documenti Web ipertestuali ha richiesto lo sviluppo di specifiche tecnologie in grado di scalare su larga scala, sia in termini di numero di documenti che di numero di query al secondo. L'uso di algoritmi innovativi e di tecnologie di programmazione avanzate ha consentito di ottenere velocità di risposta superiori a quelle dei principali prodotti commerciali [3]. L'utilizzo di tecniche di grid computing invece consente di scalare nelle altre due dimensioni [4].
Le tecniche di Information Retrieval sono il presupposto minimo delle tecniche di analisi di testi, le quali intendono andare ben al di là degli obiettivi di recupero di documenti.
Ad esempio le tecniche di analisi di testo possono portare alla realizzazione di sistemi di Question Answering.
Un sistema di IR, viene interrogato con una query espressa ticipamente in una notazione artificiale come combinazione booleana di termini e la risposta viene fornita come elenco di documenti che contengono i termini, nella combinazione richiesta. Resta compito dell'utente scorrere i documenti restituiti per verificare se contengono le informazioni richieste.
Un sistema di QA invece viene interrogato sottoponendo una domanda in linguaggio naturale e si assume il compito di estrarre brevi frasi di senso compiuto (< 50 caratteri) che rispondono alla domanda. Sistemi di Question Answering vengono messi a confronto ogni anno alla TREC (Text Retrieval Conference) che si tiene a Washington presso il NIST. Il confronto viene misurato sulla qualità delle risposte che i vari sistemi sono in grado di fornire ad un insieme di circa 500 domande in lingua inglese, estraendo le risposte da una raccolta di circa 2 milioni di documenti provenienti dalla collezione Aquaint, formata principalemente da articoli di giornali.
Il gruppo Adaptive Computing del Dipartimento di Informatica partecipa da alcuni anni alla competizione con un proprio sistema denomiato PiQASso (Pisa Question Anwering System) [2] che nel 2001 si è classificato tra i primi dieci sistemi al mondo e primo in Europa.

Obiettivi

Il progetto si pone l'obiettivo di mettere a punto le tecnologie necessarie per l'analisi di testi in lingua italiana necessarie per la realizzazione di un sistema di Question Answering basato su relazioni semantiche.

Un sistema di Question Answering si compone di due parti:

- la costruzione della base documentale
- il trattamento delle domande

Verrà progettata una versione del sistema di Question Answering PIQASso, che incorpora tecniche di analisi liguistica più sofisticate e in particolare un parsing della lingua italiana, con la seguente architettura:

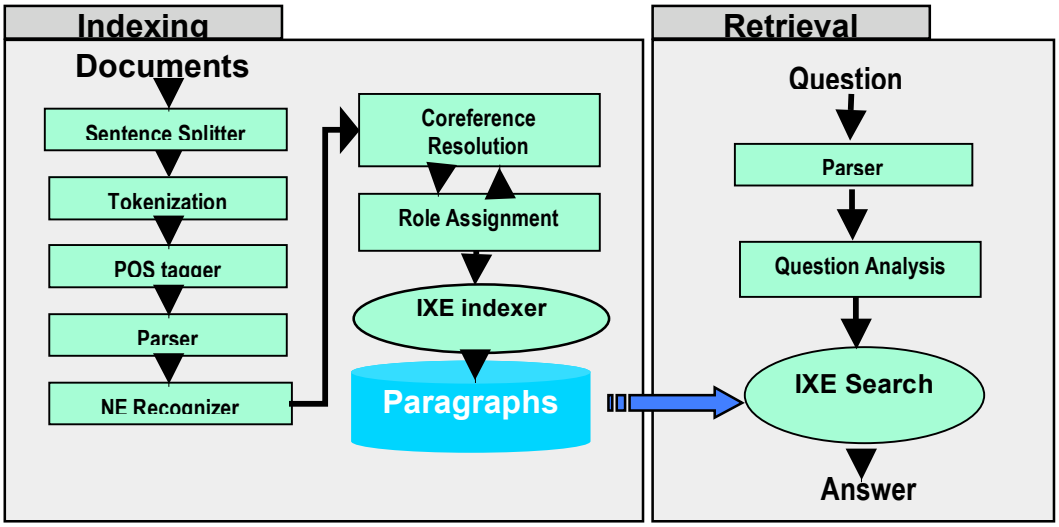


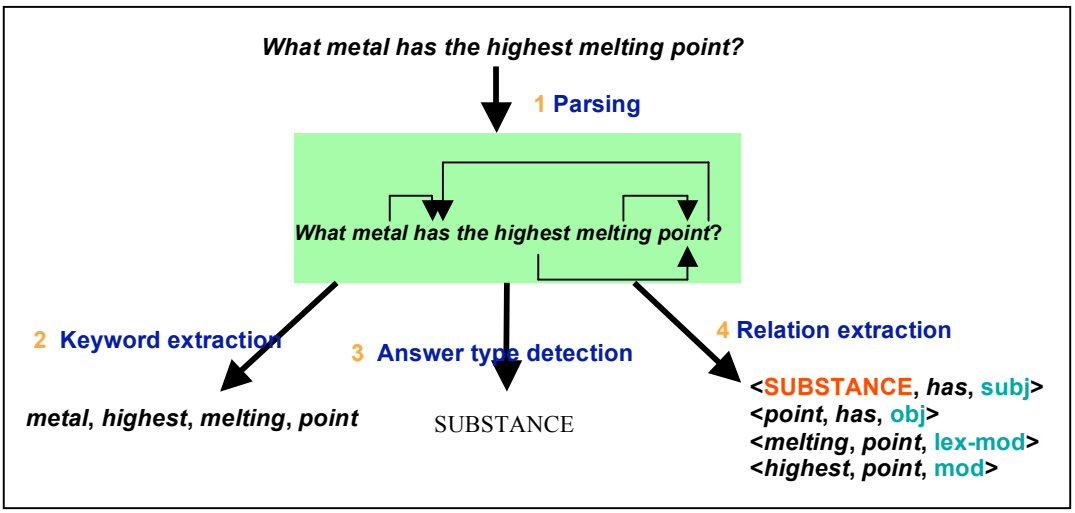
Figura 1. Architettura del sistema di Question Answering PIQASso**.**

In questa versione la costruzione della base documentale si limita alla lettura dei testi, la loro suddivisione in paragrafi e la costruzione di un indice full text contente informazioni sulla suddivisione in paragrafi.

Il trattamento delle domande consiste in:

1. parsing della domanda mediante il parser a dipendenze *Indeparser*
2. classificazione della domanda
3. formulazione dell'interrogazione da sottoporre all'indice dei paragrafi
4. estrazione di paragrafi candidati e per ciascuno di questi:
5. parsing mediante *Indeparser*
6. verifica di compatibilità tra tipo della domanda e tipo presente nella risposta
7. analisi delle corrispondenze semantiche tra domana e risposta
8. assegnazione di punteggio di vicinanza e di frequenza della risposta

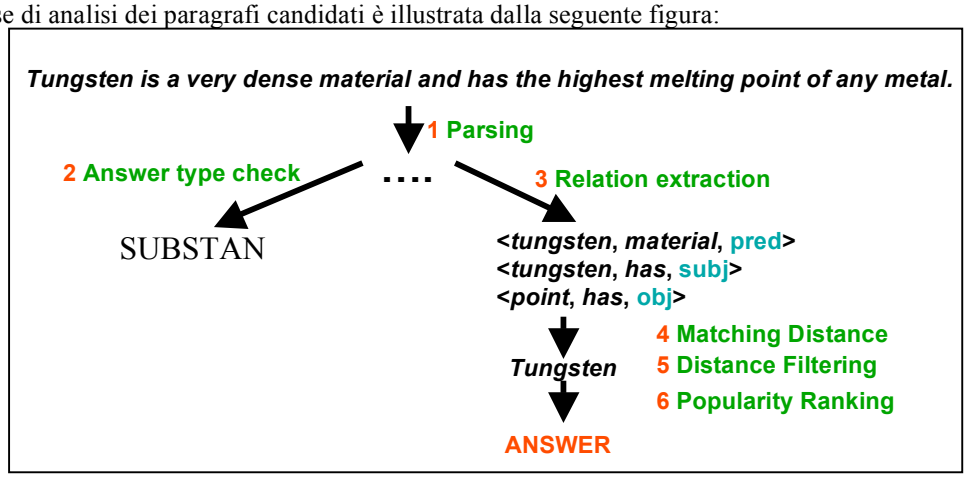
Il procedimento di analisi delle domande è illustrato nella seguente Figura:



i cui passi consistono in:

1. parsing
2. keyword extraction
3. answer type detection
4. relation extraction

La fase di analisi dei paragrafi candidati è illustrata dalla seguente figura:



le cui fasi consistono in:

1. parsing
2. answer type check
3. relation extraction
4. matching distance computation
5. distance filtering
6. popularity ranking

In questo progetto si prevede di mantenere l'impianto complessivo della precedente architettura di PiQASso, intervenendo in modo sostanziale sulla parte di analisi dei testi e di estrazione di conoscenze semantiche.
La motivazione di questa scelta è dettata dall'analisi delle prestazioni attuali del sistema, di cui si intende migliorare la selettività nell'estrazione di paragrafi candidati. Inoltre, le tecniche di estrazione di conoscenze semantiche sono utili anche per altri tipi di applicazioni di Knowledge Management.
Dall'esperienza svolta, effettuare questo tipo di analisi a tempo di trattamento della domanda risulta essere troppo oneroso e produce tempi di risposta elevati. Se si anticipa questa attività al tempo di costruzione della base di documenti ci si può aspettare un aumento di circa un ordine di grandezza nel tempo di indicizzazione, il quale risulta tuttavia tollerabile, in quanto l'operazione viene svolta una tantum sull'intera collezione di documenti e comunque ci si possono attendere prestazioni di indicizzazione intorno a 8 GB di testo all'ora.

Costruzione della base di documenti

In particolare, nella costruzione della base documentale verranno svolte le seguenti fasi:

1. acquisizione dei testi

2. preprocessing
3. estrazione di conoscenze
4. assegnazione di ruoli semantici
5. risoluzione di anafore
6. indicizzazione

Acquisizione testi

I testi vengono acquisiti da varie fonti, in vari formati, tra cui:

1. siti Web, HTML
2. news feed, RSS
3. documenti interni (Word, PDF, PostScript)
4. altre fonti in formato testo (mail)
5. Legacy DB o Warehouse

Per l'acquisizione si farà uso dell'architettura modulare di IXE, con l'uso di *pluggable document readers* e del suo *crawler* parallelo per l'acquisizione da Web.

Preprocessing del testo

Una semplice suddivisione in parole dei documenti, col criterio di caratteri delimitatori (spazio o punteggiatura) non è sufficiente per gli scopi di analisi dei testi, vista l'estrema varietà e ambiguità di notazioni usate nei testi.
Data la varietà e non rigida notazione che si ritrova nei testi, è di fatto impossibile usare tecniche puramente algoritmiche o di *pattern matching* per svolgere questo compito come anche gran parte degli altri di questa fase.
Pertanto verranno utilizzate tecniche di apprendimento automatico, in particolare basate sulle tecniche di classificazione di Maximum Entropy [8].
La scelta della tecnica della Maximum Entropy è dettata dalla flessibilità, ossia dalla possibilità di utilizzare per la classificazione un ampio numero di feature, anche di tipo diverso e non omogeneo tra loro e di non dover assumere l'indipendenza tra le feature, come richiesto da tecniche come Naive Bayes, utilizzati in altri prodotti attualmente in commercio.
Le tecniche di apprendimento automatico richiedono l'utilizzo di corpus linguistici annotati con tag a mano da persone, sulla base dei quali riconoscere gli elementi costitutivi delle frasi.
In particolare, il sistema di tokenizzazione sarà istruito nel riconoscere acronimi, numeri, abbreviazioni.
Per l'apprendimento verrà costruito un corpus opportuno per la lingua italiana.
Il secondo compito di apprendimento riguarda la scomposizione dei testi in frasi (Sentence Splitter), riconoscendo l'uso della punteggiatura e altre forme di suddivisione implicita, quali le forme tabellari o gli elenchi.
Anche per questo compito si farà uso di un sistema di apprendimento automatico.
La successiva fase di analisi del testo coinvolge l'utilizzo di tecniche di Natural Language Processing, in particolare si utilizzerà un Part-Of-Speech tagger per ottenere due aspetti per ciasun termine:

1. la categoria grammaticale (articolo, aggettivo, nome, verbo, ecc.)
2. il lemma della parola.

Per questo compito verrà utilizzato un tagger basato su tecniche di Decision Tree, anche questo appositamente allenato sulla lingua italiana [4].
Anche per la messa apunto di questo strumento è previsto lo sviluppo di un opportuno corpus di apprendimento e l'utilizzo di un completo dizionario dei lemmi della lingua italiana.
Infine, si utilizzerà opzionalmente un chunker per raggruppare sequenze di parole in gruppi quali gruppi verbali e gruppi nominali.
Riassumendo, la fase di preprocessing dei testi comprende le seguenti attività:

- Paragraph Detection
- Sentence Detection
- Tokenization
- POS Tagging
- Dependency Parsing

Estrazione di conoscenze

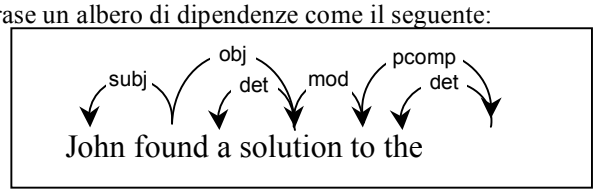
Il livello successivo di analisi dei testi ha il compito di estrarre conoscenze dai testi ed in particolare di estrarre le cosiddette Named Entity, ossia una parola od una frase che denota un nome proprio di persona, organizzazione, lougo, prodotto o un'espressione temporale o numerica.
Una tipica classificazione delle Named Entity consiste in 6 categorie principali (Time, Location, Organization, Person, Product, Event) e oltre una trentina di sottotipi.

L'identificazione di Named Entity è un aspetto essenziale per molti dei task di Knowledge Management, in particolare per il Question Answering, in quanto la risposta ad una domanda consiste frequentemente di una named entity. (Quanto costa un biglietto Roma-Milano? Chi è il presidente di Telecom?).
Inoltre il riconoscimento di NE è utile ad altri compiti di analisi, quali ad esempio l'estrazione di correlazioni (es. ruolo di una persona in un'organizzazione), la individuazione di eventi (Quando Telecom è stata acquisita da Pirelli?).
L'estrazione di NE con la tecnica della Maximum Entropy richiede la determinazione di un opportuno insieme di feature da riconoscere nei testi e di disporre di un training-set formato da testi annotati mediante queste stesse feature.
Con la nostra sperimentazione sono state selezionate le seguenti classi di feature:

1. word-level (es. presenza di maiuscole, minuscole, cifre, ecc.)
2. punteggiatura
3. categoria grammaticale (POS tag)
4. designatori di categoria
5. suffissi di categoria
6. termini interni a una sequenza in minuscolo
7. presenza in dizionari controllati

Parser della lingua italiana

L'architettura del sistema di Question Answering PIQASso prevede l'utilizzo di un parser del linguaggio naturale che viene utilizzato nella fase di analisi della domanda, per individuare gli elementi costituivi essenziali e il focus della domanda e per analizzare i paragrafi estratti dal motore di ricerca considerati contenere con maggiore probabilità la risposta alla domanda.
L'attuale versione del sistema utilizza Minipar [10], un parser a dipendenze per la lingua inglese, che produce per ciascuna frase un albero di dipendenze come il seguente:



Recentemente è stata proposta una tecnica innovativa per realizzare parser a dipendenze che si basa sull'induzione statistica e su tecniche di apprendimento automatico [5]. La tecnica evita la necessità di costruire ed utilizzare una grammatica formale del linguaggio, compito che rimane estremamente ardua, ed utilizza apprendimento da corpus di testi annotati.
Si intende sfruttare l'esperienza e risultati di precedenti attività svolte dal partner del progetto, Istituto di Linguistica Computazionale del CNR, per mettere a punto un corpus di documenti annotati con dipendenze da utilizzare nella fase di apprendimento del parser. A partire dal corpus SI-TAL, si prevede di mettere a punto dei programmi automatici che utilizzano euristiche per trasformare e completare il materiale in formato SI-TAL nel formato CoNLL.
Inoltre verrà messo a punto uno strumento visuale per analizzare le annotazioni così ottenute e effettuare eventuali correzioni manuali, in modo da ottenere un corpus completo ed affidabile.
Il corpus così prodotto potrà essere messo a disposizione della comunità scientifica internazionale, in modo ad esempio che anche la lingua italiana possa essere inclusa nelle sperimentazioni sui parser a dipendenze effettuate nell'ambito della conferenza CoNLL (Conference on Natural Language Learning) che quest'anno comprende 13 lingue, dal portoghese al cinese, dall'arabo al giapponese.
Il prototipo realizzato dai proponenti ha ottenuto risultato superiori alla media nella Shared Task della CoNLL 2006. Si prevede di svolgere ulteriori messe a punto dell'algoritmo, in particolare sperimentando la combinazione di più classificatori, per ottenere una accuratezza superiore al 90% nel parsing della lingua italiana.

Assegnazione di ruoli semantici

Un *ruolo semantico* nel linguaggio consiste in una relazione tra un costituente sintattico e un predicato. Tipici ruoli semantici includono Agente, Paziente, Strumento, ecc. ed anche ruoli aggiuntivi che indicano aspetti di Luogo, Tempo, Modi, Causa, ecc. Riconoscere ed assegnare etichette semantiche è essenziale per rispondere a domande di tipo "Chi", "Quando", "Cosa", "Dove" e "Perché" in compiti di Information Extraction, Question Answering e Summarization ed in generale in tutti i compiti di trattamento del linguaggio naturale in cui è richiesta una qualche forma di interpretazione semantica.

In particolare, il compito che si prevede di svolgere consiste nell'assegnare ruoli semantici ai costituenti di una frase in corrispondenza di determinati verbi. Ad esempio per il verbo *vendere*, il ruolo di *acquirente*, *venditore* e di *oggetto* e di *prezzo*.
Anche in questo compito si farà uso di tecniche di apprendimento e di corpus di apprendimento annotati con i tag di ruolo.
Si prevede di effettuare sperimentazione sulla base del corpus fornito da Penn TreeBank II usando i PropBank Frames [9], un insieme di ruoli semantici per i verbi della lingua inglese.

Risoluzione di anafore

Altro compito che verrà svolto durante l'analisi di testi è la risoluzione di anafore, ossia riferimenti abbreviati ad costituenti presenti in frasi precedenti, ad esempio pronomi. La risoluzione dell'anafora è un classico problema di NLP, tuttora soggetto di ricerca. Anche in questo caso si adotteranno tecniche di apprendimento, che possono produrre risultati con una precisione intorno all'80%.

Costruzione di indici arrichiti con metadati

I testi analizzati mediante gli strumenti descritti nelle fasi precedneti verranno indicizzati, costruendo indici arricchiti con metadati ricavati da queste analisi.
Gli indici conterranno quindi non soltanto le informazioni sui termini presenti e sulla loro posizione all'interno del documento (*posting lists*), come nei normali sistemi per la ricerca *full-text*. Nell'indice saranno rappresentati anche i metadati corrispondenti ai tag di tipo sintattico e semantico ricavati in precedenza. Sarà pertanto possibile effettuare ricerche con condizioni del tipo:

```
text matches person: bush

text matches location:* && bin-laden

text matches role:president && organization:apple
```

che selezionano i documenti che contengono la parola "bush" usata come nome di persona, documenti in cui compare "bin Laden" e una qualche indicazione di luogo e infine documenti in cui si parli del presidente di Apple, intesa come organizzazione.
L'indice dei documenti conterrà anche informazioni sulla suddivisione in frasi dei documenti, determinata attraverso il citato Sentence Splitter statistico, in modo da consentire il recupero di singoli paragrafi rilevanti alla risposta, anziché di interi documenti. Poiché termini rilevanti alla domanda possono apparire anche in paragrafi limitrofi, non è sufficiente adottare la soluzione di indicizzare i testi a livello di frasi anziché di documenti. Mantenendo nell'indice la struttura in documenti e l'informazione sulla divisione in paragrafi è possibile tener conto della presenza di termini in paragrafi vinici assegnando a tali termini una pesatura inferiore.
Per questo compito si utilizzerà una versione modificata del motore di ricerca IXE [3], sfruntttando le possibilità di specializzazione derivanti dalla sua architettura a oggetti estendibile.
PiQASso fa inoltre uso di WordNet [11], una base lessicale per la lingua inglese organizzata in una ontolgia di *synset*, ossia gruppi di significati.
Per adattare il sistema alla lingua italiana, si prevede di utilizzare un parser statistico multilingua, realizzandone una versione specifica per l'italiano.
Quanto a WorNet, si utilizzerà utilizzerà ItalWordNet [6], la versione per l'italiano prodotta dall'Istituto di Linguistica Computazionale del CNR, partner del progetto.

Topic Detection

Un'altra interessante applicazione delle tecniche fin qui discusse riguarda l'analisi di flussi di notizie giornalistiche, allo scopo di individuare *topic* (argomenti), in particolare quando emergono, riconoscerne duplicati e sovrapposizioni, e tracciarne l'evoluzione nel tempo. Tecniche di analisi statistica invece possono venire utilizzate per il calcolo del ranking della rilevanza o interesse di una notizia rispetto ad altre nello stesso arco temporale.
Uno compito interessante su cui misurare l'efficacia delle tecniche linguistiche riguarda la *First Topic Detection*, ovvere individurare quando una notizia introduce per prima un argomento nuovo.
Anche per questo compito si prevede di utilizzare tecniche di NLP per l'estrazione di metadati semantici (Named Entity, lexical chains) da usare come feature nella classificazione, per riconoscere somiglianze tra gli argomenti di varie notizie.
Il compito richiede sistemi non soltanto accurati ma anche efficienti, in grado di trattare un flusso di informazioni continuo come quello generato da agenzie di stampa o siti di noitizie sul Web, trasmessi sotto forma di feed RSS.

Riferimenti

[1] G. Attardi, S. Di Marco, D. Salvi, Categorisation by context, *Journal of Universal Computer Science*, 4(9),719-736, 1998.

[2] G. Attardi, A. Cisternino, F. Formica, M. Simi, A. Tommasi, C. Zavattari, PIQASso: PIsa Question Answering System *Proceedings of Text Retrieval Conference (Trec-10)*, 599-607, NIST, Gaithersburg (MD), November 13-16, 2001.

[3] G. Attardi, A. Cisternino, Template Metaprogramming an Object Interface to Relational Tables, *Reflection 2001*, *LNCS* **2192**, 266-267, Springer-Verlag, Berlin, 2001.

[4] G. Attardi, V. Sinha, "Design of a Web Switch Architecture", Deliverable Progetto FIRB Grid.it, 2003.

[5] G. Attardi, Experiments with a multilanguage projective dependency parser, CoNLL-X, 2006.

[6] R. Bartolini, Lenci A., Montemagni S., Pirrelli V., Hybrid Constraints for Robust Parsing: First Experiments and Evaluation, in *Proceedings of LREC 2004*, Fourth International Conference on Language Resources and Evaluation, 26-28 May 2004, Centro Cultural de Belem, Lisbon, Portugal, pp. 795-798, 2004.

[7] Italian TreeTagger, http://medialab.di.unipi.it/Resource/POS/index.html.en.

[8] Adwait Ratnaparkhi. (1999). Learning to Parse Natural Language with Maximum Entropy Models. *Machine Learning*, 34, 151–175.

[9] PropBank Frames, http://www.lsi.upc.es/~conll04st/resources/pb-frames.tar.gz.

[10] D. Lin, LaTaT: Language and Text Analysis Tools, *Proc. Human Language Technology Conference*, San Diego, California, March 2001. http://hlt2001.org.

[11] G. Miller, Five papers on WordNet. *Special issue of International Lexicography* 3(4), 1990.

[12] J. C. Reyner and A. Ratnaparkhi, A Maximum Entropy Approachg to Identify Sentence Boundaries. *Computational Language*, 1997.

[13] G. Schmid, TreeTagger – a language independent part-of-speech tagger, 1994. Available: http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

Allegato B

Programma di autovalutazione

Il progetto prevede la realizzazione di tecnologie e sistemi software secondo le specifiche e gli standard di alcune delle principali iniziative di valutazione internazionali del settore scientifico.
Le valutazioni consistono nella sottomissione di alcuni *run* di esecuzione sperimentali su un benchmark costruito ogni anno dagli organizzatori. A ciascun partecipante vengono forniti una base di dati di partenza su cui allenare il proprio sistema e un task che il sistema deve svolgere nel corso di meno di una settimana.
I risultati dei run vengono raccolti dagli organizzatori e sottoposti ad analisi da parte di un gruppo di valutatori, sulla base dei quali vengono calcolate delle metriche analitiche standard (*precision*, *recall*, *accuracy*) che consentono di valutare comparativamente la qualità dei risultati ottenuti da ciascun sistema.
Questo tipo di valutazione è da considerarsi una valutazione di tipo scientifico e oggettivo: pertanto più che di una autovalutazione, il progetto verrà sottoposto ad una vera e propria valutazione scientifica.
In particolare il parser a dipendenze verrà sottoposto alla valutazione effettuata nell'ambito della Conference on Natural Language Learning, che si svolge annualmente, con scadenza di sottomissione intorno a marzo di ogni anno.
L'obiettivo del progetto è di ottenere risultati di precisione e accuratezza nell'analisi di lingue romanze (italiano, spagnolo e portoghese) superiori alla media.
Per quanto riguarda il Question Answering, la principale valutazione a livello internazionale dei sistemi di Question Answering viene svolta nell'ambito della conferenza TREC (Text Retrieval Conference), che si svolge presso il National Institute of Standards ogni anno a novembre, con scadenza di sottomissione degli esperimenti ad agosto di ogni anno.