ONTOTEXT
Dal Testo alla Conoscenza per il Web Semantico

Programma di Ricerca FUP-2004
Provincia Autonoma di Trento

Accesso all'informazione basato su conoscenza: funzionalità del portale *ONTOTEXT* (Versione 31/12/2006)

Luisa Bentivogli, Bernardo Magnini
ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive 18, Povo (38050) – Trento, Italia

{bentivo, magnini}@itc.it

Technical Report

15 giugno 2006

Indice

Abstract

Il progetto Ontotext nasce per studiare e sviluppare tecniche innovative di estrazione della conoscenza al fine di produrre nuova e coerente informazione destinata al Web Semantico. La creazione del portale ONTOTEXT rappresenta lo scenario applicativo del progetto Ontotext, il cui obiettivo finale consiste nel costruire un portale in grado di presentare l'informazione nel modo più possibile flessibile ed aderente alle richieste dell'utente. La principale e innovativa caratteristica del portale ONTOTEXT consiste nel permettere un accesso all'informazione non più basato esclusivamente sui testi, come avviene attualmente nei maggiori portali a disposizione sul Web, bensì basato e guidato dalla conoscenza estratta dai testi stessi.

In questo documento sono presentate le funzionalità del portale dimostrativo di ONTOTEXT che sarà rilasciato al 31/12/2006. Verrà descritta nel dettaglio l'interfaccia utente del portale e analizzati una serie di problemi relativi a come interpretare le domande dell'utente, come trovare le risposte nella base di conoscenza, quali informazioni presentare e come presentarle al meglio.

1 Introduzione

La creazione del portale ONTOTEXT rappresenta lo scenario applicativo del progetto Ontotext, un progetto FU-PAT iniziato nell'ottobre del 2004 che coinvolge ITC-irst e ISTI-CNR. L'obiettivo generale del progetto consiste nello studiare e sviluppare tecniche innovative di estrazione della conoscenza al fine di produrre nuova e coerente informazione destinata al Web Semantico.

L'approccio più comune alla creazione di un Web Semantico si basa sull'annotazione delle risorse disponibili in Rete rispetto a concetti e relazioni definiti all'interno di ontologie, che vengono usate come strumenti concettuali per la comunicazione della conoscenza. L'annotazione semantica delle risorse Web è intesa a facilitarne l'accesso sia da parte di esseri umani sia da parte di agenti artificiali.

L'informazione contenuta nei documenti annotati può essere però incoerente e frammentaria e può cambiare nel tempo. Di conseguenza, il suo sfruttamento può essere complesso, sia da parte di esseri umani che da parte di agenti artificiali. Pur aderendo alla prospettiva del Web Semantico, Ontotext mira a mitigare questo problema integrando nel Web Semantico un nuovo tipo di fonte di informazione risultante da un processo di estrazione di conoscenza guidato da ontologie. Nella nostra prospettiva, la conoscenza contenuta nelle risorse annotate viene estratta e organizzata in una base di conoscenza strutturata, che permette una rappresentazione ed un aggiornamento coerenti dell'informazione, garantendone allo stesso tempo la tracciabilità rispetto alle fonti. I fatti così raccolti possono poi essere impiegati per perfezionare ed estendere le ontologie esistenti. Sia la base di conoscenza che le ontologie estese vengono a loro volta rese disponibili al Web Semantico.

Lo scenario concreto in cui le nuove tecnologie sviluppate verranno testate sarà l'acquisizione automatica di diversi tipi di informazione da articoli del giornale locale *L'Adige*, con il quale è stato concluso un accordo che autorizza l'ITC-irst a processare i dati delle ultime dieci annate del quotidiano da loro fornite. Le potenzialità di integrazione di queste tecnologie saranno dimostrate mediante la realizzazione del portale Web ONTOTEXT, che permetterà ai cittadini di consultare la conoscenza estratta dai testi attraverso un'interfaccia facile da usare e orientata all'utente.

2 La conoscenza in Ontotext

La principale e innovativa caratteristica del portale ONTOTEXT consiste nel permettere un accesso all'informazione non più basato esclusivamente sui testi, come avviene attualmente nei maggiori portali a disposizione sul Web, bensì basato e guidato dalla conoscenza estratta dai testi stessi.

Il progetto Ontotext si pone come obiettivo principale di estrarre e organizzare infomazioni su alcune particolari ***categorie ontologiche***: entità, oggetti temporali, relazioni, eventi, argomenti (topics) e opinioni.

Le ***entità*** denotano oggetti o insiemi di oggetti in un dominio di riferimento. Ontotext è interessato alle seguenti entità: "Persone" (es. Paolo Maldini), "Organizzazioni" (es. Vodafone), "Luoghi" (es. Mar Rosso), "Entità Geo-Politiche" (es. Roma). Per ogni entità individuata, viene compilata una *Scheda Ontologica* contenente le principali informazioni estratte per l'entità stessa, ad esempio per le entità di tipo "Persona": Nome, Cognome, Sesso, Professione, Affiliazione ecc.

Una ***relazione*** è una coppia ordinata di entità, temporalmente collegata (es. il numero di abitanti di una città in un dato anno).

Un ***evento*** è qualcosa che accade, coinvolge un certo numero di partecipanti e ha come risultato un cambiamento di stato. L'omicidio è un classico esempio di evento.

Gli ***argomenti*** (topics) sono insiemi di documenti (articoli di giornale nel nostro caso) che riguardano lo stesso avvenimento. Ad esempio "Giro d'Italia 2004", "Dimissioni di Collina". Gli argomenti hanno un certo grado di attivazione in un certo periodo, che dipende da due fattori: (i) la dimensione dell'argomento (il numero dei fatti di cui consiste) e (ii) la frequenza dell'argomento.

Infine, le ***opinioni*** sono giudizi soggettivi espressi riguardo entità, relazioni, eventi e argomenti. Come primo passo verso l'obiettivo finale di identificare e connotare le opinioni presenti nei testi, l'attività di ricerca in questo ambito è attualmente focalizzata sull'individuazione nei testi dei termini "soggettivi", cioè di quei termini che veicolano non un contenuto oggettivo bensì un'opinione (es. stima vs. cielo) e sul riconoscimento della polarità di questi termini, ovvero se la connotazione del termine è positiva o negativa (es. stima vs. disprezzo).

Per la fine del progetto (30 settembre 2007) verrano studiate le tecnologie e sviluppati gli algoritmi necessari ad estrarre ed organizzare automaticamente la conoscenza relativa a tutte categorie ontologiche di Ontotext descritte sopra. Tale conoscenza sarà estratta dai testi del ***corpus Adige-500.000,*** un corpus formato dalle ultime dieci annate del quotidiano locale *L'Adige,* per un totale di circa 500.000 articoli. Dati e conoscenza acquisita saranno poi messi a disposizione tramite il portale ONTOTEXT. È prevista dal progetto una fase intermedia di sperimentazione i cui risultati saranno resi disponibili al 31/12/2006 tramite il portale dimostrativo ONTOTEXT che viene descritto nelle sezioni seguenti.

3 Il portale dimostrativo ONTOTEXT (versione 31/12/2006)

Le funzionalità descritte in questo documento si riferiscono al portale intermedio dimostrativo che verrà rilasciato al 31 dicembre 2006 e che darà accesso all'informazione resa disponibile all'interno del progetto per quella data. Nelle due sezioni successive sono presentate le informazioni disponibili e le funzionalità che il portale dimostrativo dovrà avere al 31 dicembre 2006.

3.1 Dati disponibili

Alla data intermedia del 31 dicembre 2006 saranno disponibili in Ontotext i dati descritti di seguito.

Corpora:

- Corpus Adige-500 (I-CAB). 525 articoli (di cui 335 utilizzati come *training set* e 190 come *test set* per gli algoritmi automatici) su cui sono state effettuate una serie di operazioni ***manuali***:
    - annotazione di tutte le espressioni temporali
    - annotazione di tutte le entità
    - annotazione di tutte le relazioni di tipo "Personal-Social" sussistenti tra le entità di tipo "Persona", in particolare le relazioni *familiari* (genitore/figlio, ecc.) , *di lavoro* (colleghi, capo) e di tipo *personale durevole* (vicini di casa, compagno di classe)
    - per tutte le entità, creazione di una *Scheda Ontologica* contenente una serie di informazioni rilevanti per quella data entità
- Corpus Adige-50.000. 52.000 articoli annotati automaticamente con informazione di tipo linguistico (lemma e parte del discorso), entità nominate ed espressioni

temporali. Su questo corpus di medie dimensioni vengono attualmente testate tutte le procedure automatiche prima di passare al corpus finale.

- Corpus Adige-500.000. È il corpus finale di Ontotext, composto di 500.000 articoli. Al 31/12/2006 sarà annotato con informazione di tipo linguistico (lemma e parte del discorso), entità nominate ed espressioni temporali mentre per la fine del progetto tutte le procedure funzioneranno su questo corpus che sarà interamente accessibile tramite il portale.

Ontologie:

- Ontologia delle Persone
- Ontologia delle Organizzazioni
- Ontologia dei Luoghi
- Ontologia delle Entità Geo-Politiche

Procedure automatiche:

- Processori linguistici per l'annotazione automatica dei diversi tipi di informazione linguistica necessaria in Ontotext (lemma, parte del discorso, collocazioni, sensi di parola)
- Sistema di Named Entity Recognition per il riconoscimento delle entità nominate nei testi
- Sistema di riconoscimento e normalizzazione delle espressioni temporali nei testi
- Algoritmi per il riconoscimento di argomenti (topics) e classificazione dei testi in base all'argomento di cui trattano
- Algoritmi per il riconoscimento di opinioni positive e negative nei testi

Come vedremo nel dettaglio in seguito, alcune funzionalità del portale dimostrativo accederanno al corpus I-CAB e quindi alle informazioni codificate manualmente nei testi o estratte manualmente dai testi, mentre altre sfrutteranno già il corpus finale Adige-500.000 e le informazioni estratte tramite le procedure automatiche dai 500.000 testi.

Dal punto di vista delle categorie ontologiche studiate in Ontotext, al 31 dicembre 2006 non saranno a disposizione informazioni sugli eventi.

3.2 Funzionalità del portale

Al 31/12/2006 saranno disponibili dal portale dimostrativo le seguenti funzionalità:

1) Citografo
2) Ricerca di una o più parole
3) Ricerche avanzate (parole in una determinata parte del discorso, lemmi, lemmi in una determinata parte del discorso, sottostringhe, ricerca di più parole con restrizioni sulla modalità di ricerca)
4) Ricerca di collocazioni
5) Accesso alle schede informative delle collocazioni
6) Ricerca/browsing di entità
   a) entità nominate
   b) menzioni di entità
   c) co-referenze di entità a livello di singolo documento
   d) co-referenze di entità a livello di corpus
7) Accesso alle schede ontologiche delle entità
8) Ricerca/browsing di argomenti

9) Ricerca/visualizzazione di relazioni tra persone
10) Ricerca di opinioni
11) Accesso ai testi e ai diversi livelli di annotazione

Le funzionalità di ricerca/browsing di menzioni di entità (5b) e di co-referenza (5c e 5d), di accesso alle schede ontologiche delle entità (6) e alle relazioni tra persone (8) sono disponibili solo per i dati ricavati *manualmente* dal corpus I-CAB, e quindi su un totale di 500 documenti. Tutte le altre funzionalità utilizzano invece informazioni estratte *automaticamente* dal corpus Adige-500.000.

4 L'interfaccia utente

Data la conoscenza a disposizione, unitamente ai testi da cui è stata estratta, la sfida di Ontotext consiste nel costruire un portale in grado di presentare tale conoscenza nel modo più possibile flessibile ed aderente alle richieste dell'utente. In questa sezione descriveremo in dettaglio l'interfaccia utente del portale ONTOTEXT e analizzeremo una serie di problemi relativi a come interpretare le domande dell'utente, come trovare le risposte nella base di conoscenza, quali informazioni presentare e come presentarle al meglio.

4.1 Accesso al portale: ricerca di informazioni da parte dell'utente

L'utente ha la possibilità di effettuare due tipi di ricerca:
(a) eseguire una ricerca standard, tipo Google, inserendo una o più parole nell'apposito campo
(b) navigare direttamente le categorie ontologiche di Ontotext

La pagina di accesso al portale, da cui sono disponibili contemporaneamente le due modalità di ricerca, è riportata in Figura 1.

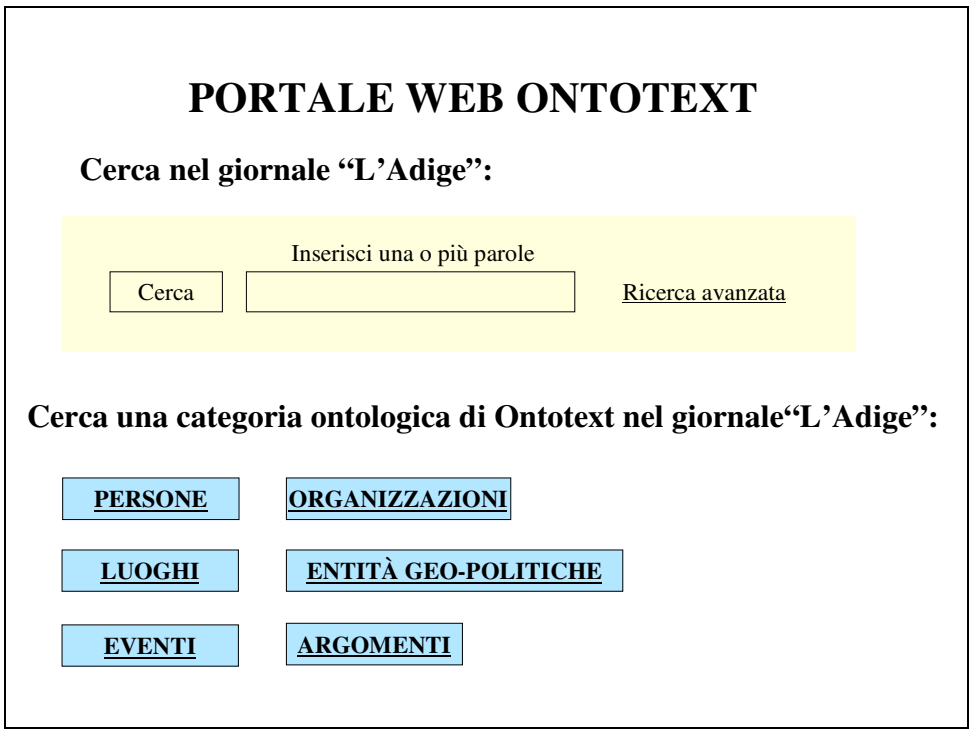

Figura 1: pagina di accesso al portale Ontotext

4.1.1 Query tramite parole chiave

Nella prima modalità di ricerca, l'utente inserisce una o più parole nell'apposito campo. L'utente può fare vari tipi di query, ad esempio una domanda vera e propria: "qual è attualmente il sindaco di Trento?"; oppure inserire una serie di parole chiave: "Elenco iscritti partito verdi"; oppure essere ancora più generico e inserire semplicemente una parola: "margherita". È il sistema che si occupa poi di interpretare la query dell'utente in modo da restituire le informazioni più aderenti alla richiesta.

È inoltre prevista una funzionalità di ricerca avanzata in cui l'utente può ad esempio cercare sottostringhe di parole o fare ricerche più raffinate dal punto di vista linguistico, come cercare una parola in una certa parte del discorso o un lemma, in una determinata parte del discorso o meno. Questa modalità è rappresentata in Figura 2.

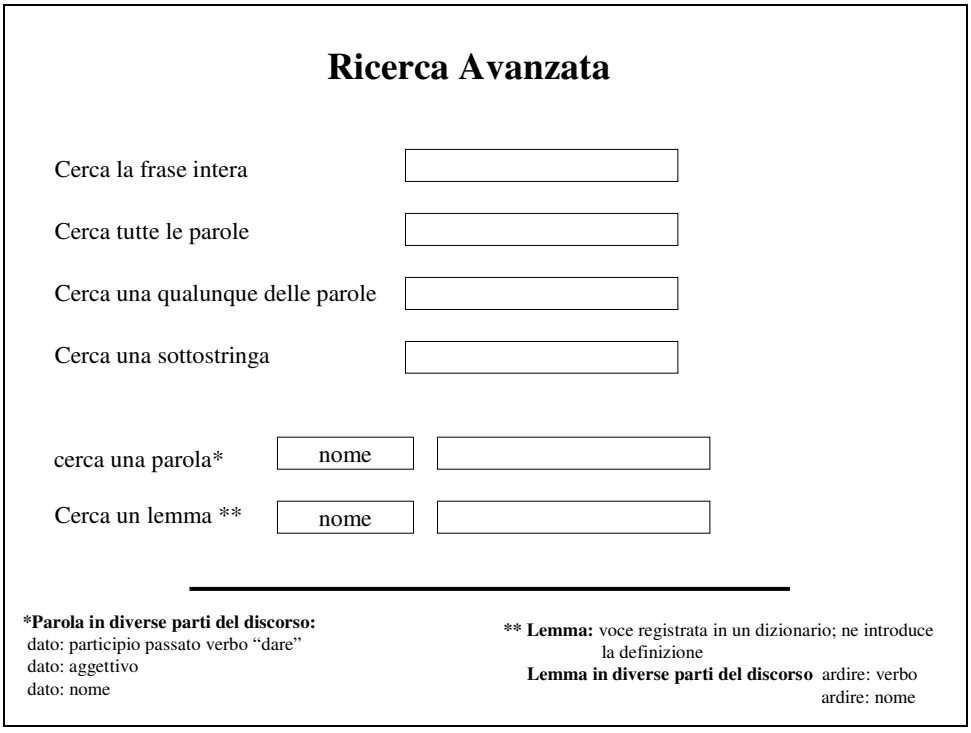

Figura 2: modalità di ricerca avanzata

4.1.2 Browsing delle categorie ontologiche di Ontotext

Nella seconda modalità di ricerca, l'utente può fare direttamente il browsing delle categorie di Ontotext. Ogni categoria è strutturata in una gerarchia e l'utente può navigarla, scegliendo le sottocategorie a cui è interessato. In alternativa, può ignorare la gerarchia e visualizzare l'elenco completo in ordine alfabetico degli appartenenti alla categoria selezionata. Una volta selezionata una determinata entità o argomento, si raggiunge direttamente la relativa scheda informativa descritta in dettaglio nelle sezioni 4.3.1 e 4.3.2.

Ad esempio, come mostrato in Figura 3, l'utente interessato alle "organizzazioni" di tipo "istruzione" può cliccare sulla sottocategoria nella parte destra della pagina e nella parte sinistra compaiono tutte le entità di quel tipo. Cliccando su una di queste si raggiunge la scheda che raccoglie tutte le informazioni a disposizione del sistema su quella entità.

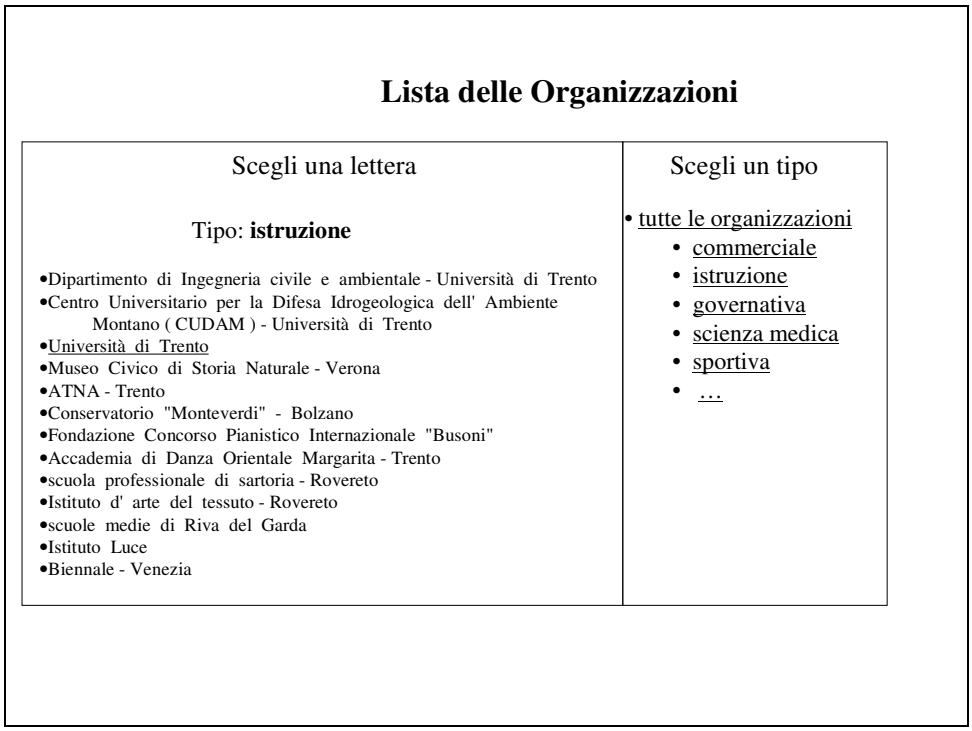

Figura 3: funzionalità di browsing delle categorie di Ontotext

4.2 La risposta del sistema alla query dell'utente

Nel caso in cui l'utente abbia deciso di fare una query cercando parole chiave, il sistema interpreta la query e fornisce una risposta. I risultati ottenuti vengono raggruppati dal sistema in maniera intelligente:

(i) viene visualizzato l'elenco delle *entità* che soddisfano la query.

(ii) viene visualizzato l'elenco degli *argomenti* che soddisfano la query.

(iii) viene presentato l'elenco delle *collocazioni*[1] contenenti la/e parola/e della query. Le collocazioni rappresentano un livello intermedio tra la conoscenza e il testo puro.

(iv) vengono presentati tutti i risultati ottenuti, senza tener conto della conoscenza del sistema. La query viene trattata come un insieme di parole che vengono cercate nei documenti. Il risultato viene visualizzato o sotto forma di *concordanze*[2] (quando viene trovata esattamente la parola o la frase inserita dall'utente) o di *snippets*[3] di documenti (quando nei documenti vengono trovate separatamente le singole parole che compongono la query e non la query stessa così come è stata digitata).

[1] Una collocazione è una sequenza di parole che co-occorrono abitualmente e il cui significato complessivo non è completamente ricostruibile dal significato delle singole parole che la compongono. Ad esempio "senso unico" , "montagne russe".

[2] Per concordanza si intende la lista di tutte le occorrenze di una data parola in un testo, assieme al contesto in cui quella parola occorre. La parola cercata è centrata rispetto alla riga ed evidenziata graficamente, il contesto è rappresentato da un certo numero di parole che precedono e che seguono la parola cercata.

[3] Uno snippet è l'insieme delle porzioni del testo in cui compaiono le parole della query, riunite in un unico paragrafo ai fini della presentazione all'utente.

Tutte queste informazioni vengono recuperate dal corpus Adige-500 e sono il risultato di annotazioni automatiche dei testi.

Vediamo ad esempio in Figura 4 il risultato fornito dal sistema alla query costituita da una singola parola: "margherita". Due *entità* di tipo diverso (Persone e Organizzazioni) e un *argomento* soddisfano la query; oppure l'utente ha la possibilità di vedere tutti i risultati assieme. In questo esempio il sistema non ha trovato nel corpus nessuna collocazione in cui figurasse la parola "margherita", quindi il livello delle collocazioni non viene presentato.

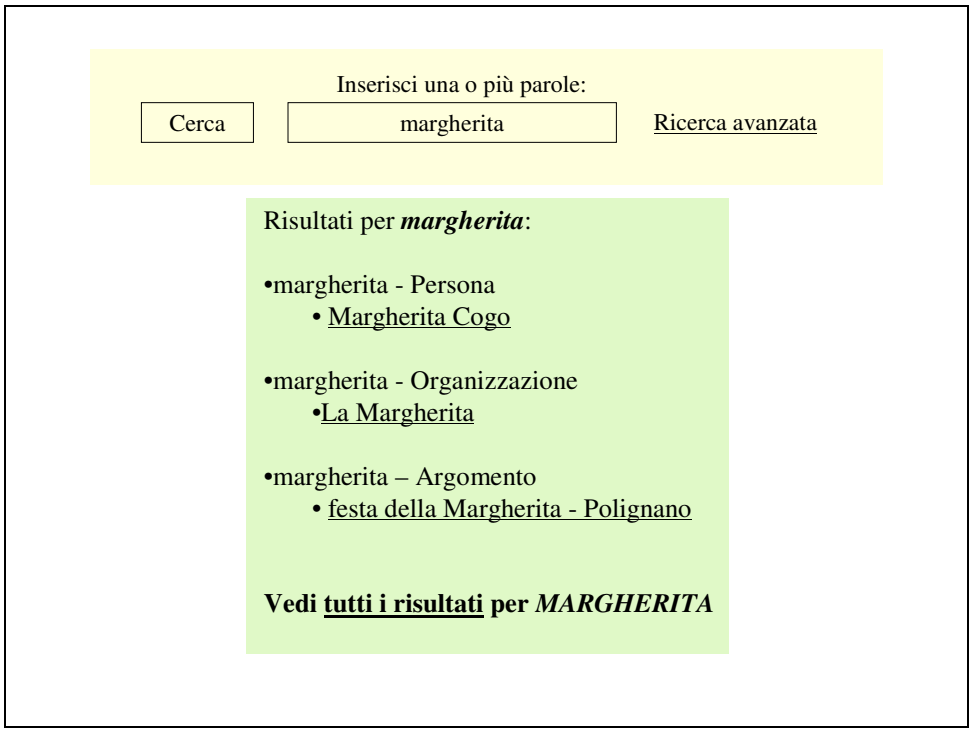

Figura 4: risposta del sistema alla query dell'utente "margherita"

Nell'esempio in Figura 5, data la query "rosso" il sistema presenta sia entità (di tipo Persone e Luoghi) sia una serie di collocazioni di cui la parola "rosso" fa parte.

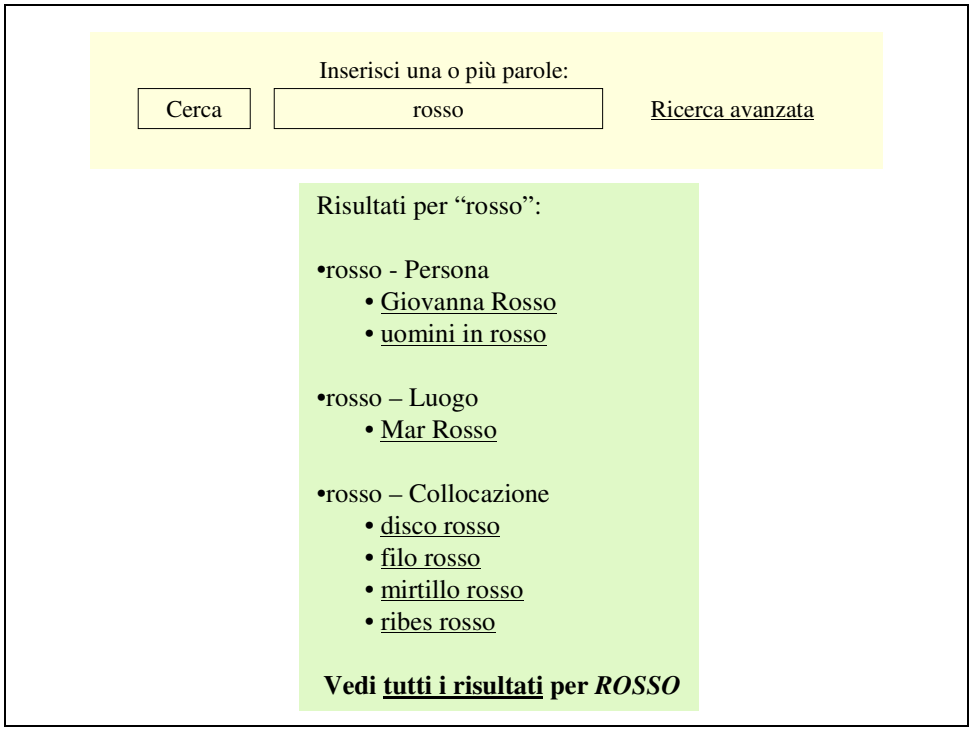

Figura 5: risposta del sistema alla query dell'utente"rosso"

Infine, nell'esempio in Figura 6 il sistema interpreta la query complessa dell'utente e riesce a trovare un'entità di tipo Organizzazioni (il partito politico dei verdi) che la soddisfa. Restituisce inoltre un elenco di documenti in cui compaiono le singole parole che compongono la query.

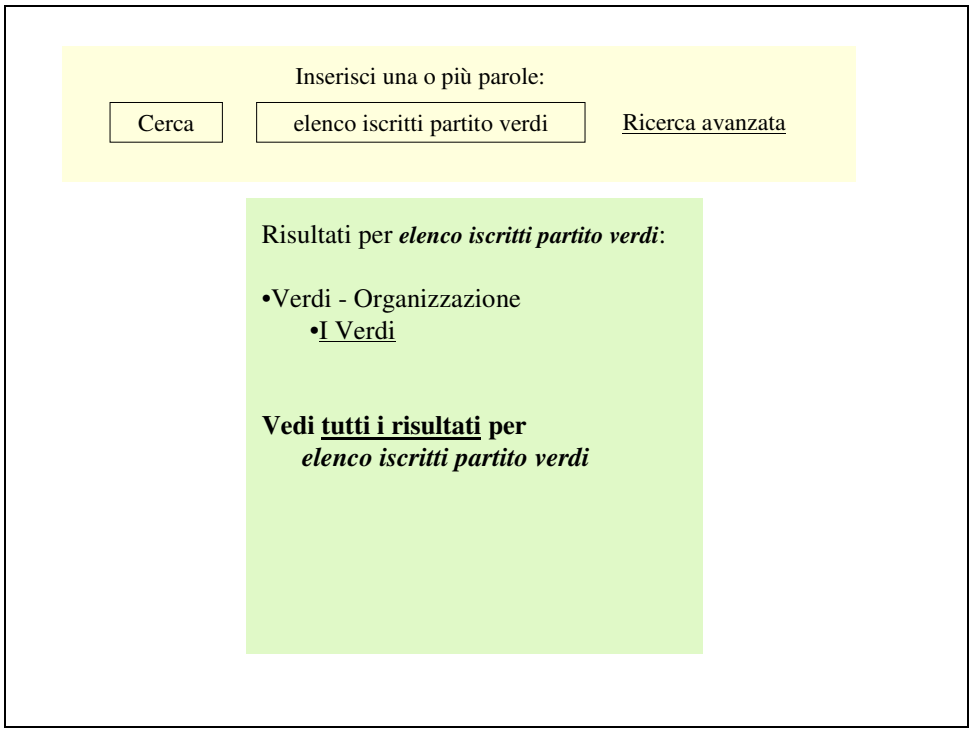

Figura 6: risposta del sistema alla query dell'utente"elenco iscritti partito verdi"

4.3 Presentazione dell'informazione

Sia nel caso di una query a parole chiave (cfr. sezioni 4.1.1 e 4.2) sia nel caso di browsing diretto delle categorie ontologiche di Ontotext (cfr. sezione 4.1.2.), il sistema presenta all'utente tutta la conoscenza a disposizione su entità ed argomenti, fornendo inoltre l'accesso ai singoli documenti annotati. Inoltre, in caso di query a parole chiave, sono attive anche la funzione di visualizzazione di collocazioni e quella di "concordancer", che dà all'utente la possibilità di visualizzare tutti i documenti che contengono le parole della sua query, indipendentemente dalle conoscenze possedute dal sistema. Nelle sezioni che seguono, le informazioni presentate e le modalità di presentazione vengono descritte dettagliatamente.

4.3.1 Informazioni su un'entità

L'utente clicca su una determinata entità e il sistema restituisce una scheda che contiene tutta la conoscenza posseduta su quella entità. La schermata è suddivisa in quattro frame contenenti:

1. scheda ontologica dell'entità. La scheda ontologica di un'entità contiene le principali informazioni estratte per l'entità stessa, contestualizzate temporalmente. Ad esempio per le persone sono presenti informazioni come Nome, Cognome, Sesso, Professione, Affiliazione ecc. e, dove necessario come ad esempio per Professione e Affiliazione, è presente anche l'informazione del periodo a cui quell'informazione si riferisce. Ogni valore degli attributi della scheda ontologica è cliccabile e rimanda all'articolo da cui quel valore è stato estratto.
2. le concordanze dell'entità, ordinate dall'articolo più recente a quello meno recente. Ogni concordanza è collegata all'articolo da cui è stata estratta. In questo frame è disponibile anche la funzionalità di visualizzazione delle opinioni positive e negative sull'entità in esame, contestualizzate temporalmente. Se l'entità è una persona, è possibile avere anche le opinioni (positive e negative) espresse da quell'entità, sempre temporalmente contestualizzate.
3. accesso al citografo relativo all'entità. Il citografo è un grafico che mostra l'andamento temporale delle citazioni dell'entità presa in considerazione nei vari articoli di giornale.
4. possibilità di visualizzare altre entità collegate tramite una relazione all'entità in esame e di navigare la porzione di ontologia in cui si trova l'entità stessa.

I frame 1 e 4 contengono informazione estratta manualmente dal corpus I-CAB e sono quindi disponibili solo ed esclusivamente se l'entità presa in considerazione è presente nel corpus I-CAB.

Per quanto riguarda il frame 2, se nel corpus I-CAB è presente l'entità presa in considerazione l'utente può scegliere se visualizzare le *concordanze dell'entità* nel corpus I-CAB dove le co-referenze a livello di corpus sono risolte, oppure se visualizzare le *concordanze a livello di Entità Nominate* estratte automaticamente dal corpus Adige-500.000, dove la co-referenza delle entità a livello di corpus non è stata codificata. Se l'entità non è presente in I-CAB saranno disponibili solo le concordanze di Entità Nominate estratte da Adige-500.000.

Anche il frame 3 è sempre disponibile ma, poiché il citografo necessita di una grossa quantià di dati e quindi lavora su Adige-500.000 dove le co-referenze delle entità non sono state risolte, non funziona a livello di entità bensì di occorrenza di Entità Nominata.

Nell'esempio in Figura 7 è mostrata la scheda personale che il sistema presenta all'utente che ha cliccato sull'entità di tipo persona "Margherita Cogo" (cfr. Figura 4).

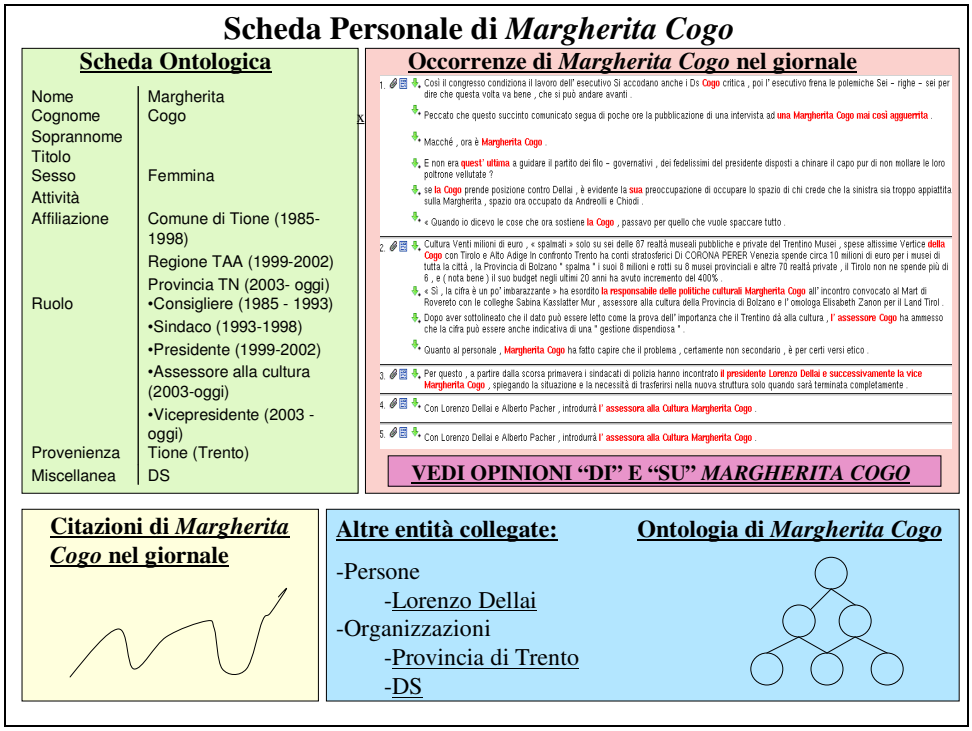

Figura 7: Scheda informativa di un'entità

4.3.2 Informazioni su un argomento (topic)

L'utente clicca su un determinato argomento e il sistema restituisce una scheda strutturata in modo parallelo a quella per le entità. Anche la scheda per gli argomenti è suddivisa in quattro campi contenenti:

1. la scheda informativa dell'argomento contenente le informazioni principali sull'argomento stesso
2. l'elenco degli articoli che parlano dell'argomento
3. il grafico che riporta l'andamento temporale dell'attivazione dell'argomento
4. la lista di altri argomenti correlati all'argomento in esame e la relativa ontologia

La scheda informativa sugli argomenti utilizza informazioni estratte automaticamente da Adige-500.000.

Nell'esempio in Figura 8 è mostrata la scheda che il sistema presenta all'utente che abbia cercato informazioni sull'argomento "strage di Beslan".

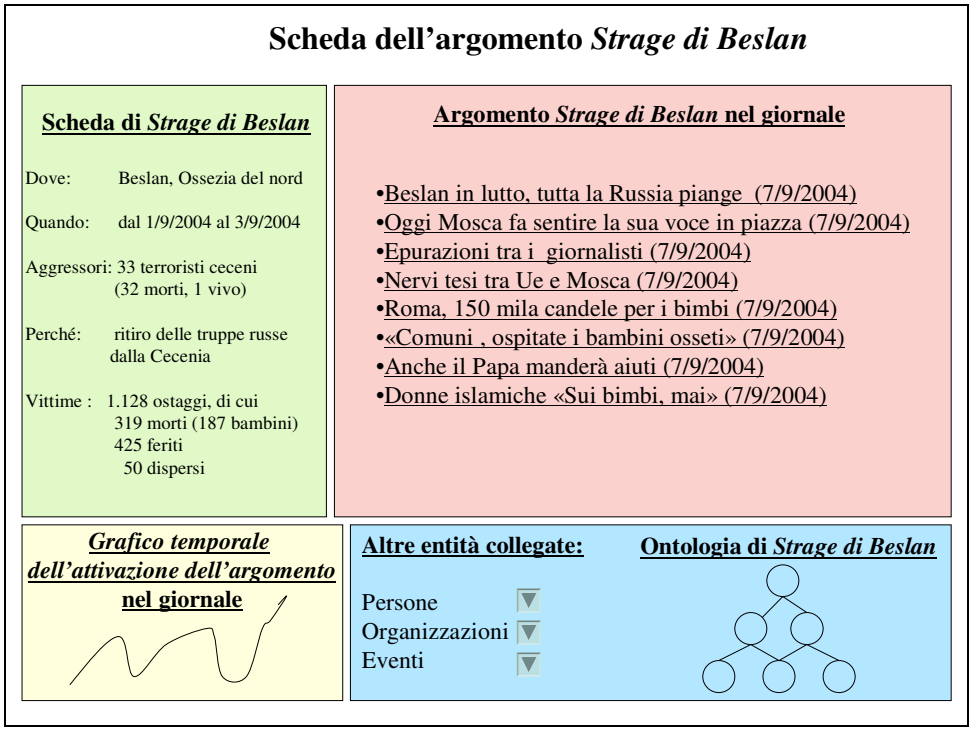

Figura 8: Scheda informativa di un argomento

4.3.3 Informazioni su una collocazione

Quella sulle collocazioni è un'informazione di tipo linguistico che viene presentata all'utente in una apposita scheda strutturata in modo parallelo a quella per le entità. Anche la scheda per le collocazioni è suddivisa in quattro campi contenenti:

1. la definizione della collocazione (tratta dal database lessicale MultiWordNet o da un dizionario)
2. le concordanze della collocazione. Ogni concordanza è collegata all'articolo da cui è stata estratta.
3. accesso al citografo relativo alla collocazione
4. le relazioni semantiche relative al concetto e la porzione di gerarchia (navigabile) di MultiWordNet in cui si trova il concetto.

La scheda informativa sulle collocazioni utilizza informazioni estratte automaticamente da Adige-500.000.

Nell'esempio in Figura 9 è mostrata la scheda che il sistema presenta all'utente che abbia cliccato sulla collocazione "ribes rosso" (cfr. Figura 5).

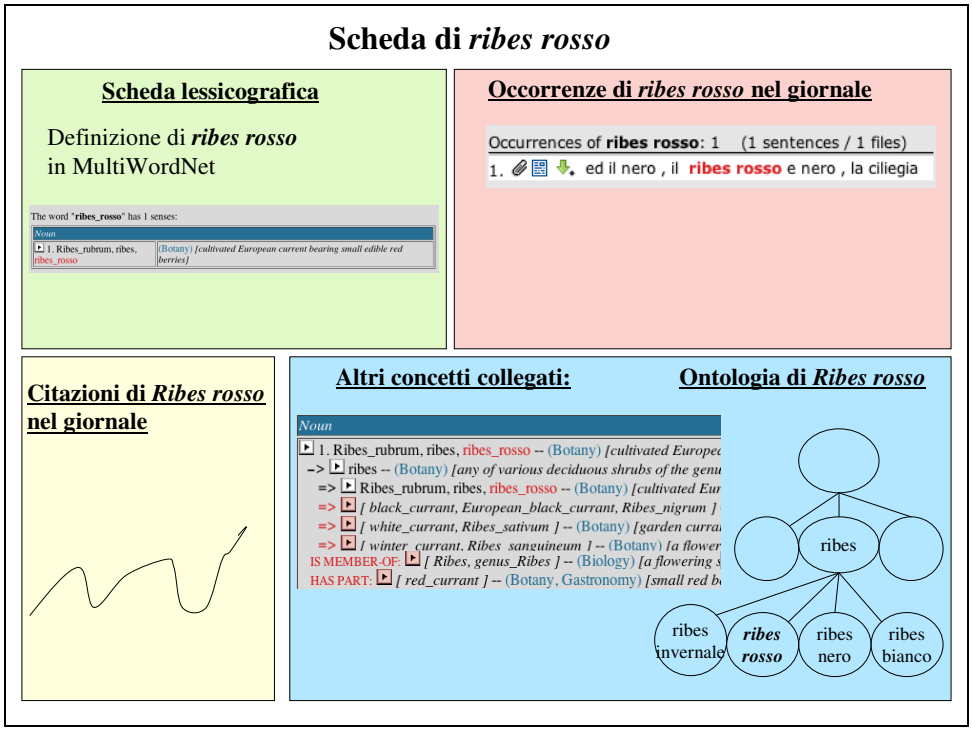

Figura 9: scheda informativa di una collocazione

4.3.4 Tutti i risultati

Come già visto nella sezione 4.2, in caso di query a parole chiave, è attiva la funzionalità "tutti i risultati", che dà all'utente la possibilità di visualizzare tutti i documenti che contengono le parole della sua query, indipendentemente dalle conoscenze possedute dal sistema.

Sono previste due modalità di visualizzazione di tutti i risultati della query. Se il sistema trova esattamente la parola o la frase inserita dall'utente, vengono restituiti i documenti sotto forma di *concordanze* di quella parola o frase. Se invece nei documenti vengono trovate le singole parole che compongono la query e non la query stessa così come è stata digitata il sistema restituisce *snippets* di documenti. La prima modalità è esemplificata in Figura 10, mentre la seconda in figura 11.

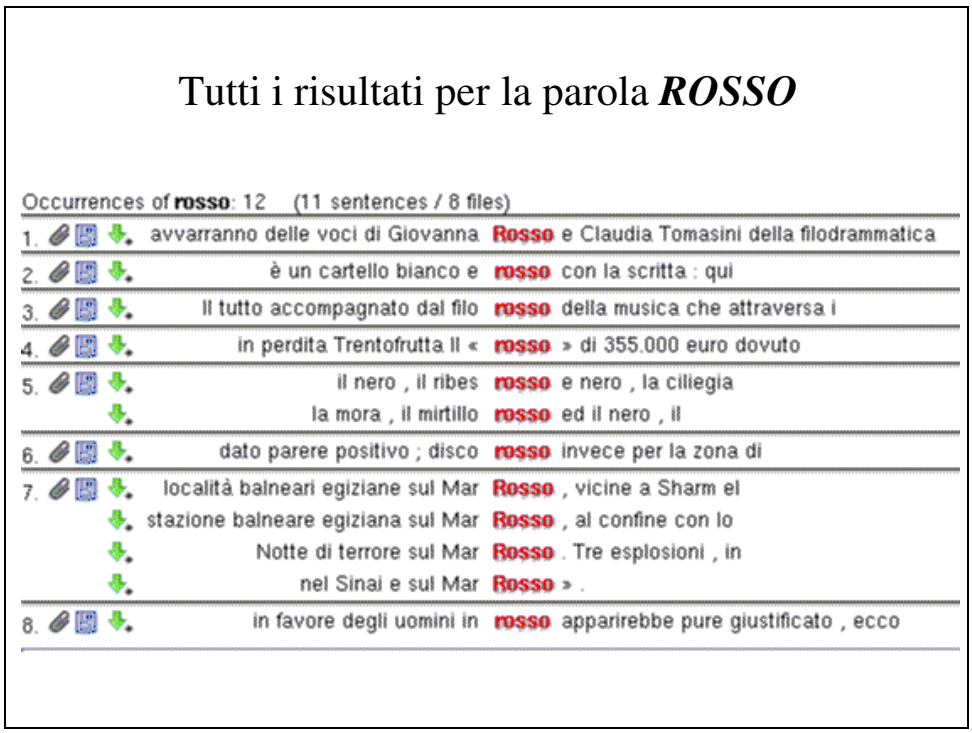

Figura 10: risultato della query per "rosso" (concordanza)

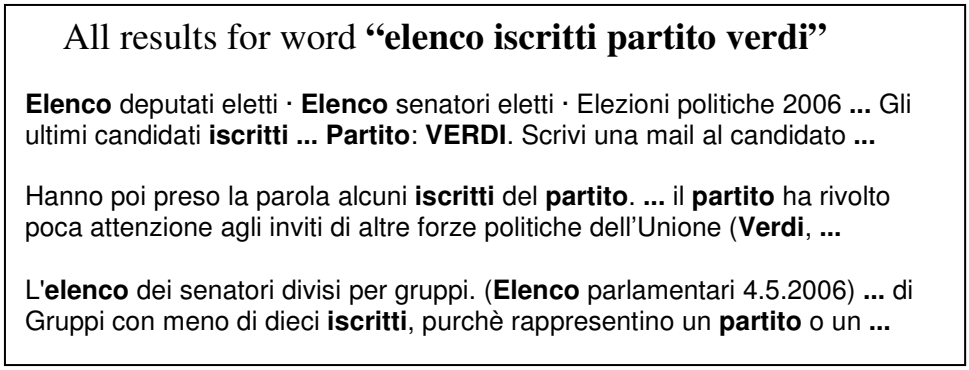

Figura 11: risultato della query per "elenco iscritti partito verdi" (snippets)

4.3.5 Accesso ai testi annotati

Una caratteristica fondamentale in Ontotext è la tracciabilità della conoscenza estratta automaticamente dai testi. Per questo motivo da ogni informazione fornita è possibile risalire alle fonti da cui è stata estratta. All'utente viene presentato l'articolo di giornale con evidenziata la porzione di informazione tramite cui si è arrivati a quel testo. È inoltre possibile visualizzare i diversi tipi di annotazione presenti nel documento: annotazione linguistica (sempre automatica) e annotazione semantica (manuale per il corpus I-CAB e automatica per il corpus finale Adige-500.000).

Da questa pagina è inoltre disponibile la funzionalità "mostra i testi simili" tramite la quale si accede direttamente ad altri articoli che parlano dello stesso argomento.

La Figura 12 mostra un esempio di testo annotato raggiunto partendo dalla scheda informativa di Margherita Cogo (cfr. Figura 7) in cui sono evidenziate tutte le occorrenze dell'entità di tipo Persona "Margherita Cogo".

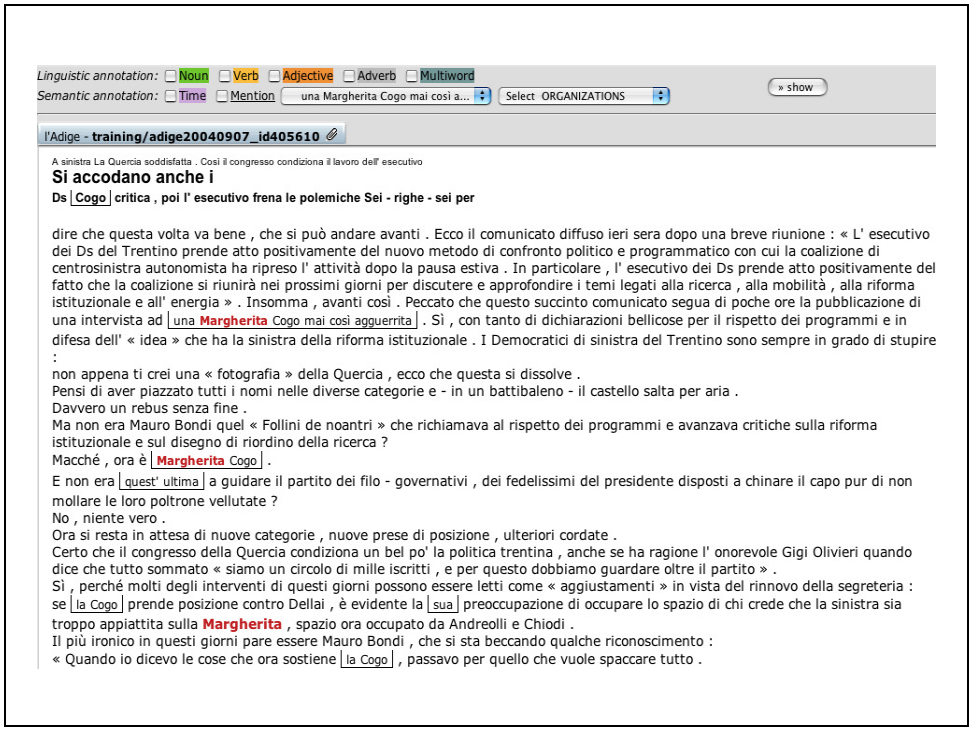

Figura 12: visualizzazione di un testo annotato

5 Prospettive future

Il progetto Ontotext è in corso di sviluppo. Per la fine del progetto sono previste ulteriori attività. Il corpus Adige-500.000 sarà interamente processato per estrarre tutta la conoscenza prevista in Ontotext, eventi compresi. Il database dei fatti (attualmente composto dalle informazioni estratte manualmente dal corpus I-CAB) sarà enormemente arricchito con la conoscenza estratta automaticamente dai 500.000 testi del corpus Adige-500.000. Il portale Ontotext rilasciato in versione dimostrativa per il 31 dicembre 2006 sarà aggiornato e arricchito durante tutto lo svolgimento del progetto fino ad arrivare alla versione definitiva con la fine del progetto stesso.

Bibliografia

L. Bentivogli, E. Pianta. Detecting hidden multiwords in bilingual dictionaries. In *Proceedings of the tenth EURALEX International Congress*, Copenhagen, Denmark, August 14-17, 2002, pp. 785-793

A. Esuli and F. Sebastiani. 2006. Determining Term Subjectivity and Term Orientation for Opinion Mining.In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).

B. Magnini, E. Pianta, O. Popescu, M. Speranza: Ontology Population from Textual Mentions: Task Definition and Benchmark, to appear in *Proceedings of the OLP-2 workshop, 2nd workshop on Ontology Learning and Population*, Sydney, Australia, 2006.

B. Magnini, E. Pianta, C. Girardi, M. Negri, L.Romano, M. Speranza, V. Bartalesi-Lenzi, and Rachele S.i. I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006* - 5th Conference on Language Resources and Evaluation, 22-28/5/2006, Genova (Italy).

Magnini B., Negri M., Pianta E., Romano L., Speranza M., and Sprugnoli R. From Text to Knowledge for the Semantic Web: the ONTOTEXT Project. In *Proceedings of SWAP 2005, Semantic Web Applications and Perspectives*, Trento, Italy, 14-15-16 December, 2005.

E. Pianta, L. Bentivogli, C. Girardi, B. Magnini. Representing and Accessing Multilevel Linguistic Annotation using the MEANING Format. *In Proceedings of the EACL-06 Workshop on Multi-dimensional Markup in Natural Language Processing (NLPXML-2006)*, Trento, Italy, April 4, 2006.

Saquete E., Martinez-Barco P., Munoz R., Negri M., Speranza M., and Sprugnoli R. Multilingual Extension of a Temporal Expression Normalizer using Annotated Corpora. In *Proceedings of the EACL 2006 Workshop on Cross-Language Knowledge Induction*, Trento, Italy, April 3, 2006.

Siti Web

- http://www.elpais.es/ (quotidiano El Pais)
- http://www.google.com (motore di ricerca GOOGLE)
- http://vivisimo.com (Vivisimo, clustering automatico di pagine web)