Annotazioni linguistiche: una rassegna

Mattia Gentilini

Relatore: prof. Fabio Vitali

Tesi di Laurea in Laboratorio di Tecnologie Web

27 marzo 2003

Annotazione Linguistica

- Notazione analitica o descrittiva applicata a risorse linguistiche
- **Risorsa Linguistica (LR)**: un qualsiasi contenuto linguistico, scritto, parlato, o non verbale (gesti, segni ecc.)
- In informatica si può presentare come testo, immagini, clip audio oppure video
- Le annotazioni linguistiche esistono da sempre, poiché legate alla comunicazione
- Il mondo di oggi è comunicazione globale, grazie a Internet, dunque le annotazioni sono destinate a diventare sempre più importanti
- Le annotazioni informatiche sono nate prima del WWW (1987-1990)

Approccio all'annotazione

- È possibile specificare vari tipi d'informazione attraverso le annotazioni
- È possibile una distinzione tra tre gruppi:
  - Fonetico/Ortografico
  - Sintattico
  - Strutturale/Semantico

Approccio Fonetico/Ortografico

- Notazioni per indicare la pronuncia e la scrittura corretta delle parole
- Informazioni esatte ed indiscutibili
- Uso di caratteri speciali (simboli dell'alfabeto fonetico, dieresi, lettere accentate, ecc.)
- Per questo motivo spesso si applica a lingue singole
- Usi: correttori ortografici e traduttori, sintetizzatori e analizzatori vocali, OCR

Approccio Sintattico

- Indicazione del ruolo delle parole nella frase (grammatica) oppure analisi della struttura della frase/periodo (logica)
- Non sempre le informazioni si possono fornire automaticamente
- Più “universali” delle annotazioni fonetico/ortografiche, non necessitano di caratteri particolari
- Usi: verificatori grammaticali (meno affidabili di quelli ortografici), supporto a traduttori e sintetizzatori vocali per aumentare la qualità

Approccio Strutturale/Semantico

- Molte interpretazioni possibili
  - Struttura dei costituenti della risorsa e loro significato (titoli, paragrafi, tabelle, liste…). Noto come markup
  - Significato di certe parti del documento, che forniscono informazione sul testo (Who, What, Where, When, How, Why negli articoli di giornale)
  - Fornitura di informazioni sulla risorsa nel suo complesso, indicate separatamente (meta-informazioni): data, autore, revisione ecc.
- Usi: markup strutturale per la presentazione dei documenti (es. XML), catalogazione per creare alberi navigabili di risorse e meta informazioni (Semantic Search Engines), migliori traduttori e sintetizzatori, applicazioni di IA

Legame Risorsa-Informazione

- È possibile dare più informazioni su di una risorsa, di vario tipo (analisi logica ***e*** semantica), dello stesso tipo ma di diversi fornitori (commenti politici su un certo discorso), oppure sequenziali (revisioni di un documento)
- Oppure si può legare la stessa annotazione a risorse diverse, descritte nello stesso modo (ad es. due suoni con la stessa pronuncia, due voci grammaticalmente equivalenti, due articoli dello stesso giornalista…)
- Inoltre è possibile segmentare una risorsa in più parti, per fornire informazioni diverse su ogni parte (punti, intervalli, parole, frasi, periodi ecc.)
- Altrimenti l'informazione riguarda la risorsa nel suo complesso (data, autore, revisione ecc.)
- Infine, di solito l'annotazione si applica a risorse esistenti
- Ma può anche accadere il contrario: le risorse servono da esempio per le annotazioni (dizionari), oppure sono create automaticamente (traduttori, sintetizzatori, IA)

Implementazione

- Risorsa: testo (soprattutto markup o testo semplice), immagine, audio, video (a volte contengono meta-informazioni). Il testo è di norma l'unico tipo di risorsa annotabile internamente
- Informazione: markup SGML/XML, markup non standard, markup interno alla risorsa. L'interoperabilità è determinante
- Meta-informazione: RDF/RDF Schema, Topic Maps. Dublin Core come set standard di informazioni
- Strumenti di gestione: alcuni sistemi possiedono strumenti e/o interfacce, spesso in Java, per creare e gestire risorse

Esempi di annotazione

- **TEI** – Text Encoding Initiative (1990 – TEI Consortium)
- **ATLAS** – Architecture and Tools for Linguistic Analysis Systems (2000 – NIST, LDC/University of Pennsylvania, MITRE)
- **CES** – Corpus Encoding Standard (2000 – MULTEXT, EAGLES, Vassar College)
- **DAISY** – Digital Accessible Information SYstem (1998 – DAISY Consortium)
- **BAS Partitur** – Bavarian Archive for Speech Signals (1996 – IPSK/Università di Monaco)

TEI

http://www.tei-c.org/

- Linguaggio di markup strutturale per risorse testuali
- Specializzato in diversi tipi di testo: Prosa, Poesia, Teatro, Discorso, Dizionari, DB terminologici
- Collegamenti di vario tipo, note, citazioni, bibliografia
- Iniziativa di lunga esperienza (1987), prima applicazione di SGML
- Passaggio ad XML in corso
- Nessuno strumento di gestione proposto da TEI

TEI: Esempio

Collezione di racconti brevi

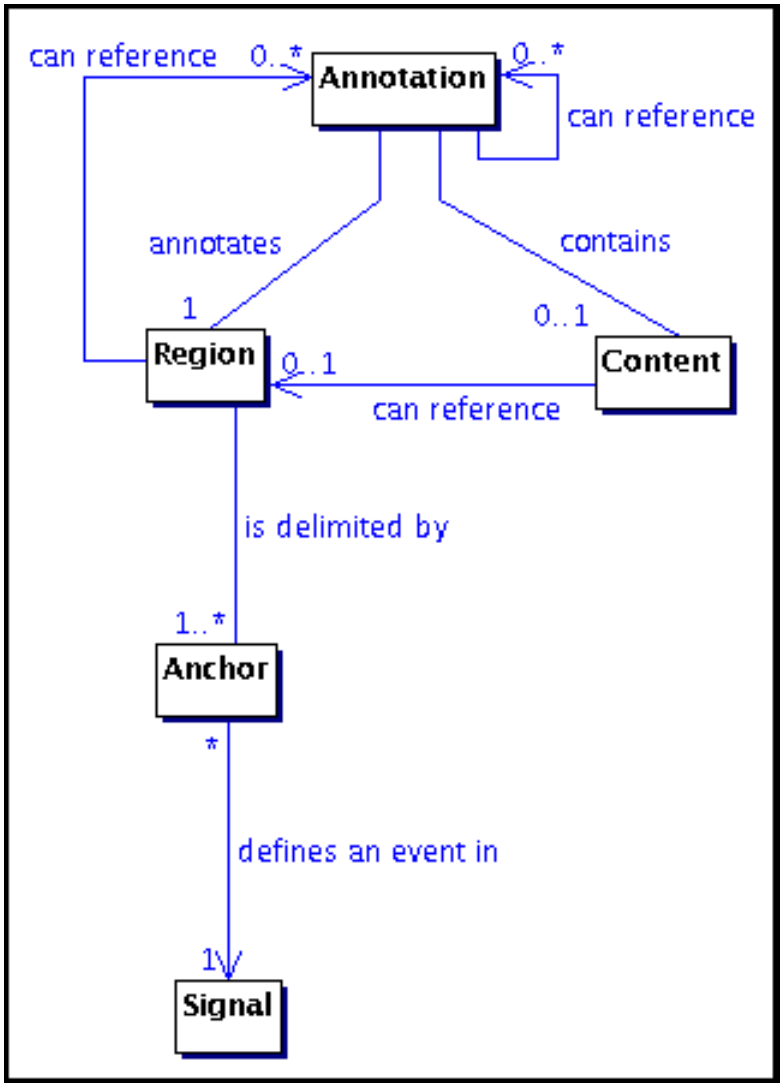
```
<TEI.2>
<teiHeader>
        <!-- header information for the whole collection -->
</teiHeader>
<text>
        <front>
                <docTitle><titlePart> The Adventures of Sherlock
                        Holmes </titlePart></docTitle>
                <docImprint>First published in <title>The
                        Strand</title> between July 1891 and
                        December 1892</docImprint>
                <!-- Any other front matter specific to the
                collection here ... -->
        </front>
        <group>
                <text>
                        <front>
                                <head rend="italic">Adventures of
                                        Sherlock Holmes</head>
                                <docTitle><titlePart>Adventure I.
                                        &mdash;</titlePart> <titlePart>A
                                        Scandal in Bohemia
                                        </titlePart></docTitle>
                                <byline>By A. Conan Doyle.</byline>
                        </front>
                        <body>
                                <p>To Sherlock Holmes she is always
                                <emph>the</emph> woman. [...] </p>
                        </body>
                </text> [...]
```

ATLAS

http://www.nist.gov/speech/ATLAS/

- Linguaggio di markup e API per gestire risorse di ogni tipo
- Approccio estremamente astratto, basato su Annotation Graphs
- Segnale, Àncora, Regione, Contenuto, Annotazione
- Infrastruttura di definizione tipi (MAIA)
- Meta-informazioni Dublin Core
- Linguaggio per esprimere interrogazioni (AQL)
- Implementazione XML (AIF)
- API per interfacciarsi ad applicazioni, formati di I/O e segnali
- Proposta ancora in via di definizione

ATLAS: Schema



ATLAS: Esempio

Commenti su un discorso politico

```
<Corpus id="USApresident0001" type="spoken" AIFVersion="1.1"
        xmlns="http://www.nist.gov/speech/atlas"
        xmlns:xlink="http://www.w3.org/1999/xlink">
        <Metadata>
                <dc:title>Discorso del presidente USA</dc:title>
                [...]
        </Metadata>
        <SimpleSignal id="audio1" type="audio" mimeClass="audio"
                mimeType="wav" encoding="PCM" xlink:type="simple"
                xlink:ref="USApresident0001.wav" track="ALL"/>
        <AnchorSet containedType="offset">
                <Anchor id="Aaudio1-intro" type="offset">
                        <SignalRef xlink:type="simple"
                        xlink:href="#xpointer(/Corpus/SimpleSignal[1])"/>
                        <Parameter type="offset" role="start"
                        unit="seconds">10</Parameter>
                        <Parameter type="offset" role="end"
                        unit="seconds">23</Parameter>
                </Anchor>
        </AnchorSet>
        <RegionSet containedType="interval">
                <Region id="Raudio1-intro" type="interval">
                        <AnchorRef xlink:type="simple"
                xlink:href="#xpointer(/Corpus/Anchorset[1]/Anchor[1])"/>
                </Region>
        </RegionSet>
        <Analysis id="comments" type="phrase" role="none">
                <AnnotationSet containedType="phrase">
                        <Annotation>
                                <RegionRef xlink:type="simple"
                                xlink:href="#xpointer(Region[id()="Raudio1-intro"])"/>
                                <Content><Feature role="news">
                                        <Parameter type="text" role="author"
                                        unit="none">CNN</Parameter>
                                        <Parameter type="text" role="content"
                                        unit="none">President wanted to...</Parameter>
                                </Feature>
                                <Feature>[...]</Feature></Content>
                        </Annotation>
                </AnnotationSet>
        </Analysis>
</Corpus>
```

CES

http://www.cs.vassar.edu/CES/

- Linguaggio di markup basato sulle Guidelines di TEI (quindi risorse testuali) e EAGLES (categorie grammaticali)
- Tre tipi di informazione
  - Documento: Markup strutturale
  - Annotazione: Grammatica di parole e frasi
  - Allineamenti: Paralleli fra testi simili (es. traduzioni in più lingue)
- Livelli di conformità del documento
- Sviluppato in SGML, passato poi ad XML per intero (Schema, XLink)
- Manca documentazione per il nuovo formato

CES: Esempio di annotazione

```
<!DOCTYPE cesAna PUBLIC "-//CES//DTD cesAna//EN“>
        <cesAna version="1.5" type="SENT TOK LEX DISAMB" doc=MyText1>
        <cesHeader version="2.3"> ... </cesHeader>
        <chunkList>
                <chunk doc="MyText1" from='1.2.1\1'>
                        <s>
                                <tok class='tok' from='1.2.1\1'>
                                        <orth>Les</orth>
                                        <disamb> <ctag>DMP</ctag> </disamb>
                                        <lex>
                                                <base>le</base>
                                                <msd>Da-fp--d</msd>
                                                <ctag>DFP</ctag>
                                        </lex>
                                        <lex>
                                                <base>le</base>
                                                <msd>Da-mp--d</msd>
                                                <ctag>DMP</ctag>
                                        </lex>
                                        <lex>
                                                <base>le</base>
                                                <msd>Pp3fpj-</msd>
                                                <ctag>PPJ</ctag>
                                        </lex>
                                        <lex>
                                                <base>le</base>
                                                <msd>Pp3mpj-</msd>
                                                <ctag>PPJ</ctag>
                                        </lex>
                                </tok> [...]
```

DAISY

http://www.daisy.org/

- Linguaggio di markup per la definizione di Digital Talking Books (DTB), strumenti pensati per garantire accessibilità delle risorse a persone con handicap di lettura
- Risorse audio (MP3, WAV) e testuali (XHTML)
- SMIL per sincronizzare le risorse
  - Modello temporale con durate e istanti di inizio/fine impliciti, espliciti, desiderati ed effettivi
  - Possibilità di paralleli, alternative, sequenze
- Navigation Control Center (NCC) in XHTML ridotto, per consultazione strutturale (a titoli e blocchi) e sequenziale (a pagine)
- Meta-informazioni Dublin Core e proprie

BAS Partitur

http://www.phonetik.uni-muenchen.de/Bas/BasFormatseng.html#Partitur

- Linguaggio per l'analisi di clip audio in formati definiti da BAS
- Segmentazione del segnale in parole
- Trascrizione fonetica in vari formati
- Ortografia con supporto ai caratteri speciali
- Analisi logica e grammaticale ad albero
- Elenco di proprietà (*tier*) associate a istanti, intervalli e parole
- Annotazioni scritte in testo semplice, non si utilizza markup SGML/XML

Verso uno Standard

- Ampia richiesta di Annotazioni (LR Community)
- Tante proposte diverse disponibili
- Nessuna ha raggiunto quel grado di accettazione diffusa tale da renderla standard
- Perché?
  - Molte proposte sono limitate nell'applicazione, oppure di recente introduzione e quindi non ancora conosciute
  - Manca una proposta stabile e conosciuta (come TEI), adattabile ad ogni situazione (come ATLAS).

La sfida di ISO

- Le annotazioni, pur avendo la rete come futuro, devono soddisfare applicazioni anche esterne ad essa
- ISO ha un'esperienza di standard che copre tutti i campi dell'attività umana, e ha già fatto proposte nel campo delle annotazioni (TMF/GMT), dunque è più adatta di W3C ed IETF per proporre uno standard
- Sono stati creati da cinque WG all'interno del TC37/SC4, per lo studio di risorse ed annotazioni. Di questi il WG1 è quello avente il compito più generale di standardizzazione
- Obiettivi: analisi delle proposte esistenti, creazione di un'architettura astratta, definizione di categorie di contenuto e meccanismi di istanziazione

Requisiti dello Standard

- Innanzitutto è bene rifarsi ad altri standard esistenti (in particolare XML e sua suite), decidendo cosa richiedere e cosa consigliare
- Rappresentare ogni tipo di annotazione, definendo formalmente le categorie
- Standard esplicito, possibilità di verificare la correttezza a prescindere dal contenuto
- Informazioni incomplete o errate, dilazionate nel tempo e di diversa granularità
- Uniformità della politica e dei meccanismi
- Adattabilità, in particolare verso le proposte esistenti

Conclusioni

- Le annotazioni linguistiche hanno una grande importanza:
  - Esistono da sempre
  - Ampia fascia di utenti, a vari livelli
  - Ampia varietà di proposte, per numero e usi
  - Soddisfano esigenze sentite da molto tempo (accessibilità per tutti e duratura)
  - ISO ha sentito la necessità di standardizzare. Il lavoro è appena iniziato