Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

Caso di studio e GATE

Rocco Tripodi

rocco@unive.it

I-CAB 1

I-CAB (Italian Content Annotation Bank)
- Corpus italiano costituito da articoli di giornale
- Categorizzazione degli articoli
- Attualità, Cultura, Economia, Sport, Cronaca di Trento
- Annotazione manuale
- Annotazione semantica
- Espressioni temporali
- Persone
- Organizzazioni
- Entità geo-politiche
- Luoghi

I-CAB 2

Annotazione automatica delle informazioni di basso livello
tokenizzazione, lemmatizzazione, collocazioni, POS Tagging

Annotazione *stand-off*
Separazione tra il testo e le annotazioni
Separazione dei differenti livelli di annotazione
(Benveniste, Jackenoff)
Creazione di applicazioni modulari
Indipendenza di ogni modulo di annotazione (Persone, Luoghi, ecc.)

Annotazione conforme allo *standard* TEI

Gerarchizzazione dei livelli
1. Annotazione ortografica (token identificati dalla posizione dei caratteri)
2. Annotazione morfo-sintattica (informazioni collegate ai token)
3. Multi-world (collegate alle parole)

Annotazione di alto livello

Annotazione di alto livello

Obiettivi del corpus

Adagiare le informazioni su una struttura ontologica
annotare documenti con informazione semantica e relazionale,
facilitare la interoperabilità da tale informazione
estendere le ontologie per le annotazioni del Web Semantico

Espressioni temporali (TE)

Annotate in base allo standard TIMEX2
Sono considerate TE sia periodi temporali
<tre anni>, <due settimane>, ecc

che espressioni puntuali
<oggi>, <6 maggio 2003>, ecc

che espressioni temporali vaghe come
<recentemente>, <prossimamente>, ecc

Annotazione temporali

TIMEX2 prevede una forma specifica da dare ai valori temporali tramite *val*

Es: <6 maggio 2003> → …val="2003-05-06"…

Es: <3 anni> → … val="P3Y"…

Ulteriori attributi

MOD: usato per i modificatori temporali. Valori possibili sono:

APPROX <verso mezzanotte>

MORE THAN (<più di 3 ore>)

START (<i primi anni '70>)

ANCHOR VAL: contiene la forma normalizzata di TE che funge da ancora

ANCHOR DIR: direzione di una TE, con valoti AFTER e BEFORE.

Es: *sarò in vacanza per <due mesi>*

VAL="P2M" ANCHOR_VAL= "2004-05-06" e ANCHOR DIR="AFTER"

SET: identifica le TE relative a *set of times*, ripetersi di azioni.

Es: <ogni anno> è annotata con SET="YES".

Annotazione delle entità

Etichette delle entità

PER – ORG – LOC – GPE – MIX

Ogni entità a sua volta viene accompagnata da ulteriori marche semantiche che specificano il tipo di riferimento dell'espressione:
SPC (Specific referential)
Es: <L' [avvocato] di Giovanni> ha vinto la causa
GEN (Generic referential)
Es: <Gli [avvocati]> non lavorano gratis>
USP (non generici né specifici)
Es: <100.000 [persone]>)
NEG (Negative)
Es: <Nessun [avvocato]>

La menzione copre l'intero sintagma nominale dal quale si estrae l'entità

Caratteristiche sintattiche

NAM: nomi propri (<Totti>, <UE>)

NOM: costrutti nominali (<i [bambini] buoni>, <l' [azienda]>);

PRE: pre-modificatori (il <brasiliano> Ronaldo)

BAR: costrutti nominali non introdotti da pre-modificatori ed articoli (<[poliziotti] di quartiere>)

HLS: costrutti nei quali la testa nominale non è espressa (<Il più [forte] di tutti>);

WHQ: pronomi interrogativi e relativi (<Chi> è lì?)

PRO: pronomi, personali (<tu>) e indefiniti (<qualcuno>);

PTV: partitivi (<[alcune] delle scuole>);

APP: costruzioni appositive (<il Po, fiume italiano>)

ARC: costruzioni appositive con una relativa adiacente (<L'ex direttore, Rossi, che faceva spesso tardi>.

Articolazione delle entità 1

Entità persona

INDIVIDUAL singolo individuo <George W. Bush>
GROUP: più di una persona <la tua [famiglia]>, <Alice e suo figlio>
INDEFINITE: dal contesto non è possibile giudicare se si tratta di una più persone <Mi chiedo [chi] arriverà>

Entità Organizzazione

GOVERNMENT <I [Carabinieri]>
COMMERCIAL <La [Microsoft]>
EDUCATIONAL <l'[Università di Pisa]>
MEDIA <National Geographic>
RELIGIOUS <La [Chiesa Valdese]>
SPORTS <La [Juventus]>
MEDICALSCIENCE <Il [laboratorio] di analisi>
NONGOVERNMENTAL <La [Croce Rossa]>
ENTERTAINMENT <La [compagnia] teatrale>
MIXED entità costituite da gruppi di ORG con diverso sottotipo

Articolazione delle entità 2

Entità geo-politiche

regioni geografiche caratterizzate dalla presenza di gruppi sociali e/o politici.
CONTINENT (<Asia>)
NATION (<Italia>, <Stati Uniti>)
STATE-OR-PROVINCE (<Florida>)
COUNTY-OR-DISTRICT (<Canton Ticino>)
POPULATION-CENTER (<Trento>)
GPE-CLUSTER (gruppi di GPE <Unione Europea>)
SPECIAL (GPE a cui è difficile applicare un'etichetta <La [Palestina]>)
MIXED (<Gli Stati Uniti e l'UE>)

Per questo tipo di menzioni si deve annotare anche il ruolo previsto per l'entità

Ci si riferisce al Governo o alla popolazione?

Allo stesso modo si deve annotare la metonimia. Per esempio quando si utilizza la capitale di uno Stato per riferirsi al Governo della nazione

Riferimenti

Endofora: espressione che si riferisce a qualcosa nello stesso testo

Anafora: co-riferisce con qualcosa che la precede (già espressa)

Es: Francesca non è andata a lavorare, lei era al parco

Catafora: co-riferisce con qualcosa che segue (ancora non espresso)

Es: Un giovane matematico, Kurt Gödel, rivoluzionò la logica a 25 anni

Exofora: co-riferisce con qualcosa che non è stato espresso

Es: Lei era al parco

Omofora: riferisce con qualcosa dedotto dal contesto

Es: The Queen

GATE: General Architecture for Text Engineering

Meta – programma

GATE Developer (Graphical User Interface)

Sviluppato in Java dall'Università di Sheffield dal 1995

Usato per lavori nel campo del NLP e Information Extraction

Composto da una serie di plugins che si possono attivare e disattivare per comporre applicazioni personalizzate

I plugins consentono di compiere analisi su diversi livelli del linguaggio dalla morfologia alla semantica - dalla tokenizzazione al recupero delle entità

Sono presenti plugins che consentono di caricare e gestire ontologie o interrogare motori di ricerca

Le annotazioni sono costruite tramite JAPE (Java Annotation Patterns Engine) Linguaggio che consente di creare regole per l'annotazione (espressioni regolari)

GATE Teamware: piattaforma di collaborazione