ITALIAN CONTENT ANNOTATION BANK (I-CAB):

Named Entities

Bernardo Magnini*, Emanuele Pianta*, Manuela Speranza*,
Valentina Bartalesi Lenzi**, and Rachele Sprugnoli**[1]

** FBK-irst, Povo 38050 (Trento) Italy*
{magnini | pianta | manspera }@fbk.eu

*** CELCT, Trento 38100 Italy*
{sprugnoli | bartalesi}@celct.it

August 2011

The content of the document is the result of the discussion which has taken place within the working group of FBK-irst and CELCT composed by Valentina Bartalesi Lenzi, Christian Girardi, Bernardo Magnini, Matteo Negri, Emanuele Pianta, Lorenza Romano, Manuela Speranza, Rachele Sprugnoli, Roberto Zanoli.

[1] This work has been supported by the ONTOTEXT (From Text to Knowledge for the Semantic Web) project, funded by the Autonomous Province of Trento under the FUP-2004 research program.

TABLE OF CONTENTS

Abstract

This document reports on the annotation of Named Entities for the Italian Content Annotation Bank (I-CAB) being developed at FBK-irst in conjunction with CELCT.

I-CAB is a corpus of Italian news annotated with semantic information at different levels. The first level is represented by Temporal Expressions, the second level is represented by different types of Entities (both Named and not-Named), and the third level is represented by Relations between Entities (e.g. the affiliation relation connecting a person to an organization).

For the annotation of I-CAB, we followed a policy of reusing already available mark-up languages. In particular, we adopted the annotation schemes developed for the ACE Entity Detection and Time Expressions Recognition and Normalization tasks for English. In this document we describe the annotation of Named Entities.

1. INTRODUCTION

This report presents the annotation of Named Entities in the Italian Content Annotation Bank (I-CAB), a corpus of semantically annotated documents for Italian containing annotations of Temporal Expressions (Lavelli et al. 2005), of different types of Entities (both Named and not-Named) and of a number of selected Relations among such Entities.

Following a policy of reusing already available mark-up languages, the annotation activity has been carried out adopting the formalisms developed within the American ACE program. However, due to the differences between English and Italian, part of the work has been dedicated to the revision and adaptation to Italian of the annotation guidelines.

The creation of I-CAB is part of the three-year project Ontotext funded by the Autonomous Province of Trento. Ontotext focuses on the study and development of innovative knowledge extraction techniques to produce new or less noisy information to be made available for the Semantic Web. Within the new research area of Ontology-Based Knowledge Extraction, Ontotext addresses three key research aspects: annotating documents with semantic and relational information, providing an adequate degree of interoperability of such relational information, and updating and extending the ontologies used for Semantic Web annotation. The concrete evaluation scenario in which algorithms will be tested with a number of large-scale experiments is the automatic acquisition of information about people from newspaper articles.

This document is structured as follows. Section 2 describes the annotation of Person Entities (PER), Organization Entities (ORG), Geo-Political Entities (GPE) and Location Entities (LOC), giving plenty of examples. Section 3 presents some special cases of annotation. Appendix A presents data about the inter-annotator agreement.

2. ANNOTATION OF NAMED ENTITIES

Annotators should tag all Named Entities within a document and, for each Entity, they identify its extent and semantic type. The full extent of a Named Entity consists of the entire proper name of the Entity. In case of ambiguous structures, the extent annotated should be the maximal extent.

In the ACE project, seven semantic types of Entities were identified (LDC 2005, p. 4):

- Person: a single individual or a group of humans.
- Organization: corporations, agencies, and other groups of people defined by an established organizational structure.

- Geo-Political Entity: geographical regions defined by political and/or social groups (e.g. a nation, its region, its government, or its people).
- Location: geographical Entities such as geographical areas and landmasses, bodies of water, and geological formations.
- Facility: buildings and other permanent man-made structures and real estate improvements.
- Vehicle: physical devices primarily designed to move an object from one location to another.
- Weapon: physical devices primarily used as instruments for physically harming or destroying other Entities.

In I-CAB we have restricted our annotation to four of the semantic types defined above:

- Person (PER)
- Organization (ORG)
- Geo-Political Entity (GPE)
- Location (LOC)

Nicknames

Entities have to be tagged also when they are mentioned by a nickname. This happens more often with Person Entities:

[*Pinturicchio*]$_{\text{PER}}$ *sta giocando bene*

Il [*Pupone*]$_{\text{PER}}$ *gioca sempre nella Roma*

Lo chiamano [*Little John*]$_{\text{PER}}$

[*Aladino*]$_{\text{PER}}$ *non è il vero nome, glielo hanno appiccicato*

However, the same holds for all other types of Named Entities (e.g. ORG):

Il [*Carroccio*]$_{\text{ORG}}$ *ha molti sostenitori nell'Italia Settentrionale*

Il [*Toro*]$_{\text{ORG}}$ *è in testa alla classifica*

Acronyms and abbreviations

Acronyms are annotated as Named Entities.

L'[*ONU*]$_{\text{ORG}}$ *compie 60 anni il 24 ottobre*

When names of Entities appear in abbreviated form, they are still annotated. In the following example, we consider "Università di Trento" as an abbreviated form of "Università degli Studi di Trento".

L'[*Università ii Trento*]$_{\text{ORG}}$ *avrà un nuovo rettore*

Meta-information

We do not use any special tag to mark meta-information. As a consequence, when the names of the journalists who authored the news stories appear inside the text (be it their full name or just their initials), they are simply tagged as PER.

Titles of books, CDs and exhibitions

Names of Entities appearing in book or CD titles, in names of organizations or events, etc., are not annotated.

2.1 Person Entities

Person Entities include both individuals and groups of people (including family names).

[*Laura*]$_{\text{PER}}$ *vive a Roma*

[*Carlo Azeglio Ciampi*]$_{\text{PER}}$ *è nato nel 1920*

I [*Lunelli*]$_{\text{PER}}$ *non sono originari di Pisa*

Il presidente [*Napolitano*]$_{\text{PER}}$ *è di origine campana*

Papa [*Giovanni Paolo II*]$_{\text{PER}}$ *ha viaggiato molto*

L'avvocato [*Rossi*]$_{\text{PER}}$ *si è trasferito di recente*

[*Sophia Loren*]$_{\text{PER}}$*, famosa attrice italiana, è apparsa ieri in televisione*

[*Luigi Rossi*]$_{\text{PER}}$ *in qualità di produttore ha realizzato molti film*

Una nuova Lazio con [*Toni*]$_{\text{PER}}$ *punta centrale*

L'ex direttore, [*Rossi*]$_{\text{PER}}$*, che faceva spesso tardi, è stato licenziato*

Quella ragazza si chiama [*Francesca*]$_{\text{PER}}$

[*Fassino*]$_{\text{PER}}$ *(Ds),* [*Rutelli*]$_{\text{PER}}$ *(Margherita) e* [*Fini*]$_{\text{PER}}$ *(AN)*

2.2. Organization Entities

Each organization named in a document is annotated as a Named Entity of type Organization (ORG). Organization Entities must be a formally established association. (LDC 2005, p. 7).

Typical examples are:

- Government organizations, i.e. organizations (also military organizations) that are connected to the structure or affairs of the government of a state.

  La [*Camera*]$_{\text{ORG}}$ *è composta da 630 membri*

  L'auto è stata posta sotto sequestro dalla [*Guardia di Finanza*]$_{\text{ORG}}$

- Commercial organizations, i.e. organizations that are focused primarily upon providing ideas, products, or services for profit.

  Stabile [*L'Espresso*]$_{\text{ORG}}$ *(+0,13%), brillante* [*Mediaset*]$_{\text{ORG}}$ *(+1,22%)*

  Sono Rita e Tamara che gestiscono il bar [*Stazione*]$_{\text{ORG}}$

- Educational organizations, i.e. institutions that are focused primarily upon the promulgation of learning/education.

  L'[*Università di Pisa*]$_{\text{ORG}}$ *avrà un nuovo rettore*

- Entertainment organizations, i.e. organizations whose primary activity is entertainment.

  I [*Rem*]$_{\text{ORG}}$ *presenteranno il loro nuovo album*

- Non Governmental organizations, i.e. several types of organizations whose main role is advocacy, charity or politics (in a broad sense):
    - (Para-)Military organizations

    Nel muro di silenzio delle nuove [*Brigate Rosse*]$_{ORG}$ *si è finalmente aperto uno squarcio*
    - Political parties

    Fini ([*AN*]$_{ORG}$*) si è incontrato ieri con Bertinotti (*[*Rifondazione Comunista*]$_{ORG}$*)*
    - Professional Regulatory

    È stato appena eletto il nuovo presidente dell'[*Ordine degli Avvocati*]
    - Charitable and no-profit organizations

    L'iniziativa è stata promossa da «[*Un ponte per…*]$_{ORG}$*»*

    Gli [*Amici della Neonatologia Trentina*] *organizzano un incontro in mattinata*
    - International Regulatory and Political Bodies

    Il cardinale sottolinea il ruolo che deve avere l'[*Onu*]$_{ORG}$
    - Labor and industrial unions

    Arrivato l' ok della [*CGIL*]$_{ORG}$ *al piano industriale 2005 - 2008*

    È andato a buon fine l'accordo con [*Assoartigiani*]$_{ORG}$
- Media organizations, i.e. organizations whose primary interest is the distribution of news or publications. They can be private or public.

    Arrestato l'inviato di [*Al Jazira*]$_{ORG}$
- Religious organizations, i.e. organizations that are primarily devoted to issues of religious worship.

    L'[*Arcidiocesi di Trento*]$_{ORG}$ *propone un nuovo convegno ecumenico*

    Era presente anche Igor Vyzhanov del [*Patriarcato Ortodosso di Mosca*]$_{ORG}$
- Medical science organizations are those whose primary activity is the application of medical care or the pursuit of scientific research. They can be private or public.

    La rassegna è stata promossa dal [*CNR*]$_{ORG}$

    Il giovane venne soccorso dall'elicottero del [*118*]$_{ORG}$
- Sports organizations, i.e. organizations that are primarily concerned with governing or participating in organized sporting events. They can be professional, amateur, or scholastic.

    Ho colto al volo l'opportunità di giocare in [*A1*]$_{ORG}$

    [*Juve*]$_{ORG}$ - [*Roma*]$_{ORG}$ *1 – 1*

    [*Luna Rossa*]$_{ORG}$ *è in testa alla classifica*

More examples:

La [*Microsoft Corporation*] *ha sede a Redmond, USA*

La [*Trentino Trasporti*] *e altre ditte trentine sono in crescita*

Il [*Gruppo Folkloristico Canazei*] *terrà un'esibizione in piazza*

L'[Associazione Pranic Healing] *presenta la propria attività*

Una nuova [*Lazio*] *con Toni punta centrale*

Gli [*Outline*]*, il gruppo musicale che si è appena esibito, sta avendo un enorme successo*

[*Margherita*] *e* [*Ds*] *propongono nuovi emendamenti*

Nuovo corteo di [*CISL*] *e* [*UIL*] *che scendono in piazza*

[*Izvestia*] *è un quotidiano autorevole nel panorama russo*

La chiameranno [*Fondazione Kessler*]

Il [*Santa Chiara*]*, l'ospedale cittadino, apre un nuovo reparto*

La [*Roma*]*, padrona di casa, che ha nettamente vinto, ha contestato l'arbitraggio*

La [*Serie A*] *inizia oggi il suo campionato*

La [*Brigata Friuli*] *(300 uomini) tornerà presto in Italia*

[*Inter*] *(30 punti),* [*Palermo*] *(27) e* [*Roma*] *(26)*

Foreign organizations have been annotated if they were the literal translation of the original name, whereas they have not been annotated if they were considered a cultural transposition of the concept expressed by the original word. Following this rule, *Dipartimento di Stato Americano* is annotated as ORG since it is the direct translation of *U.S. Department of State*. On the other hand, *Federazione Olandese* is not annotated as a Named Entity because the official name of the Dutch football association is *Koninglijke Nederlandse Voetbalbond*, whose literal translation would be *Lega Calcio Olandese*.

2.3. Geo-Political Entities

Geo-Political Entities (GPE) are composite entities comprised of a population, a government, a nation (or province, state, county, city, etc.), and a physical location (LDC 2005, p. 13).

We annotate different types of Geo-Political Entities:

- Continents ([*Asia*])
- Nations ([*Italia*], [*Stati Uniti*])
- States and provinces ([*Florida*], [*Toscana*], [*Alto Adige*])
- Counties and districts ([*Canton Ticino*])
- Population centers ([*Trento*])
- Groups of Entities which act like Geo-Political Entities ([*Unione Europea*])

In the case of *Nickname Metonymy*, a special case of metonymy which occurs when the proper name of a Geo-Political Entity is used to refer to an organization, we follow the ACE-LDC guidelines by annotating that Entity as ORG (LDC 2005, p. 33). The most common example of *Nickname Metonymy* is when the name of a Geo-Political Entity is used to refer to a sport team:

La [*Russia*]$_{ORG}$ *ha conquistato la medaglia d'oro*

Another example of *Nickname Metonymy* is when the name of a Location is used to refer to an organization. For example, the address of the headquarters of an organization can be used to refer to the organization itself, as in the following sentence, where *Via Segantini* refers to the organization *Federazione Trentina della Cooperazione*:

[Via Segantini]$_{ORG}$ *non è d'accordo*

2.4. Location Entities

Location Entities (LOC) are places defined on a geographical or astronomical basis which are mentioned in a document and do not constitute a political entity (LDC 2005, p. 19).

These include, for example:

- Addresses ([*Via Nazionale 12*])
- Celestial bodies (*Il pianeta* [*Venere*])
- Water-Bodies (*Il* [*Po*], *Il Mar* [*Mediterraneo*])
- Natural land regions (*Il Monte* [*Bondone*])

3. SPECIAL CASES

3.1 Person versus no Named Entity annotation

Names of people are not annotated if they refer to something which is not a person. For instance, in *la legge Bossi-Fini* or *il governo Prodi*, the names do not identify persons anymore (i.e. the relation with the person has been lost), but they are used respectively as the name of the law and as the name of the government themselves.

On the other hand, when the names of people are introduced by a preposition, they are annotated as Named Entities of type PER:

La legge di [*Bossi*] *e* [*Fini*]
Il governo di [*Prodi*]

3.2 Organization versus no Named Entity annotation

Political Parties

In general, political parties are Entities of type ORG. However, when the text mentions the people belonging to the party instead of the party itself, we have two cases, depending on the context:

- if the text refers to the party as a whole, or to its directives, we have an ORG
  I [*Ds*]$_{ORG}$ *hanno votato contro l'emendamento*
- if the text refers to a specific subset of the people belonging to the party (especially if they behave somehow differently from the others) the entity is not annotated

  I Verdi hanno abbandonato l'aula
  In this example, the text refers to the people of the green party who were present in the chamber and left it.

Sport teams

If the text refers to sport teams in general, their management or their administration, they are annotated as ORG; if instead the text refers to the players, they are not annotated.

La maglietta della [*Juventus*]$_{\text{ORG}}$ *è in vendita nei negozi specializzati*
La Juventus ha fatto due gol ieri

Police and similar organizations

If the text refers to police organizations as a whole, they are annotated as ORG; if instead the text refers to the single persons in a special activity, they are not annotated.

Cambio al vertice nella [*polizia*]$_{\text{ORG}}$
La polizia ha perquisito l'abitazione

Please notice that, *polizia francese* is not annotated as a Named Entity because it is considered as a cultural transposition of the concept expressed by *Gendarmerie* and not as its literal translation (see how we deal with foreign organizations, Section 2.2).

APPENDIX A: Inter-annotator agreement

Inter-annotator agreement for Person, Organization, Location and Geo-Political Entities has been evaluated on the dual annotation of four different sets of ten news stories randomly chosen from I-CAB (see Table 1). We have used the Dice coefficient computed as in (1), where *C* is the number of common annotations (i.e. both annotator have identified the same extent for an Entity), while *A* and *B* are respectively the number of the Named Entities annotated by the first and the second annotator.

(1) $Dice=2C/(A+B)$[2]

	Tokens	**Dice**	**Annotator 1**	**Annotator 2**
PER	5,406	0.96	114 Named Entities	111 Named Entities
ORG	3,878	0.84	62 Named Eentities	57 Named Entities
GPE	4,581	0.97	65 Named Entities	65 Named Entities
LOC	5,660	0.89	9 Named Entities	9 Named Entities

Table 1. Data about inter-annotator agreement.

[2] Please notice that the Dice coefficient has the same value as the F1-Measure computed considering one of the two annotators as reference.

REFERENCES

- Bernardo Magnini, Matteo Negri, Emanuele Pianta, Manuela Speranza and Rachele Sprugnoli. *Italian Content Annotation Bank (I-CAB): Temporal Expressions (V. 2.0)*. ITC-irst Technical Report, September 2006.
  On-line: http://ontotext.fbk.eu/Publications/TIMEX2_V2.0-TR.pdf
- (LDC 2005) Linguistic Data Consortium, *Automatic Content Extraction English Annotation Guidelines for Entities*, version 5.6.1 2005.05.23.
  On-line: http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v5.6.1.pdf

WEB SITES

- ACE: http://www.nist.gov/speech/tests/ace/index.htm
  http://www.ldc.upenn.edu/Projects/ACE/
- Ontotext: http://ontotext.fbk.eu/