THE ANNOTATION OF NAMED ENTITIES IN KIND

(KESSLER ITALIAN NAMED-ENTITIES DATASET)

A guideline for Italian Named Entities annotation

Version 1.0

TABLE OF CONTENTS

1. INTRODUCTION

The present text displays some revised guidelines for the annotation of named entities based on the ones extrapolated by the Italian Content Annotation Bank (I-CAB), a corpus of semantically annotated documents for Italian containing annotations of different types of mentions of entities. The present proposal aims at simplifying the current annotation's guidelines in favor of more intuitive and general ones.

2. ANNOTATION OF NAMED ENTITIES

Annotators should tag all named entities within a document and, for each entity, they have to identify its extent and semantic type. The full extent of a named entity consists of the entire proper name of the entity.

In the present text we restrict our annotation to three semantic types (instead of the seven previously identified) which are:

2.1. <u>Person</u>: a single individual or a group of humans.

2.2. <u>Organization</u>: corporations, agencies, and other groups of people defined by an established organizational structure.

2.3. <u>Location</u>: geographical entities such as geographical areas and landmasses, bodies of water, and geological formations or geographical regions defined by political and/or social groups.

Nicknames

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

Il [Pupone]$_{B-PER}$ gioca sempre nella Roma

Degno di nota è il [Rivoltoso]$_{B-PER}$ [Sconosciuto]$_{I-PER}$

However, the same holds for all other types of Named Entities:

Il [Toro]$_{B-ORG}$ è in testa alla classifica

Lo stesso Cuomo ha detto male dell'operato del sindaco della [Grande]$_{B-LOC}$

[Mela]$_{I-LOC}$

Acronyms

Acronyms are annotated as Named Entities.

L'[ONU]$_{B-ORG}$ compie 60 anni il 24 ottobre

[ATAC]$_{B-ORG}$ approva il bilancio 2019

Meta-information

We do not use any special tag to mark meta-information. Therefore, 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's titles are annotated only when they refer to a real place, character or group or to fictitious ones which refer without any doubt to a person, a location, or an organization, as we can see in the following examples:

Diedero vita recentemente a "[Aspettando]$_{O}$ [Godot]$_{B-PER}$", il classico dell'assurdo

Cosa ne pensi di "[Vita]$_{O}$ [di]$_{O}$ [Galileo]$_{B-PER}$" di Brecht?

Il ruolo di [Serebriakov]$_{B-PER}$ in "[zio]$_{O}$ [Vanja]$_{B-PER}$" di [Cechov]$_{B-PER}$

2.1 Person Entities

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

[Laura]$_{B-PER}$ vive a Roma

[Carlo]$_{B-PER}$ [Azeglio]$_{I-PER}$ [Ciampi]$_{I-PER}$ è nato nel 1920

Il presidente [Napolitano]$_{B-PER}$ è di origine campana

Papa [Giovanni]$_{B-PER}$ [Paolo]$_{I-PER}$ [II]$_{I-PER}$ ha viaggiato molto

[Sophia]$_{B-PER}$ [Loren]$_{I-PER}$, famosa attrice italiana, è apparsa ieri in televisione

Quella ragazza si chiama [Francesca]$_{B-PER}$

L'artista [Christo]$_{B-PER}$ è morto ieri a New York

L'agente ha tenuto il ginocchio sul collo di [Floyd]$_{B-PER}$, soffocandolo

As we can see from the examples above, appositions such as *[papa], [presidente], [sindaco], [arcivescovo]* as well as functions and title, generally placed before the proper name of a person, do not have to be annotated.

We also annotate animals when they have a proper noun, such as in:

La morte di [Priscilla]$_{B-PER}$, il suo bulldog di 3 anni

Partecipa la Torre con [Tremendo]$_{B-PER}$ su [Guess]$_{B-PER}$ (where Guess is the horse e Tremendo is the jockey)

2.2. Organization Entities

Each organization named in a document is annotated as a named entity of type Organization (ORG). Organization's entities must be a formally established association. 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 and the government itself.

La [Camera]$_{B-ORG}$ è composta da 630 membri

L'auto è stata posta sotto sequestro dalla [Guardia]$_{B-ORG}$ [di]$_{I-ORG}$ [Finanza]$_{I-ORG}$

Il [governo]$_{B-ORG}$ [italiano]$_{I-ORG}$ si è espresso a favore

Stiamo parlando con il [consolato]$_{B-ORG}$ [americano]$_{I-ORG}$ per risolvere la situazione

Permettendogli così di informare l'[ambasciata]$_{B-ORG}$ [americana]$_{I-ORG}$

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

I ricavi di [Zoom]$_{B-ORG}$ sono incrementati del 169%

In this case, we are talking about Zoom as a society, so it has to be annotated as ORG, but in the case the context refers to Zoom as the application has not to be annotated, as we can see in the following instance:

Le iscrizioni ad applicazioni di videoconferenze (come la stessa Zoom, Google Meet etc.)

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

L'[Università]$_{B-ORG}$ [di]$_{I-ORG}$ [Pisa]$_{I-ORG}$ avrà un nuovo rettore

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

I [Rem]$_{B-ORG}$ 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):

- o Political parties

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

- o Professional Regulatory

È stato appena eletto il nuovo presidente dell'[Ordine]$_{B-ORG}$ [degli]$_{I-ORG}$ [Avvocati]$_{I-ORG}$

- o Charitable and no-profit organizations

La ONLUS [Cooperazione]$_{B-ORG}$ [Internazionale]$_{I-ORG}$ [Solidarietà]$_{I-ORG}$ [Sanitaria]$_{I-ORG}$ ([C.I.S.S.]$_{I-ORG}$) ha effettuato una missione umanitaria

- o International Regulatory and Political Bodies

Il cardinale sottolinea il ruolo che deve avere l'[Onu]$_{B-ORG}$

- o Labor and industrial unions

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

Media organizations, i.e. organizations whose primary interest is the distribution of news or publications, such as newspapers, tv news, magazines, websites and TV and radio channels.

Si tratta di Domenico Quirico de [La]$_{B-ORG}$ [Stampa]$_{I-ORG}$

Bruno Girodon, cameraman per [France]$_{B-ORG}$ [2]$_{I-ORG}$

Ha distrutto file per il sito di [Wikileaks]$_{B-ORG}$

L'hanno detto al [TG1]$_{B-ORG}$

Può scegliere la diretta di [Rai]$_{B-ORG}$ [2]$_{I-ORG}$

Realizzata da Niccolò Caranti per [Sanbaradio]$_{B-ORG}$

Presentazione del libro sul sito [teopol.it]$_{B-ORG}$

Religious organizations, i.e. organizations that are primarily devoted to issues of religious worship.

Le posizioni della [Chiesa]$_{B-ORG}$ non hanno alcuna collocazione politica

fa riferimento alla [Curia]$_{B-ORG}$ [Vescovile]$_{I-ORG}$ dell'[Arcidiocesi]$_{B-ORG}$ di Bologna

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]$_{B-ORG}$

Il giovane venne soccorso dall'elicottero del [118]$_{B-ORG}$

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]$_{B-ORG}$

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

Brand names in general and means of transport with brand name referring to the industrial or commercial organization.

Viaggiavano su una [Fiat]$_{B-ORG}$ punto

Creò anche la linea [Chicco]$_{B-ORG}$ per [Artsana]$_{B-ORG}$

More examples:

La [Microsoft]$_{B-ORG}$ [Corporation]$_{I-ORG}$ ha sede a Redmond, USA

L'[Associazione]$_{B-ORG}$ [Pranic]$_{I-ORG}$ [Healing]$_{I-ORG}$ presenta la propria attività

Una nuova [Lazio]$_{B-ORG}$ con Toni punta centrale

Gli [Outline]$_{B-ORG}$, il gruppo musicale che si è appena esibito, sta avendo un enorme successo

[Margherita]$_{B-ORG}$ e [Ds]$_{B-ORG}$ propongono nuovi emendamenti

Nuovo corteo di [CISL]$_{B-ORG}$ e [UIL]$_{B-ORG}$ che scendono in piazza

La chiameranno [Fondazione]$_{B-ORG}$ [Kessler]$_{B-ORG}$

Presidente del [consiglio]$_{B-ORG}$ [di]$_{I-ORG}$ [amministrazione]$_{I-ORG}$ della [Fondazione]$_{B-ORG}$ [Teatro]$_{I-ORG}$ [alla]$_{I-ORG}$ [Scala]$_{I-ORG}$

2.3. 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.

These include, for example:

Addresses *([Via]$_{B-LOC}$ [Nazionale]$_{I-LOC}$ [12]$_{I-LOC}$, [Autostrada]$_{B-LOC}$ [M1]$_{I-LOC}$)*

Celestial bodies and the universe *(Il pianeta [Venere]$_{B-LOC}$, siamo solo un piccolissima parte dell'[universo]$_{B-LOC}$)*

Water-Bodies *(Il [Po]$_{B-LOC}$, Il [Mar]$_{B-LOC}$ [Mediterraneo]$_{I-LOC}$)*

Natural land regions *(Il [Monte$_{B-LOC}$ [Bondone]$_{I-LOC}$)*

We also annotate as Locations (LOC) that composite entities comprised of a population, a government, a nation (or province, state, county, city, etc.), and a physical location (which have been previously annotated as Geo-Political Entities), such as:

Continents *([Asia]$_{B-LOC}$)*

Nations *([Italia]$_{B-LOC}$, [Stati]$_{B-LOC}$ [Uniti]$_{I-LOC}$, Si trova in Italia ([IT]$_{B-LOC}$))*

States, regions and provinces (including the origin code) *([Florida]$_{B-LOC}$, [Toscana]$_{B-LOC}$, [provincia]$_{B-LOC}$ [di]$_{I-LOC}$ [Mantova]$_{I-LOC}$, [Alto]$_{I-LOC}$ [Adige]$_{I-LOC}$, [regione]$_{B-LOC}$ [del]$_{I-LOC}$ [Caucaso]$_{I-LOC}$, [Regello]$_{B-LOC}$ ([FI]$_{B-LOC}$)*

Counties and districts *([Canton]$_{B-LOC}$ [Ticino]$_{I-LOC}$, [quartiere]$_{B-LOC}$ [di]$_{I-LOC}$ [Monterosso]$_{I-LOC}$, [comune]$_{B-LOC}$ [di]$_{I-LOC}$ [Firenze]$_{I-LOC}$, [prefettura]$_{B-LOC}$ [di]$_{I-LOC}$ [Kagawa]$_{I-LOC}$, nel [cosentino]$_{B-LOC}$)*

Population centers *([Trento]$_{B-LOC}$)*

Groups of Entities which act like Location entities *([Unione]$_{B-LOC}$ [Europea]$_{I-LOC}$)*

Facilities *([Stadio]$_{B-LOC}$ [Olimpico]$_{I-LOC}$, [Torre]$_{B-LOC}$ [Eiffel]$_{I-LOC}$, [Muro]$_{B-LOC}$ [di]$_{I-LOC}$ [Berlino]$_{I-LOC}$)*. Facilities have to be annotated only if they have a proper name:

[ospedale]_{B-LOC} *[Torregalli]_{I-LOC}* is annotated, while *ospedale di* [Firenze]_{B-LOC} is not (only *[Firenze]* is annotated), or *[Teatro]_{B-LOC}* *[degli]_{I-LOC}* *[Arcimboldi]_{I-LOC}* since *[Teatro]* is part of the name, while in *cinema [Odeon]_{B-LOC}* we only annotate *[Odeon]*.

Restaurants, bar, pub: (*[Chalet]_{B-LOC}* *[delle]_{I-LOC}* *[rose]_{I-LOC}*)

Other cases (*stazione di [Roccasecca]_{B-LOC}, [stazione]_{B-LOC} [Termini]_{B-LOC}; [Duomo]_{B-LOC} [di]_{I-LOC} [Milano]_{I-LOC}, [Abbazia]_{B-LOC} [di]_{I-LOC} [Westminster]_{I-LOC}; [carcere]_{B-LOC} [di]_{I-LOC} [Regina]_{I-LOC} [Coeli]_{I-LOC}, aeroporto di [Napoli]_{B-LOC}, [Aula]_{B-LOC}/[Sala]_{B-LOC} + proper noun, [chiesa]_{B-LOC} [di]_{I-LOC} [San]_{I-LOC} [Giovanni]_{I-LOC} [Battista]_{I-LOC}, [tunnel]_{B-LOC} [della]_{I-LOC} [Manica]_{I-LOC}, [Africa]_{B-LOC} [Australe]_{I-LOC}, [Europa]_{B-LOC} [Settentrionale]_{I-LOC}, Crisi del [Levante]_{B-LOC}, [Sud]_{B-LOC} [Italia]_{I-LOC}*)

It is possible to find a LOC entity contained in a prize name or in the name of a race/competition. We annotate them when the award ceremony or race/competition takes place in the location described, such as in *Premio [Roma]_{B-LOC}, Coppa [Italia]_{B-LOC}* or *Rally [Dakar]_{B-LOC}*.

In the case of *metonymic location names*, a special case of metonymy which occurs when the proper name of a location entity is used to refer to an organization, we annotate the entity as ORG. The most common example of *metonymic location name* is when the name of a location entity is used to refer to a sport team:

La [Russia]_{B-ORG} ha conquistato la medaglia d'oro

Another example of *metonymic location name* can also be the situation in which the address of the headquarters of an organization is 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]_{B-ORG} non è d'accordo

In general, we can say that the cases in which the LOC entity is treated as an ORG entity include all the situations in which the LOC entity is the agent of an action, such as in the following examples:

La [Lombardia]_{B-ORG} si dice contraria al provvedimento

La [Germania]_{B-ORG} si ritira dalla trattativa

3. SPECIAL CASES

3.1 Organization versus no Named Entity annotation

Police and similar organizations

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

- *Cambio al vertice nella [polizia]$_{B\text{-}ORG}$*
- *La polizia ha perquisito l'abitazione*
- *Un agente di [polizia]$_{B\text{-}ORG}$/il veicolo della [polizia]$_{B\text{-}ORG}$*
- *Ha assaltato a colpi di fucile una pattuglia dei [carabinieri]$_{B\text{-}ORG}$*
- *Simbolo della dedizione al lavoro dei [vigili]$_{B\text{-}ORG}$ [del]$_{I\text{-}ORG}$ [fuoco]$_{I\text{-}ORG}$*

Events or meetings

Names of meetings, reunions, assemblies or events have not to be annotated but there are some special cases in which the name of the meeting/event also represents the people/the organization which is there reunited, and so the entity has to be annotated as ORG, such as in the following case:

- *Il [G8]$_{B\text{-}ORG}$ si è espresso a favore del cambio di direzione*
- *Il G8 ha avuto luogo settimana scorsa*

3.2 Entities with multiple type annotations

There are some cases in which we have an ORG entity and a LOC or PER entity both contained in the same expression, as we can see in the following examples:

- *[Procura]$_{B\text{-}ORG}$ di [Trani]$_{B\text{-}LOC}$*
- *[polizia]$_{B\text{-}ORG}$ di [Gotham]$_{B\text{-}LOC}$ [City]$_{I\text{-}LOC}$*
- *[governo]$_{B\text{-}ORG}$ [Berlusconi]$_{B\text{-}PER}$*
- *[amministrazione]$_{B\text{-}ORG}$ [Bush]$_{B\text{-}PER}$*
- *[regime]$_{B\text{-}ORG}$ di [Assad]$_{B\text{-}PER}$*

It is also possible to find in the same expression an apposition/function/title referring to a person followed by an ORG entity. In this case only the ORG entity has to be annotated:

-presidente del [Consiglio]$_{B\text{-}ORG}$

-presidente della [Repubblica]$_{B\text{-}ORG}$

-direttore della [Scala]$_{B\text{-}ORG}$

-capo dello [Stato]$_{B\text{-}ORG}$

3.3 Lists of entities

We can also find lists of entities, in which different elements referring to one entity are listed, separated by commas or conjunctions. In these cases only the element which is immediately adjacent to the entity has to be annotated :

- *[Italia]$_{B\text{-}LOC}$ [settentrionale]$_{I\text{-}LOC}$ e meridionale*
- *X, XI e [XII]$_{B\text{-}LOC}$ [arrondissement]$_{I\text{-}LOC}$*

3.4 Person versus no Named Entity annotation

Names of people are not annotated if they do not explicitly refer to the person. For instance, in *legge Prodi*, the name of the law is always associated with the person of Prodi while in the case of *Premio Nobel* and *Scala Richter* the relation with the person has been lost and so the name is not annotated, as we can see below:

- *legge [Prodi]$_{B\text{-}PER}$*
- *Premio Nobel*
- *Scala Richter*
- *Premio Pulitzer*