EuroParl Named Entity Annotation Guidelines[1]

Version 1.0

Authors:
Andreea Bodnari
Pierre Zweigenbaum

Table of contents

[1] Guidelines inspired from the MUC-2007 and the 2011 i2b2 annotation guidelines, and Tim Finin et al. work in “Annotating named entities in Twitter data with crowdsourcing”.

1. Overview

Rationale The task of this project is to capture two layers of information about expressions occurring inside a document. The first layer captures expressions as they occur inside a document, based on their type. The second layer, the coreference layer, links together all expressions of a given type that are identical to each other.

Document Structure: This guidelines describe the specific type of information that should be annotated for named entity extraction and coreference resolution and provides examples similar to those that may be found in the EuroParl documents. The instances that should be marked along with the examples in the surrounding text that should be included in the annotations are described. Instances in this guideline marked in **BLUE** are correctly annotated named entities. Instances marked in **RED** are terms that should not be marked. Coreference pairs will be linked by a connecting line.

Annotation Tool: www.notableapp.com/

2. General Guidelines for Named Entities

1. **What things to annotate**
    a. Only complete noun phrases (NPs) and adjective phrases (APs) should be marked. Named entities that fit the described rules, but are only used as modifiers in a noun phrase should not be annotated.
        - Media-conscious David Servan-Schreiber was not the first ...
        - Deaths were recorded in Europe. Various European capitals ...

2. **How much to annotate**
    a. Include all modifiers with named entities when they appear in the same phrase except for assertion modifiers.
        - some of our Dutch colleagues
        - Committee on the Environment
        - no criminal court
    b. Include up to one prepositional phrase following a named entity. If the prepositional phrase contains a named entity by itself, but it is the first prepositional phrase following a named entity, then it is included as a prepositional phrase and not annotated as a stand-alone named entity.
        - President of the council of Ecuador
        - President of Ecuador
        - members of the latest strike
    c. Include articles and possessives
        - ***the*** European Union
        - ***an*** executive law
        - ***his*** proposed law

3. Categories of Entities

Concepts are defined in three general categories that are each annotated separately: Location, Organization, and Person. Named entities of other entity types should be ignored.
In general, an entity is an object in the world like a place or person and a named entity is a phrase that uniquely refers to an object by its proper name (“*Hillary Clinton”*), acronym (“*IBM”*), nickname (“*Opra”*) or abbreviation (“*Minn.*”).

3.1. Location

Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments. Compound expressions in which place names are separated by a comma are to be tagged as the same instance of Location (see “*Kaohsiung, Taiwan*”, “*Washington, D.C.*”). Also tag ‘generic’ entities like “*the renowned city*”, “*an international airport*”, “*the outbound highway*”.

3.2. Organization

Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure. Some examples are businesses (“*Bridgestone Sports Co.*”), stock ticker symbols (“*NASDAQ”*), multinational organizations (“*European Union”*), political parties (“*GOP”*) non-generic government entities (“*the State Department”*), sports teams (“*the Yankees”*), and military groups (the Tamil Tigers). Also tag ‘generic’ entities like “*the government*”, “*the sports team*”.

3.3. Person

Person entities are limited to humans (living, deceased, fictional, deities, ...) identified by name, nickname or alias. Include titles or roles (“*Ms.”*, “*President*”, “*coach”*) and names of family members (“*father*”, “*aunt*”). Include suffixes that are part of a name (*Jr.*,*Sr.* or *III*). There is no restriction on the length of a title or role (see “*Saudi Arabia’s Crown Prince Salman bin Abdul Aziz*”). Also tag ‘generic’ person expressions like “*the patient*”, “*the well-known president*”.

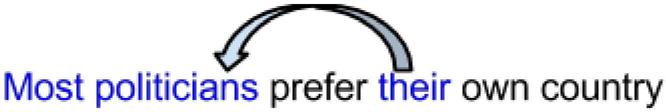
NOTE: some expressions tend to be ambiguous in the category to which they belong (see “*Paris”*, both the capital of France (Location) and a proper name (Person); “*Peugeot”*, both an organization (Organization) and a proper name (Person)). We ask that you specifically disambiguate those cases, and annotate the expression with the category best defined by the context in which it is used.

4. General Guidelines for Coreference Resolution

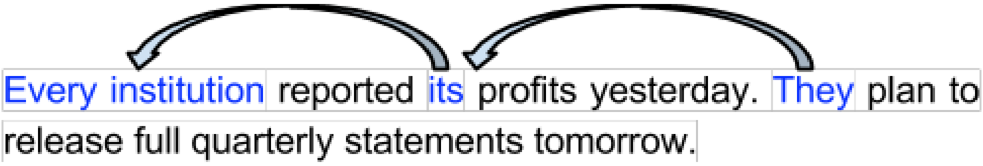
The general principle for annotating coreference is that two named entites are coreferential if they both refer to an identical expression. Only named entities of the same type can corefer. Named entities should be paired with their nearest preceding coreferent named entity.

NOTE: For ease of annotation, the pronouns in each document have been annotated. If a pronoun is involved in a coreference relation with a named entity annotated in step 1, then a coreference link should be created. See the examples below for when a pronoun should be linked to a named entity.

1. **Bound Anaphors** Mark a coreference link between a "bound anaphor" and the noun phrase which binds it.

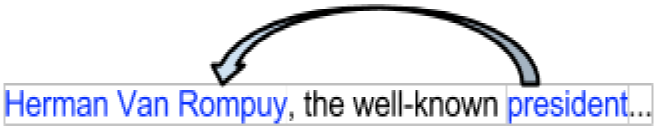

Most politicians prefer their own country

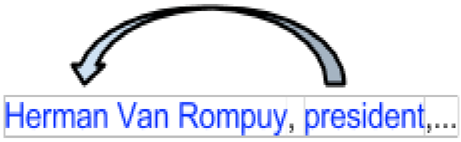
-

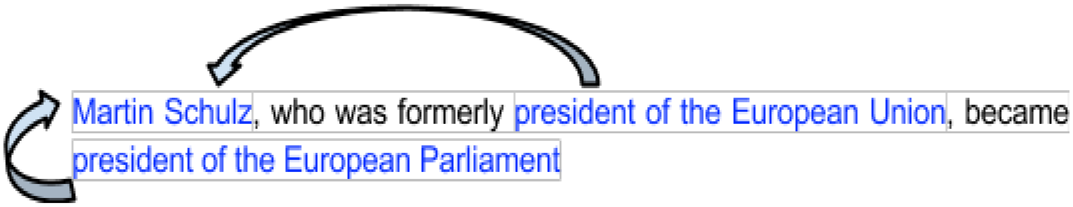

Every institution reported its profits yesterday. They plan to release full quarterly statements tomorrow.

-

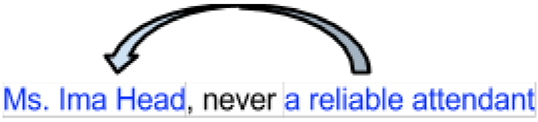
2. **Apposition** Typical use of an appositional phrase is to provide an alternative description or name for an object. In written text, appositives are generally set off by commas


Herman Van Rompuy, the well-known president...

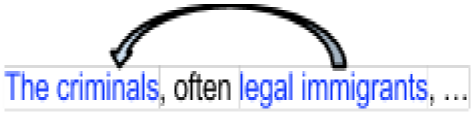

Herman Van Rompuy, president,...


Martin Schulz, who was formerly president of the European Union, became president of the European Parliament

Mark negated appositions:

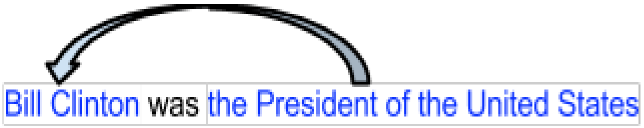

Ms. Ima Head, never a reliable attendant

Also mark if there is only partial overlap between the named entities:

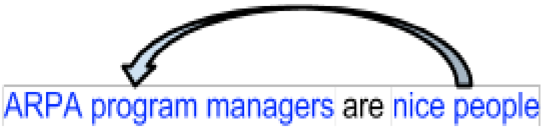


The criminals, often legal immigrants, ...

3. **Predicate Nominals and Time-dependent Identity** Predicate nominals are typically coreferential with the subject.



Bill Clinton was the President of the United States



ARPA program managers are nice people

Do **NOT** annotate if the text only asserts the possibility of identity

- Phinneas Flounder may be the dumbest man who ever lived.
- Phinneas Flounder was almost the first president of the corporation.
- If elected, Phinneas Flounder would be the first Californian in the Oval Office.

4.1. Coreference Annotation Arbitration

Each batch of documents will be annotated by two independent human annotators. The merged document batches will then will then undergo arbitration by a third annotator.