LegalNERo Annotation Guide

Purpose

This short guide offers definitions and examples on how to annotate entities for the LegalNERo corpus. LegalNERo is a corpus composed of legal-domain documents in Romanian language manually annotated for named entities. The annotation process will follow the general guidelines described in this document. Annotators will identify text spans associated with one of the entity types introduced in this document. For clarification purposes, meetings will be held regularly with the annotators. As needed, additional examples may be added to this document following the meetings.

Introduction

In data mining, a named entity is a phrase that clearly identifies one item from a set of other items that have similar attributes. General examples of named entities are first and last names, geographic locations, ages, addresses, phone numbers, companies and addresses. Depending on the domain of interest, different entity types may be defined. For the purposes of the LegalNERo corpus, we are considering only the following entities: Person (PER), Organization (ORG), Location (LOC), Time expressions (TIME), Legal references (LEGAL). These will be detailed in the next section.

In certain cases, named entities may be imbricated one into another. For example, the organization “Primăria Municipiului Brașov” also contains a potential location entity “Brașov”. For our purposes, we are not considering such sub-entities. The annotation will consider only the largest text span denoting an entity. For this example, only the organization entity will be annotated. As an exception to this rule, we will consider sub-entities present in the legal reference entities. This will allow us to construct automatic named entity recognition (NER) systems able to recognize entities with or without the legal reference class. Therefore, for entities like “Legea 176/2019” annotations will contain both the entire span as a legal reference entity and the text span “2019” as a time expression.

Annotation will focus on identifying the largest text span clearly identifying the name of a named entity, without including additional information. Examples (the named entity spans are indicated between [...] and in bold): “dr. ing. [**Ion Popovici**]” (en. “Dr. Eng. Ion Popovici”), “orașul [**București**]” (en. “Bucharest city”), “competiția se va desfășura în aproximativ [**3 luni**]” (en. the competition will take place in about 3 months). General words that do not denote the name of a named entity or a clear time expression or a resolvable legal reference will not be annotated, such as “primarul” (en. “mayor”), “satul” (en. “vilage”), “comuna” (en. “commune”), “parcul” (en. “park”), “peste un timp” (en. “after a while”), “o lege” (en. “a law”), etc.

Entity Types

Person (PER)

Person entities are regularly limited to humans. A person may be a single individual or a group. By extension, if a fictional character or a reference to religious figures is present it will be annotated. However, in legal-domain text we expect to encounter only person names. The annotation will consider only the largest text span clearly identifying the person's name. Additional words that may be present, such as titles, honorifics, functions will not be annotated as part of the entity name.

Even though there are no entity subtypes, the following mentions are considered entities of type PERSON:

a) Each distinct person or set of people mentioned in a document refers to an entity of type Person.
Eg.: Președintele Agenției Naționale de Administrare Fiscală, **[Mirela Călugăreanu]**. (en. President of the National Agency for Fiscal Administration, Mirela Călugăreanu.)

b) Saints and other religious figures, including references to “God”.
Eg.: **[Sfântul Ion]** se sărbătorește la începutul anului. (en. Saint John is celebrated at the beginning of the year)
[Dumnezeu] a făcut cerul și pământul. (en. God made the heavens and the earth.)

c) Fictional characters, names of animals, and names of fictional
Animals.
Eg.: **[Spider-Man]** este un personaj popular printre tineri. (en. Spider-Man is a popular character among young people.)

d) Family names

Eg.: **[Popeștii]** și-au cumpărat un teren lângă pădure. (en. Popeștii bought a plot of land near the forest.)

Are **not** considered entities of type PERSON: occupations “măcelarul” (en. “the butcher”), family relations “tatăl” (en. “dad”), pronouns “el” (en. “he”), titles and honorifics that precede the name, unmanned/general groups of people “familie” (en. “family”), “pictorii” (en. “painters”), etc..

Organization (ORG)

Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure. The annotation process will mark text spans clearly indicating the name of an organization. Additional words present in text will not be annotated if they are not part of the organization’s name. For example (annotated text spans are given in bold and between [...]): “compania [**SC Company SA**]” (en. “company SC Company SA”), “[**Compania Apa Brașov**]” (en. “Company Apa Brașov”). In the first case, the word “compania” is not part of the organization’s name and it was not annotated. In the second case, the same word is part of the second organization’s name and it was annotated. Since we are dealing with legal-domain text (and not fictional text), if in doubt the complete organization name can be looked up on Google to make sure the proper text span is considered.

Even though there are no entity subtypes, the following mentions are considered entities of type ORGANIZATION:

a) Government organizations relating to, or dealing with the structure or affairs of government, politics, or the state, political parties.
Eg.: Hotărârea a fost emisă de **[Guvernul României]**. (en. The decision was issued by the Romanian Government.)
[Camera de comerț a României] organizează cursuri acreditate. (en. The Romanian Chamber of Commerce organizes accredited courses.)

b) Commercial organizations and nonprofit organizations
Eg.: **[Societatea Economică Eurasică (EAEC)]** s-a întrunit săptămâna trecută. (en. The Eurasian Economic Society (EAEC) met last week.)
[Crucea Roșie] a fost prezentă după izbucnirea incendiului. (en. The Red Cross was present after the fire broke out.)

c) Educational organizations
[Universitatea București] organizează concursul de admitere la masterat pe 2 aprilie. (en. The University of Bucharest is organizing the master's admission competition on April 2.)

d) Other types of organizations such as: political parties, UN, EU, etc.

[Biserica Ortodoxă Română] sprijină refugiații din Siria. (en. The Romanian Orthodox Church supports Syrian refugees.)

Are **not** considered entities of type ORGANIZATION mentions like: “angajați” (en. “employees”), “echipaj” (en. “crew”), etc.

Location (LOC)

Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations, denoted by a proper name. The annotation process will identify the name associated with a location entity, without additional words, such as the location type, unless these words are part of the official entity name. Examples (annotated text spans are given in bold and between [...]): “[**Parcul Alexandru Ioan Cuza**]” (en. “Alexandru Ioan Cuza Park”), “orașul [**București**]” (en. “Bucharest city”). Since we are dealing with legal-domain text (and not fictional text), if in doubt the complete location name can be looked up on Google to make sure the proper text span is considered.

a) Continent (Africa), Nation (“**Germania**”, en.”Germany”), County (**Argeș**), City (**Curtea de Arges**), district (**Sector 4**)
b) political entity: “**Europa de Est**” (en. “Eastern Europe”), “**Orientul Mijlociu**” (en.”the Middle East”)
c) geographical areas and landmasses: “**Transilvania**” (en. Transylvania”)
d) bodies of water (**Olt**), and geological formations “**Carpații**” (en. “Carpatians”)

Are **not** considered entities of type LOCATION general mentions such as: “stradă” (en. “street”), “pod” (en. “bridge”), “oraș”(en. “city”), “aeroport” (en. “airport”), etc.

Legal References (LEGAL)

Legal references are designations (the title of a legal document) or expressions pointing to another legal document.

Are considered LEGAL references mentions such as:

Law: “**Legea fondului funciar nr. 18/1991**.” (en.Land fund law no. 18/1991.)
Decree: “**Decretul nr. 309/2020** pentru promulgarea legii.” (en. Decree no. 309/2020 for the promulgation of the law.)
Decision: “**Decizia nr. 892/2021** referitoare la respingerea excepţiei de neconstituţionalitate.” (en. Decision no. 892/2021 regarding the rejection of the exception of unconstitutionality.)

Convention: "**Convenţia Consiliului Europei privind coproducţia cinematografică (revizuită), din 19.01.2021**." (en. Council of Europe Convention on Film Co - Production (Revised) of 19.01.2021.)
Declaration: "**Declaraţia Parlamentului României nr. 1/2020** cu privire la situaţia din Irak." (en. Declaration of the Romanian Parliament no. 1/2020 on the situation in Iraq.)

Are not considered entities of type LEGAL general mentions such as: "lege" (en. "law"), "decret" (en. "decree"), "declarație" (en. "declaration"), etc.

Time Expressions (TIME)

Time expressions tell us when something happened, or how long something lasted, or how often something occurs. Sometimes the precise date cannot be determined, allowing for expressions indicating periods of time ("next year", "in the next three months").

Examples:

a) Complete or partial dates:
"3 ianuarie 2010" (en. "January 3, 2010"), "19.01.2020", "martie 2018" (en. "March 2018"), "2017".

b) Periods of time:
"Martie-aprilie" (en. "march-april"), "3 luni" (en. "3 months"), "următorii 2 ani" (en. "The next 2 years").

c) Specific holidays indicating a time period
"Crăciun" (en. "Christmas"), "Anul nou" (en. "New year"), "Paști" (en. "Easter").