EVALITA 2016 NEEL-IT Challenge

Annotation Guidelines (version 1.1)

Introduction

The task consists of three consecutive steps:
1) extraction and typing of entity mentions within a tweet;
2) linking of each mention to an entry in the canonicalized version of DBpedia 2015-10 representing the same real world entity, or NIL in the case such entry does not exist;
3) clustering of all mentions linked to NIL. Thus, an entity which does not have a corresponding entry in DBpedia will be always referenced with the same NIL identifier.

This document introduces various definitions relevant to this task and provides a summary of the guidelines we followed to generate our gold-standard.

Basic Concepts

An entity, in the context of NEEL-IT challenge, is used in the general sense of being, not requiring a material existence but requiring to be an instance of a taxonomy class. Thus, a mention to an entity in a tweet can be seen as proper noun or an acronym referring to an entity. The extent of an entity is the entire string representing the name, excluding the preceding definite article (i.e. “il”,“lo”,“la”,“il”,“gli”,“le”,“l’”),iompound any other pre-posed (e.g. “Sig.ra.”, “Sig.”,“Prof.”, etc.) or postposed modifiers and quotation marks. Compound entities should be annotated in isolations.

Mentions and Typification

In this task we consider that an entity may be referenced in a tweet as a proper noun or acronym if:
1) it belongs to one of the categories specified in the Taxonomy (see below);
2) it can be linked to a DBpedia entry or to a NIL reference given the context of the tweet.

Notice:

- Pronouns (e.g., lui, lei): they are not considered mentions to entities in the context of this challenge.
- Misspellings: lower cased words and compressed words (e.g. “c u 2night” rather than “see you tonight”) are common in tweets. Thus, they are still considered mentions if they can be directly mapped to proper nouns.
- Complete Mentions: entity extents that are substrings of a complete named entity mention are not identified. For example, from the following tweet: "*Chameleon Launcher in arrivo anche per smartphone: video beta privata su Galaxy Note 2 e Nexus 4: Chameleon Laun...*

http://t.co/1nCDV2ji", you could not select either of the words *Galaxy* or *Note 2* by themselves. This is because they constitute a substring of the full mention *[Galaxy Note 2]*. However, in the text: "*@OfficialAnto_77 Sei il migliore sia in campo sia nella vita!! Vai Luca sei un grande!! http://t.co/lj7MGULi*" the term *[Luca]* should be selected as entity mention.

- Overlapping mentions: nested entities with qualifiers should be considered as independent entities. For example:
  Tweet: " @PiazzapulitaLA7 metteteci le mani della polverini".
  In this case the *[LA7]* entity qualifies *[Piazzapulita]*, the annotation for such a case should be:
  [La7, Organization, http://dbpedia.org/resource/La7]
  [Piazzapulita, Product, NIL]
- Noun phrases: these are not considered as entity mentions.
  - Noun phrases referring to an entity: While noun phrases can be dereferenced to an existing entity, we do not consider them as entity mentions. In such cases we only keep "embedded" entity mentions. For example:
    "*@Crazytilde no nulla solo il nome. Ma Cappellacci **pres. PDL della Sardegna** è coinvolto in in varie inchieste nella sua regione."*
    In this case ***pres. PDL della Sardegna*** refers to a Person-type however since it is not a proper noun we do not consider it as an entity mention. For that reason in this case the annotation should only contain the embedded entity *[PDL della Sardegna]*: [PDL della Sardegna, Organization, nil]
- Entity Typing by Context: entity mentions in a tweet can be typified based on the context in which they are used. For example:
  "*Difendere la libreria [Hoepli] vuol dire conoscere #Milano http://t.co/c7MddrkW"*
  In this case *[Hoepli]* mention refers to a Location rather than to an Organization-type.

Special Cases in Social Media (#s and @s)

Entities may be referenced in a tweet preceded by hashtags and @s or composed by hashtagged and @-nouns:

- *@[fattoquotidiano] i cicli e i ricicli della storia: tutto cambia per non cambiare niente*
- If the mention contains the name and surname of a person, the name of a place, or an event, etc., it should be considered as a named entity:
  @[CarlottaFerlito]
  @[LiciaRocca]
  @[iltrenoromalido]
- The mention is composed by a part of a complete name of an entity and from the context it is possible to understand that it refers to an entity:
  @[Pierferdinando] troppo legati poi alla parola #matrimonio. I cattolici in politica = danni
- The name of an entity is a substring of the mention, in this case only the substring that identify the named entity should be annotated as an entity:
  @Senatore[Monti]

- Should not be considered as named entity those aliases not universally recognizable or traceable back to a named entity, but should be tagged as entity those mentions that contains well known aliases. Then, *@ValeYellow46* should not be tagged as is not an alias for Valentino Rossi:
  @ValeYellow46 Auguri anche a te campione .
  While, *@Saturnino69* should be annotated since it refers to a well recognized alias
  @[Saturnino]69 quanto è costato ad idinner farsi pubblicizzare da te su twitter?

Hashtags can refer to entities, but it does not mean that all hashtags will be considered entities. We consider the following cases:

- Hashtagged nouns and noun-phrases. In the tweet:
  "*#coseimbarazzantisudime appena trono da scuola mi vedo allo specchio per capire in che condizioni sono stata tutta la mattinata"*
  The hashtag *#coseimbarazzantisudime* does not represent an entity as defined in the Guidelines. Thus, it should not be given as an entity mention.
- Tagged entities: If a proper noun is splitted and tagged with two hashtags, the entity mention should be splitted into two separate mentions. For example:
  "*#Emilia #Romagna"*
  In this case the annotation should be [Emilia, Location, http://dbpedia.org/resource/Emilia-Romagna] [Romagna, Location, http://dbpedia.org/resource/Emilia-Romagna]

Note that the characters # and @ should not be included in the annotation string.

Use of Nicknames

The use of nicknames (i.e. descriptive names replacing the actual name of an entity) are commonplace in Social Media. For example the use of "Hazza" for referring to "Harry Styles". For these cases we coreferenced the nickname to the entity it refers to in the context of the tweet.

quandoci sono i fuochi,con [Hazza] dovremo comportarci come i cani alzare il volume della tv e parlare fortissimo così non sentirà i spari lol
[Hazza] -> http://dbpedia.org/resource/Harry_Styles

Knowledge Base

The 2016 NEEL-IT task will be based on the Italian DBpedia 2015-10[1] as the Knowledge Base for linking. Concepts should be annotated with the canonicalized dataset of DBpedia 2015.

[1] http://wiki.dbpedia.org/dbpedia-dataset-version-2015-10

DBpedia is a widely available Linked Data dataset and is composed by a series of RDF (Resource Description Framework) resources. Each resource is uniquely identified by a URI (Uniform Resource Identifier). A single RDF resource can be represented by a series of triples of the type <S,P,O> where S contains the identifier of the resource (to be linked with a mention), P contains the identifier for a property and O may contain a literal value or a reference to another resource.

In this challenge, a mention in a tweet should be linked to the identifier of a resource (i.e. the S in a triple). Note that in DBpedia there are cases where one resource does not represent an entity, instead it represents an ambiguity case (disambiguation resource), a category, or it just redirects to another resource. In this challenge, only the final IRI properly describing a real world entity (i.e. containing their descriptive attributes as well as relations to other entities) are considered for linking. Thus, if there is a redirection chain given by the property wikiPageRedirects, the correct IRI is the one at the end of this redirection chain.

Evaluation

Participants are allowed to submit up to 3 runs of their system as TSV files. An example of the submission format will be released with the development set. We encourage participants to make available their system to the community to facilitate reuse.

We will use the TAC KBP scorer[2] to evaluate the results and in particular we will focus on:

[tagging] strong_typed_mention_match (check entity name boundary and type)
[linking] strong_link_match
[clustering] mention_ceaf (NIL detection)

The strong_typed_mention_match evaluates the micro average F-1 score for all annotations considering the mention boundaries and their types. The strong_link_match is the micro average F-1 score for annotations considering the correct link for each mention. The mention_ceaf (Constrained Entity-Alignment F-measure) is a clustering metric developed to evaluate clusters of annotations. It evaluates the F-1 score for both NIL and non-NIL annotations in a set of mentions. The final score will be computed as follows:

$$score = 0.4 \quad mention\_ceaf + 0.3 \quad strong\_typed\_mention\_match + 0.3 \quad strong\_link\_match$$

[2] https://github.com/wikilinks/neleval/wiki/Evaluation

Gold Standard (GS) Generation Procedure

The GS was generated with the help of 3 annotators. The annotation process follows three phases.
In the first one, an unsupervised annotation of the GS was performed, with the intent to extract candidate links which were meant as inputs of the second stage.
In the second stage annotations were performed by two annotators using brat[3]. The annotators were asked to analyze the entity mentions, categories and links provided in the first stage and to add, remove any others. The annotators were also asked to mark any problematic case if encountered.
In the third phase, a third annotator went through the problematic cases and, involving the two initial annotators, refined the annotation procedures.
An iterative process has then taken place looping on stage 2 and 3, till mostly all problematic cases were resolved.

Data and Annotation Format

The dataset comprises two files: the data and the annotation file. The data file is a list of tweet ids, each listed on a different line. The annotation file consists of a line for each tweet id, which is followed by the start and the end offset 6 of the annotation, the linked concept and the category. All values are separated by the TAB character. For example, for the tweet:

"@CarlottaFerlito io non ho la forza di alzarmi e prendere il libro! Help me"

The following line in the annotation file should be produced:

tweet_id 1 16 http://dbpedia.org/resource/Carlotta_Ferlito Person

We randomly selected training and test data from the TWITA dataset[4]. Because some tweets may have been removed since their publication, we will release the original text of tweets to participants that will register at Evalita 2016 - NEEL-IT task.

Taxonomy

Thing

languages

[3] brat, http://brat.nlplab.org/
[4] Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 100–107, Atlanta, Georgia, June. Association for Computational Linguistics.

ethnic groups
nationalities
religions
diseases
sports
astronomical objects

Examples:
Oggi lanciamo la pagina [Vivere la Pieve] - una bacheca aperta a tutti i pievesi, nati e d'adozione.
… l'[IMU] dovrà andare ai Comuni

Event

holidays
sport events
political events
social events

Examples:
[Calciomercato]
[JuveUdinese]
[Biella Festival 2011]

Character

fictional character
comics character
title character

Examples:
@VaultProject tipo è la [sailor pluto] della tua bigioia.
La [Befana] a Milano Marittima...

Location

public places (squares, opera houses, museums, schools, markets, airports, stations, swimming pools, hospitals, sports facilities, youth centers, parks, town halls, theatres, cinemas, galleries, universities, churches, medical centers, parking lots, cemeteries)
regions (villages, towns, cities, provinces, countries, continents, dioceses, parishes)
commercial places (pubs, restaurants, depots, hostels, hotels, industrial parks, nightclubs, music venues, bike shops)
buildings (houses, monasteries, creches, mills, army barracks, castles, retirement homes, towers, halls, rooms, vicarages, courtyards)

Examples:

Sciara su tweet was born. [Grazie all'amico e collaboratore Domenico Spampinato].Sciara parla di [Belpasso], di belpassesi e su [Belpasso].

Nella lista dei potenti della [Terra]...

Organization

companies (press agencies, studios, banks, stock markets, manufacturers, cooperatives)

subdivisions of companies

brands

political parties

government bodies (ministries, councils, courts, political unions)

press names (magazines, newspapers, journals)

public organizations (schools, universities, charities)

collection of people (sport teams, associations, theater companies, religious order, youth organizations, musical band)

Examples:

Ho accettato la candidatura alla [Camera] per [Centro Democratico]. Un impegno per crescita,lavoro,politiche sociali,cultura e ricerca.

Milanisti sez mana ? RT\"@soalMILAN: Ale ale ale [milan] ale, forza lotta vincere, non ti lascremo mai #ForzaMilan

...[Juventus]...

“Caro Presidente, ma perché i prof vengono cambiati a metà dell’anno?” - @[la_stampa] http://t.co/6SmT6uMy

Person

people’s names (titles and roles are not included, such as Dr. or President)

Examples:291210843000041472

@[CarlottaFerlito]...

...[Elisabetta Preziosa]

Product

movies

tv series

music albums

press products (journals, newspapers, magazines, books, blogs)

devices (cars, vehicles, electronic devices)

operating systems
programming languages

Examples:

[iPasta], la app per cucinare anche Tablet e Smartphone: [IPasta] è un'applicazione per tablet che può essere utiliz... http://t.co/SHuqRcxD

'[Dove osano le aquile]' e Colditz, teatro di uno dei press tour più inutili di sempre con @MatSant una vita fa #sky

Quando non ho molta socialità con gli amici mi faccio una vita su [the sims] #coseimbarazzantisudime