#Microposts2014 Challenge on Named Entity Extraction & Linking (NEEL) - Annotation Guidelines (version 1.3)

Introduction

The NEEL task consists of two stages: 1) extraction of entities mentions within a tweet; and 2) linking of each of these entities to an English DBpedia v3.9 resource.
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

We consider the definition of entity in the general sense of being, in which an object or a set objects not necessarily need to have a material existence, but which however must be characterised as an instance of a taxonomy class.

In this task we consider that an entity may be referenced in a tweet as a noun or noun phrase if:

1) it belongs to one of the categories specified in the taxonomy (see appendix);

2) it is disambiguated by a DBpedia URI within the context of the tweet. This means that all the NIL entities (i.e. entities without a disambiguation URI) are not taken into account;

3) it subsumes other entities. This means that an entity phrase, which can be composed of two or more entities, is considered a single entity if it can be disambiguated by a DBpedia URI. Notice that the longest entity phrase with a DBpedia URI will have therefore precedence over shorter and single entities:

- *[Natural History Museum at Tring]*
- *[News International chairman James Murdoch]'s evidence to MPs on phone hacking*
- [*Sony*]'s [*Android Honeycomb*] Tablet

In the latter case, since there is no DBpedia URI for [[Sony]'s [Android Honeycomb]], it is splitted into it's embedded entities.

* Notice: In this task we do not consider pronoun mentions (e.g., he, him) as entities.

Numerical expressions

We only considered numerical expressions which:

- Refer to integer numbers
- Integer expressions separated by one “-” (e.g. 0-1)

Excluded:
- Numerical expressions accompanied by currency signs (e.g. £2) were excluded.
- Expressions with decimal numbers.
- Billion cases, since the DBpedia URI is not of type
- Expressions separated by “/” or “:”
- Expressions composed of number and words (e.g.. 1-year)

Linking Data Set

English DBpedia v3.9.

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:

- *#[Obama] is proud to support the Respect for Marriage Act*
- *#[Barack Obama] is proud to support the Respect for Marriage Act*
- *@[BarackObama] is proud to support the Respect for Marriage Act*

Use of Nicknames

Nicknames occur when a name of an entity is used to refer to another entity. For these cases we coreferenced the mention to the mention it refers to in the context of the tweet.

#[Panda] with 3 straight hits to give #[SFGiants] 6-1 lead in 12th
#[Panda] -> http://dbpedia.org/page/Pablo_Sandoval
#[SFGiants] -> http://dbpedia.org/page/San_Francisco_Giants

Gold Standard (GS) Generation Procedure

The GS was generated with the help of 14 annotators, who had different backgrounds including computer scientists, social scientists, social semantic web experts, semantic web experts, and linguists.

The annotation process followed three phases. In the first one, an unsupervised annotation of the GS has been performed, with the intent to extract candidate links that were meant as inputs of the next stage.

In the second one the dataset was divided into batches so as to assign three different annotators to each batch. In this phase annotations were performed using the CrowdFlower service (http://crowdflower.com/). The annotators were asked to analyze the links provided in the first stage and to add, remove any others. The annotators were also asked to mark any ambiguous case if encountered.

In the third phase, the adjudication stage, three annotators polished the collected annotations and generated the training GS. In particular three main actions took place:

1. cross consistency check of the entity types;
2. cross consistency check of the URIs;
3. resolution of ambiguous cases raised up by the 14 annotators.

The entire process will be further explained in a paper, together with the agreement scores reached so far.

Appendix

Taxonomy

Amount
Animal
    Bird
    Insect
Event
    MilitaryConflict
    PoliticalEvent
    SportEvent
    WeatherEvent
    MeetingEvent
    BreakingNews
Function
    Job
Location
    AdministrativeRegion
    Airport
    Bridge
    Canal
    City
    Continent
    Country
    Hospital
    Island
    Museum
    Lake

- Lighthouse
- Mountain
- Park
- Restaurant
- River
- Road
- ShoppingMall
- Stadium
- Station
- Valley

Organization
- Airline
- Band
- Broadcast
- Company
- EducationalInstitution
- Legislature
- NonProfitOrganisation
- RadioStation
- SoccerClub
- SportsLeague
- SportsTeam
- TVStation
- University
- PoliticalOrganisation

Person
- Ambassador
- Architect
- Artist
- Astronaut
- Athlete
- Celebrity
- ComicsCharacter
- Criminal
- FictionalCharacter
- Mayor
- MusicalArtist
- Politician
- SoccerPlayer
- TennisPlayer

Product
- Aircraft
- Album

- Automobile
- Book
- Drug
- EmailAddress
- Magazine
- Movie
- Newspaper
- OperatingSystem
- PhoneNumber
- ProgrammingLanguage
- RadioProgram
- SchoolNewspaper
- Software
- Song
- Spacecraft
- URL
- VideoGame
- Weapon
- Website

Time
- Holiday

Cardinal Direction
Language
Nationality
Numeric Expression
- Day of a month

Religion
Season
AstronomicalObject
- Planet
- Natural Satellite

EthnicGroup
Weather
Sport Name
AstrologicalSign