Boosting Eventus ID through Natural Lenguage Processing

Pedro Armengol

2/6/2018

Abstract

The Boosting Eventus ID project is a Natural Language Processing Pipeline designed to apply Name Entity Recognition (NER) to news texts in Spanish. The software uses label data elaborated by the Researcher Javier Osorio to fit neuronal models (using the SpaCy package) to predict 7 categories (Target, source, location and so on).The accuracy of the model, defined as exact match of class as well as word start and end location, is around 14% for all the classes and around 45% for the location class. Further extensions of this research will include Entity Relationship Extraction from the news.

Introduction

More than 2 million articules are published every year on the web. Those articules posses a universe of information about facts and trends, many times, not available in structured datasets (that are prone to be use for statistical analysis). Also, those articles are in thousands of lenguages, formats and styles making a very difficult task the standarization of their content into structured information. However, through natural lenaguage processing and statistical learning techniques, I used the features of each word to predict their class (target, source, locatio and so on): where every class could be compose of one or more words.

One first attempt to systematize the information in news was the creation of the Conflict and Mediation Events Observations (CAMEO) catalog. CAMEO is a coding scheme that catalog autors and events, specialized for automatized coding and for the register of sub-national authors (previous coding schemes where capable just to deal with national level authors (they where unable to catalog asymetrical warefare for example). The cited catalog has been used in several projects including the Integrated Conflict Early Warning System a DARPA – Lockheed Martin Project joint venture project.

One step further, the Global Database of Events, Language and Tone (GDETL) is an effort to systematize the global media through the implementation of worldwide collector of news in more than 28 lenguages and machine learning techniques to systematize their information using the CAMEO labels. The GDETL website (now powered by Jinsaw - Google) claim to have systematize more than 500 million news since start operating.

However, for some lenguages, including Spanish, the GDETL database seems to perform particularly poorly (Osorio 2013).The is the reason that leaded the researcher Javier Osorio to develop techniques that could systematize news texts in Spanish. Particualrly, Osorio (2018) did a extraction of classes based on exact matches, looking for concepts related to drug traffiking in Mexican Army reports. Despite of a high the level of accuracy reached in the exercise, around 70 percent, there are concerns that this matching rules could overfit models for the particular task such that in other contexts they performs poorly.

As a contribution in the same direction, this project uses statistical learning techniques, particurly neuronal networks implemented with SpaCy, to classified the CAMEO classes that where labeled by the Osorio's research team.

Data

This research uses 724 news texts in Spanish coming from the Linguistic Data Consortium (LDC) Gigaword project. Each news text was labeled with the CAMEO classifications (particularly from the Protest sub-

Input 1

```
In [9]: training_data[0][0]
Out[9]: 'HUELGAS EN ECUADOR\n\nQUITO, Jul 12\n\n\nEste martes fue un día marcado por la
s huelgas en Ecuador, donde los hospitales estatales ecuatorianos paralizaron sus activ
idades, los maestros anunciaron un paro de dos días para la próxima semana y en la pobl
ación de La Concordia, en el litoral, sus habitantes dejaron de trabajar durante dos dí
as.\nLos cambios que se generan en el Ecuador, con el inicio de las privatizaciones, la
 reducción de la burocracia y una cerrada ofensiva a la inflación a costa de una supues
ta recesión, han generado descontento ciudadano reflejado por las huelgas hasta el punt
o que comentaristas locales subrayaron la "extraña" inexistencia de protestas en las úl
timas dos semanas.\nLos empleados de los hospitales estatales reclaman los ofrecimiento
s de mejoras económicas y sus condiciones de trabajo, una vez que desde principios de m
es se inició un sistema de cobro a los usuarios. La atención era antes gratuita, aunque
 los pacientes debían adquirir medicamentos cuando faltaban en las casas de salud estat
ales.\nLos profesores acordaron una huelga el martes y miércoles de la próxima semana,
como una acción preventiva para reclamar el pago de los atrasos salariales, lo que fue
rechazado por la ministra del ramo, Rosalía Arteaga, quien sostuvo que se está pagando
dentro de los diez primeros días del mes subsiguiente.\nLos maestros paralizaron sus ac
tividades en noviembre y diciembre pasados, causando estragos en la educación estatal.\
nLa ministra Arteaga, por otro lado, apoya la oposición de los maestros a una ley que s
e tramita en el Congreso para incluir la educación religiosa en los establecimientos es
tatales, si la mayoría de padres lo decide, pero el Gobierno apoya decididamente esta i
niciativa eclesiástica.\nAnte los reclamos de los maestros y sectores políticos que ase
guran que el restablecer la educación religiosa sería un retroceso de un siglo y se con
tradice con la libertad de culto, la iglesia pidió la suspensión del trámite legislativ
o del proyecto hasta que se concrete un debate nacional más profundo.\nMientras tanto,
en La Concordia, cuyos pobladores piden la cantonización -una nominación de la división
 geopolítica que le aseguraría recursos estatales-, las autoridades militarizaron la zo
na para evitar eventuales desmanes.\n'
```

Input 2

```
In [10]: training_data[0][1]
Out[10]:
{'entities': [(0, 7, 'material_conflict')
  (11, 18, 'location'),
  (20, 25, 'location'),
  (75, 82, 'material_conflict'),
  (86, 93, 'location'),
  (101, 125, 'source'),
  (139, 166, 'material_conflict'),
  (172, 180, 'source'),
  (181, 199, 'verbal_conflict'),
  (256, 268, 'location'),
  (276, 283, 'location'),
  (289, 299, 'source'),
  (300, 319, 'material_conflict'),
  (700, 741, 'source'),
  (742, 750, 'material_conflict'),
  (1034, 1048, 'source'),
  (1049, 1069, 'verbal_conflict'),
  (1215, 1235, 'source'),
  (1237, 1252, 'source'),
  (1197, 1210, 'verbal_conflict'),
  (1343, 1355, 'source'),
  (1356, 1383, 'material_conflict'),
  (1461, 1480, 'source'),
  (1497, 1515, 'verbal_cooperation'),
  (1519, 1531, 'target'),
  (1770, 1778, 'source'),
  (1754, 1762, 'verbal_conflict'),
  (1781, 1799, 'source'),
  (1804, 1812, 'verbal_cooperation'),
  (1820, 1844, 'verbal_cooperation'),
  (1928, 1938, 'source'),
  (1939, 1958, 'verbal_cooperation'),
  (1963, 1995, 'verbal_cooperation'),
  (2070, 2082, 'location'),
  (533, 554, 'material_conflict'),
  (573, 580, 'material_conflict'),
  (1146, 1188, 'verbal_conflict'),
  (2210, 2221, 'source'),
  (2222, 2243, 'material_conflict'),
  (2090, 2100, 'target')]}
```

Figure 1: Input: tuples structure

category). The categories are “location”,“material coperation”, “material cooperation”, “source”, “target”, “verbal cooperation” and “verbal conflict”. There around 19 labels for each news with an standar desviation of around 16 labels (there is an important variation in the number of labels for each text).

In **Figure 1** depicts the inputs of the prediction model fitted in the next section. Basically, each text is extracted and stored as the first position of a tuple where the second position of the tuple are the labels (with class and starting/ending position of the word or words that compose the class).

Pipeline

The pipeline or process of prediction if composed of four components:

- Preprocessing
- Splitting
- Training
- Test

Preprocessing basically is divided in two functons: one takes the texts with labels provided by Osorio and create tuples with them (one tuple for each news text) and the second takes the first element of the tuple and process the text to create first sentences, then tokens (one forevery words) and then assigns Part-of-speech (POS), entity recognition (ER) and dependency parsing (DP) (see **Figure 2**).

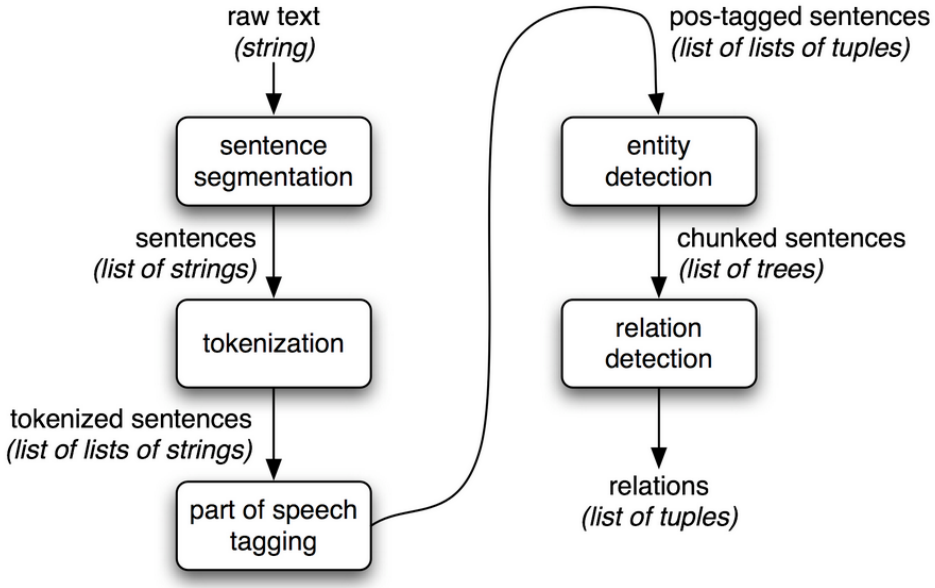

Figure 2: Preprocessing: second function funnel - source: NLTK

Spanish

MODEL	SPACY	TYPE	UAS	NER F	POS	WPS	SIZE
es_core_news_sm 2.0.0	2.x	neural	89.8	88.7	**96.9**	*n/a*	**35MB**

Figure 3: Preprocessing: accuracy, SpaCy tokenization - source: SpaCy

The implemented pipeline uses the SpaCy package for the second function of preprocessing. SpaCy, uses annotated corpuses in Spanish to predict each one of the token values (POS, ER or DP). According to the results reported in their website, the accuracy rates of those predictions are adove 89 % for each category (see *Figure3**).