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.

Introduction

Around 1 million articles are published every day on the web[1]. Those articles possess a universe of information about facts and trends, many times, not available in structured data sets (that are prone to be used for statistical analysis). Also, those articles are in thousands of languages, formats, and styles making the standardization of their content into structured information a very difficult task. However, through natural language processing and statistical learning techniques, I used the features of each word to predict their class (target, source, location and so on), where every class could be composed 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 catalogs authors 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 were unable to catalog asymmetrical warfare 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 a worldwide collector of news in more than 28 languages and machine learning techniques to systematize their information using the CAMEO labels. The GDETL website (now powered by Jigsaw - Google) claim to have systematize more than 250 million news articles dating back to 1970[2].

However, for some languages, including Spanish, the GDETL database seems to perform particularly poorly (Osorio & Reyes 2014). This is the reason that led the researcher Javier Osorio to develop techniques that could systematize news texts in Spanish. Particularly, Osorio & Reyes (2014) did develop an extraction of classes based on exact matches, looking for concepts related to drug trafficking in Mexican Army reports. Despite the high the level of accuracy reached in the exercise, around 70 percent, there are concerns that this matching rules could over-fit models for the particular task such that in other contexts they perform poorly.

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

[1] According to the worldwide blog activity published by WordPress.com

[2] The GDETL Story: https://www.gdeltproject.org/about.html

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

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

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 is 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

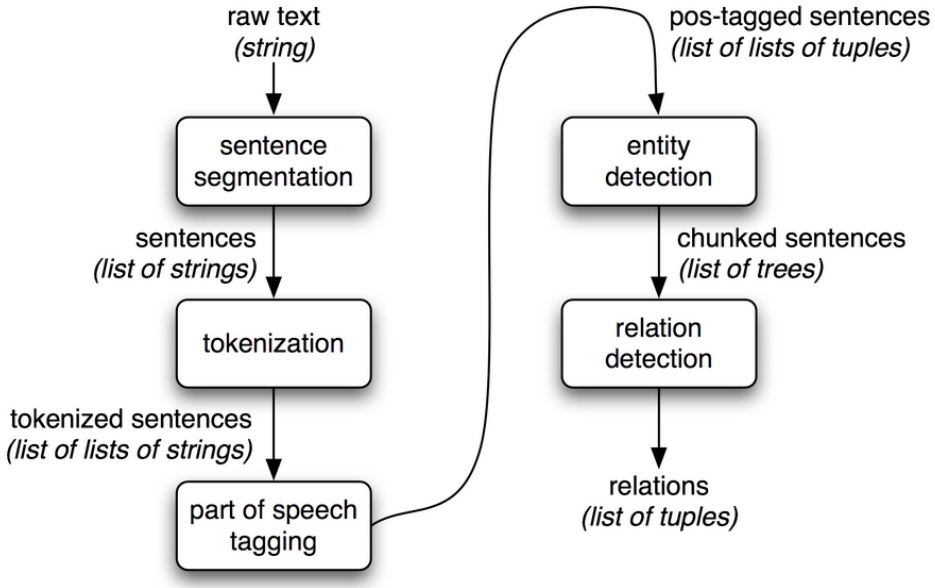

Figure 2: Preprocessing: second function funnel - source: NLTK

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

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

The implemented pipeline uses the SpaCy package for the second function of preprocessing. SpaCy, uses annotated courses in Spanish to predict each one of the token values (POS, ER or DP). According to the results reported on their website, the accuracy rates of those predictions are above 89 % for each category (see *Figure 3**).

After the preprocessing step, an **splitting** of the tuples (news texts) is done in order to obtain a training set and a test set. Before the splitting, there is a random shuffling of the documents to avoid training the model based on the order of the documents on the folder.

Later, the **training** data is used as an input of the Neuronal Network model with a gradient descent optimizer behind (see **Figure 4**).

The statistical model used to determine the weights of the tokens to predict the class is not disclosed by SpaCy, but is likely to be very similar to the Dynamic Recurrent Acyclic Graphical Neural Networks (DRAGNN) model. The framework about how a DRAGNN work is shown in **Figure 5**. Previous testings of those models have reached as high as 93% of accuracy in similar problems (where the human accuracy is 97%).

Lastly, in the **testing** section, the predicted labels are compared with the actual values (to recall: each value

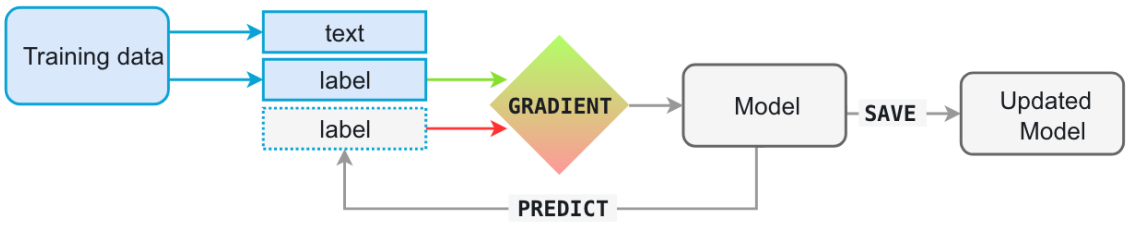

Figure 4: Training: interaction between the splitting and training components - source: SpaCy

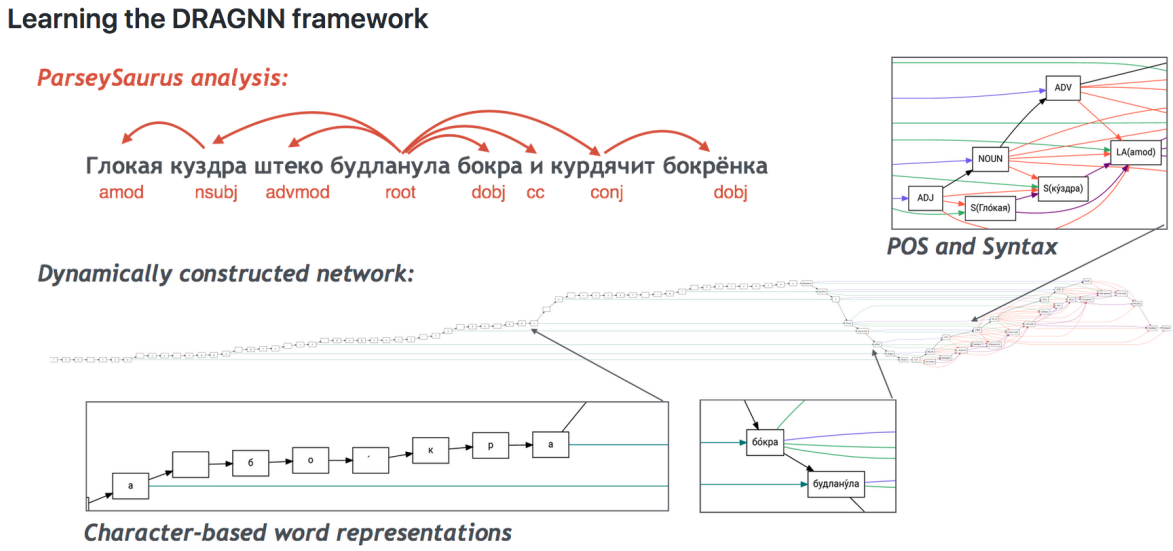

Figure 5: Model - DRAGNN - source: Source: https://github.com/tensorflow/models/tree/master/research/syntaxnet

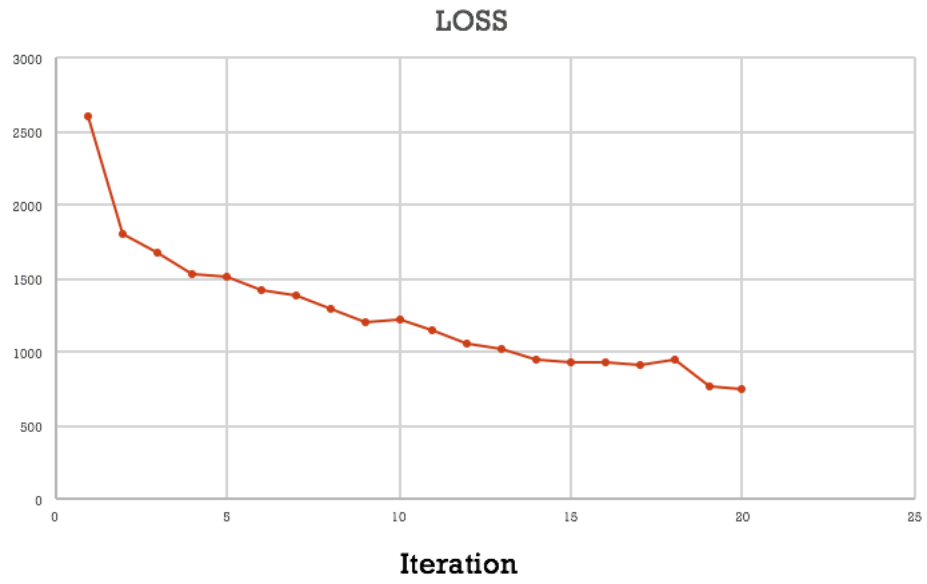

Figure 6: Training: Loss decrease - source: Source: own estimations

is represented by a class and a start/end position in the text of the word or words that compose the class). If there is an exact match (in class and position) then the label is considered as correctly predicted. Accuracy is just the sum of the correctly predicted labels over all the labels in the different texts (remember that these tests were not using in the training of the model such that is an out-of-sample prediction).

Results

So far, the described pipeline was tested with a sample of 50 news articles, where 85% of the sample was used as the training sample and 15% as the testing sample.

As a first result, the loss functions decreased in every interaction over the training set **Figure 6** (a backward and forward pass of the data). This is a healthy sign that model is updating its weights based on learning of the observed samples. The drop-out rate used for this estimation was of 0.35.

The accuracy obtained in this estimation was of 14% for all the classes but it varied by class. For example, in the class "location" the accuracy was as high as 44%.

The accuracy is not as high as predicted because in this exercise no other token values where added in the preprocessing section (just the words as tokens) and the sample of news articles is relatively low: less than 10% of the label data available.

However, despite the low accuracy of this exercise, interesting insights shed some light on the usefulness of the pipeline. For example, in **Figure 7** the class classification can be observed for one sample news article:

First, the model is able to differentiate between sources and targets (however in this case there overlap between predictions). Second, the model can detect relatively well the location of the article. It is particularly promising that the model can distinguish two different classes for a sample token depending on the surrounding context (in this case "Israel" was the source of the action and also the place of the news: Jerusalem, Israel). Third, the material conflict, verbal cooperation and so on classes are particularly hard to predict because they are, normally, long sequences of words around a noun: our model is locating the verbs but not limiting correctly the other words of such classes.

Figure 7: Testing: Class visualization - source: Source: own estimations

Conclusion

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 the 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.

Bibliography

Javier Osorio and Alejandro Reyes. 2014. Eventus ID. Supervised Event Coding From Text Written in Spanish. Version 2.0