Exploratory Analysis for Ontology Learning from Social Events on Social Media Streaming in Spanish

Enrique Valeriano
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú
enrique.valeriano@pucp.pe

Arturo Oncevay-Marcos
Departamento de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú
arturo.oncevay@pucp.edu.pe

Abstract

The problem of event analysis in Spanish social media streaming is that of difficulty on automatically processing the data as well as obtaining the most relevant information, such as mentioned by Derczynski et al. (2015). An event is defined as a real world occurrence that takes place in a specific time and space; Atefeh and Khreich (2013) identifies these occurrences by the entities that took part on it as well as the activities done in it. This project focuses on researching about the viability of modeling these events as ontologies using an automatic approach for entities and relationships extraction in order to obtain relevant information about the event in case. Spanish data from Twitter was used as a study case and tested with the developed application.

1 Introduction

According to Lobzhanidze et al. (2013), globalization and the increased use of social networks has made it possible for news and events related information to be propagated in a much faster manner to every part of the world. It is in this context that event analysis is the most relevant since, as Valkanas and Gunopulos (2013) mention, now there is more data available to study and analyze than ever before.

An event is defined as a real world occurrence that takes place in a specific time and space; Atefeh and Khreich (2013) identifies these occurrences by the entities that took part on it as well as the activities done in it. Events will be the main study object in this paper and, more specifically, event data in Spanish obtained from Twitter will be used to test the different methods and techniques exposed on each Section.

In order to effectively analyze events there are two steps that need to be taken into consideration as mentioned in Kumbla (2016): (1) event data acquisition, and (2) event data processing.

The first step is the one that benefits the most by social media streaming since more data is available, though one of the downsides to this is that the data is usually not ready to be used right away and most of the times a preprocessing step needs to happen. This step is further explained on section Section 3.

The second step will be the main focus on this paper since the biggest problem on event data analysis in Spanish is this one. In particular, automatic approaches for entities and relationships extraction will be presented on Section 4.

The remainder of this paper is organized as follows. In Section 2 some relevant related work is exposed. Later, in Section 3 the event acquisition process is further expanded upon. The ontology structure used for the events representation as well as the algorithms employed in order to obtain entities and relationships between these are further explained on Section 4. Section 5 introduces a simple application developed in order to make use of the algorithms and techniques mentioned on the previous sections. On section 6 we compare the results obtained with manually created ontologies and obtain precision and recall values for each case. Finally, concluding remarks are provided in Section 7.

2 Related Work

In Al-Smadi and Qawasmeh (2016) an unsupervised approach for event extraction from Arabic tweets is discussed. Entities appearing in the data are linked to corresponding entities found on Wikipedia and DBpedia through an ontology based knowledge base. The entities from the data are extracted based on rules related to the Arabic language.

In Derczynski et al. (2015) a comparative evaluation of different NER is done based on three different datasets. Also, some common challenges or errors when handling data from Twitter are presented as well as methods for reducing microblog noise through pre-processing such as language identification, POS-tagging and normalization.

In Ilknur et al. (2011) a framework for learning relations between entities in Twitter is presented. This framework allows for entities as well as entity types or topics to be detected, which results in a graph connecting semantically enriched resources to their respective entities. Then relation discovery strategies are employed to detect pair of entities that have a certain type of relationship in a specific period of time.

In Raimond and Abdallah (2007) an event ontology is described. This model also contains some key characteristics such as place, location, agents and products. On the other hand, event-subevent relationships are used to build the related ontologies. This model was developed for the Center for Digital Music and tested by structuring proceedings and concert descriptions.

Finally, an ontology model for events is proposed in which entities are extracted using the CMU tweet analyzer and relationships are inferred from Wikipedia, DBpedia and Web data. This approach also uses a POS-tagging step in order to obtain the initial set of entities to process.

3 Event data acquisition

3.1 Data retrieval

As it was mentioned before, nowadays there are numerous avenues for event data acquisition. For this paper Twitter was chosen as the social network to use for retrieving data since this data is easily available and a good amount of it is related to events of different categories.

Twitter's REST API was used in order to retrieve data related to these events:

1. Australian Open: 2217 tweets from 21/01/2017 to 30/01/2017
2. March against corruption in Peru: 1493 tweets from 11/02/2017 to 20/02/2017
3. Complaints about new toll in Puente Piedra: 3882 tweets from 08/01/2017 to 18/01/2017

Each dataset had a file per day with all the tweets from the day and contained only the text that represents a tweet per line.

3.2 Preprocessing

With the raw data ready to be used, the preprocessing step followed. The sequence followed is exposed below:

1. Removing punctuation and unicode only characters except written accents.
2. Tokenizing the tweets for easier use in Section 4.

Each tokenized tweet also contains a reference to the original, unprocessed tweet, which will be used on Section 5.

4 Event data processing

4.1 Ontology learning overview

Ontology learning is defined by Cimiano (2006) as the automatic acquisition of a domain model from some dataset. In this paper we focus on applying ontology learning techniques for data represented as text.

Cimiano points towards two main approaches for ontology learning:

1. Machine learning
2. Statistical approach

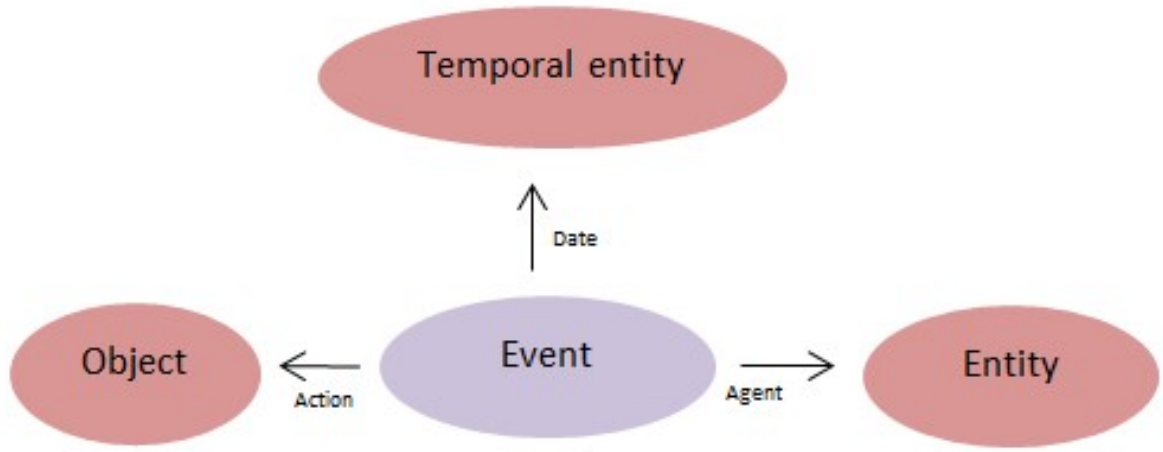
Statistical based algorithms are further discussed on Sections 4.3 and 4.4.

4.2 Ontology structure

Before we start using different techniques in order to populate an ontology or to learn entities and relationships from the data that was retrieved previously, an ontology structure had to be defined.

The ontology structure that we define will point us towards different techniques depending on the information that must be retrieved to populate this particular structure. Therefore, the proposed ontology structure in this paper is defined on Figure 1.

Figure 1: Entity structure



The ontology will be populated by such triples composed of (Entity, Temporal entity, object). Where Entity denotes a subject that interacts in the event, Temporal entity refers to the date when the particular activity takes place and object is the recipient of the activity.

4.3 Entities extraction

This was one of the main points of interest and research on this paper, how to select the most representative entities for the event in order to not overwhelm people analyzing the results but also to not present too little or irrelevant information.

In order to achieve this, two initial tools for entity retrieval were tested:

1. Stanford NER: The Stanford NER used with a trained Spanish model from late 2016 was used in order to retrieve persons, entities and organizations and group them all together as entities.

2. UDPipe: UDPipe allows to parse text in order to obtain the grammatical categories of the words in each sentence, as well as the syntactic dependencies or syntactic tree that envelops the whole sentence. The entities are obtained from the grammatical category **PROPN**.

These two approaches were then implemented and tested with each dataset and a manual comparison was made between the entities that each approach captured.

The results showed that, while the Stanford NER worked really well in the case where the tweets were news related or had a more formal undertone, such as in the case of the Australian Open, it failed to find a lot of basic entities in the other two datasets where the data was more unstructured as one would very likely find when working on social streaming. Also, the Stanford NER has heavily influenced by correct capitalization and punctuation, whereas UDPipe wasn't influenced by these factors as much.

Because of this, UDPipe was chosen as the main initial entity extraction tool moving forward.

After having a set of initial entities, further processing steps were taken to ensure a better result.

4.3.1 Entity clustering

Entity clustering was done on two stages. First, an algorithm for entity clustering was devised based on two metrics:

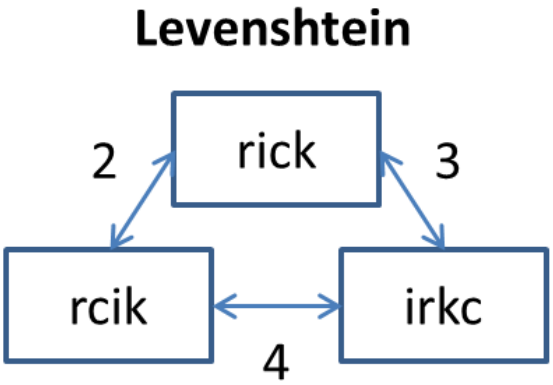
1. Normalized frequency of two entities appearing in a single tweet: The frequency of appearance between two specific entities in tweets.

2. Average Entity to entity distance in a tweet (i.e. in the sentence "Nadal venció a Federer", if both Nadal and Federer are identified as entities, they would have a distance of 3 for this tweet)

A threshold of 0.125 was set as the minimum normalized frequency for a pair of entities and a minimum average Entity to Entity distance of 1.65. These two values were set based on experimentation with the resulting clustered entities from each dataset.

After that, an approach based on Levenshtein distance (minimum amount of additions, replacements or deletions needed to turn a word into another) was employed, where two entities were clustered together if their distance was more than 0.9 times the length of the longest entity from the two. An example of this distance can be seen on Figure 2.

Figure 2: Example of Levenshtein distance



By applying this, resulting clusters such as the ones shown on Figure 3 were obtained.

Figure 3: Resulting clusters for the Australian Open case

```
australian:australianopen,australian,ãustralianopen,open,australia,
nadal:rafael,nadal,rafaelnadal,
andy:andy,murray,
serena:williams,serenawilliams,serena,wlliams,
federer:federe,roger,rogerfederer,federer,
grand:slam,grand,
mischa:mischa,zverev,mischazverev,
```

4.3.2 Formal Context Analysis (FCA)

FCA is one of the approaches for entity extraction detailed on Cimiano (2006). It is the one that garners the most focus on this book as the main set-theoretical approach based on verb-subject components.

This approach is based on obtaining the formal context for a specific domain or dataset and then proceed to use it to create a hierarchy ontology.

An example of how a formal context would look for a tourism domain knowledge can be seen on Table 1.

Table 1: Example of a tourism domain knowledge as a formal context Cimiano (2006)

	bookable	**rentable**	**rideable**
hotel	X		
apartment	X	X	
bike	X	X	X
excursion	X		
trip	X		

In this paper we use the created formal contexts to discriminate between entities based on three metrics:

$$Conditional(n,v) = P(n,v) = \frac{f(n,v)}{f(v)} \tag{1}$$

$$PMI(n,v) = log_2 \frac{P(n|v)}{P(n)} \tag{2}$$

$$Resnik(n,v) = SR(v) * P(n|v) \tag{3}$$

Where:

1. f(n,v) => Frequency of apparition of entity n with verb v
2. f(v) => Frequency of apparition of verb v with any entity

And:

$$SR(v) = \sum_n P(n|v) * log_2 \frac{P(n|v)}{P(n)} \tag{4}$$

A threshold of 0.1 as a minimum value is set for all of the three aforementioned metrics (Conditional, PMI and Resnik weights), meaning that the (entity,verb) pairs that not surpass this threshold for any of the three metrics are pruned.

4.4 Relationships extraction

In this subsection UDPipe is also used in order to extract the syntactic dependencies, in particular, the focus is to obtain 'dobj' and 'iobj' objects, which refer to direct and indirect object respectively, and then obtain the root verb they stem from.

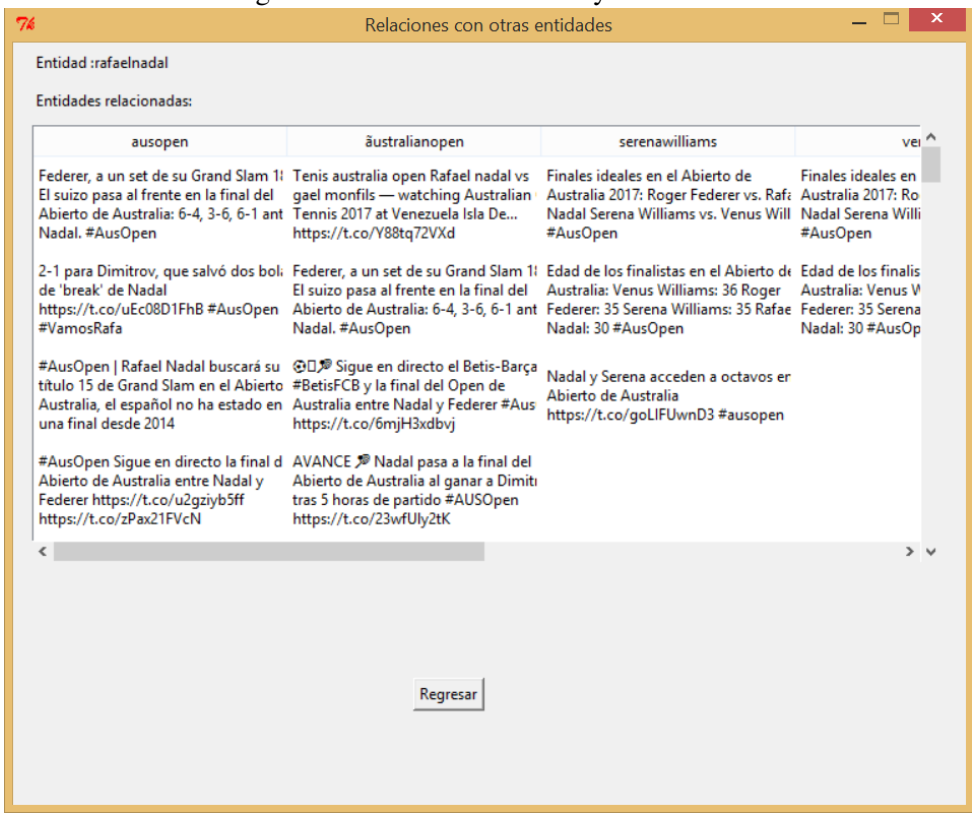
By doing this a verb can be linked to each object and furthermore, the entities related to verb, which were obtained from the Formal Context, can be linked to each object.

Doing this allows us to add activities for each entity, as well as create a relationship between two entities where one of them appears as an object in the action of another.

5 Visualization

A desktop application was developed in order to allow for easier visualization of both the ontology and the resulting activities that each entity participated in, as well as the activities that create a relationship between two particular entities.

Figure 4: Timeline for the entity rafaelnadal



On Figure 4 a timeline was given for the entity rafaelnadal on the Australian Open case, where each day has tweets that represent activities that were extracted from the dataset.

6 Verification

In order to verify the approach applied for ontology extraction, we manually created ontologies for each test case where the most relevant entities and relationships are specified based on investigation related to these cases, these ontologies can be seen on Figures 5, 6 and 7.

These ontologies were then presented to colleagues with more profound knowledge on each of the events for validation and were redone based on their feedback until they were accepted by them.

Figure 5: Ontology created for the Australian Open case

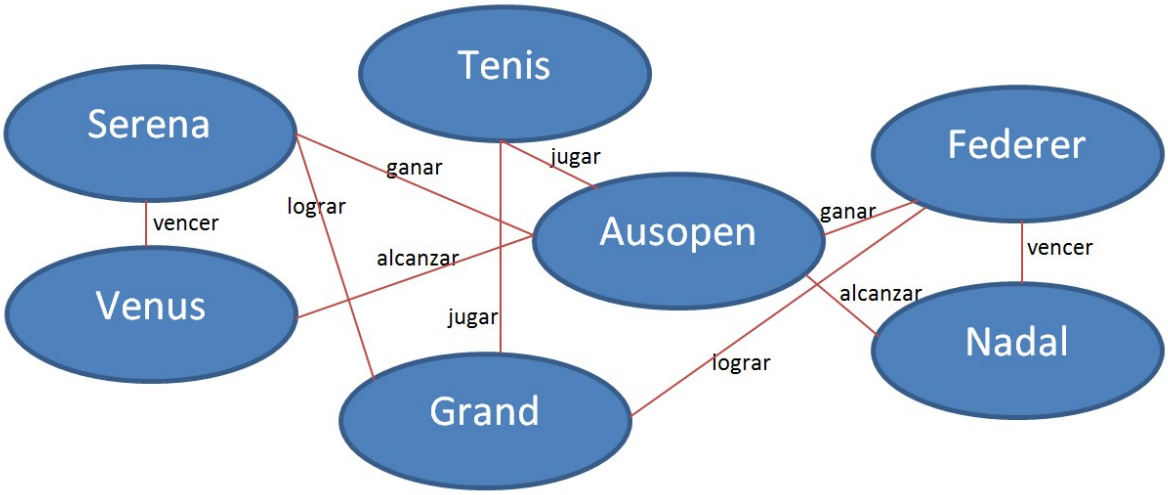


Figure 6: Ontology created for the Puente Piedra's toll case

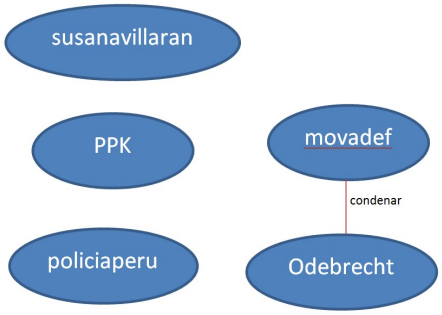
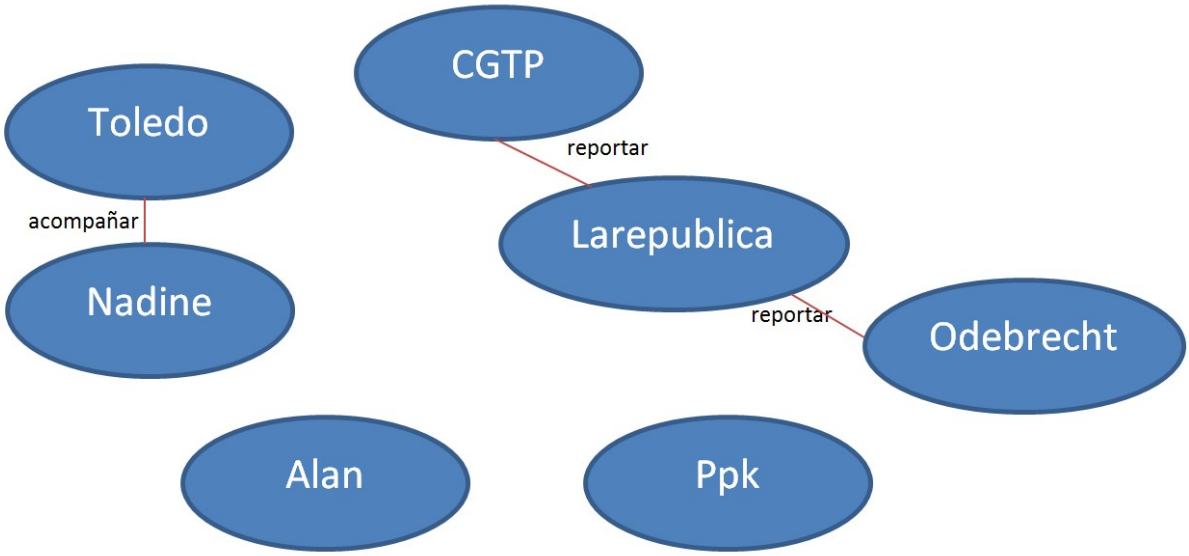


Figure 7: Ontology created for the March against the Corruption case



From these ontologies we obtained precision and recall values for both entities and relationships for each case. These can be seen on Tables 2, 3 and 4:

Table 2: Metrics for the Australian Open case

Analyzed parameter	Metric	Value
Entities	Precision	0.875
Entities	Recall	1.0
Relationships	Precision	0.952
Relationships	Recall	1.0

Table 3: Metrics for the Puente Piedra's toll case

Analyzed parameter	Metric	Value
Entities	Precision	0.556
Entities	Recall	1.0
Relationships	Precision	0.333
Relationships	Recall	1.0

Table 4: Metrics for the March against Corruption case

Analyzed parameter	Metric	Value
Entities	Precision	0.467
Entities	Recall	1.0
Relationships	Precision	0.333
Relationships	Recall	0.667

The main point of interest in these metrics lies on the precision, where the precision on the Australian Open case in quite higher than on the other two cases. From further inspection on the corresponding data we could infer that this was the case because a big part of the tweets for the Australian Open where either formal tweets made by users representing news outlets or by the players themselves. As for the other two cases, most of the tweets where a mix of news and discussion from common people about these events.

7 Conclusions and future work

We conclude that, while the methods exposed on this paper work good enough on cases such as the Australian Open one, there is still work to be done when the general public is more engaged on the event such as the cases of the Puente Piedra toll and the March against the corruption.

This paper's aim was to give a foundation and a initial stage of exploratory analysis on social media streaming in Spanish by using ontologies, after which future work could be based upon in order to expand the knowledge in the ontologies or use this analysis together with an event detection system in order to be able to both detect and analyze events in real time.

References

Al-Smadi, M. and O. Qawasmeh (2016). Knowledge-based approach for event extraction from arabic tweets. *International Journal of Advanced Computer Science and Applications 7*, 483–490.

Atefeh, F. and W. Khreich (2013). A survey of techniques for event detection in twitter. *Computational Intelligence 0*(0).

Cimiano, P. (2006). *Ontology Learning and Population from Text Algorithms, Evaluation and Applications*. 223 Spring Street, New York, NY 10013: Springer Science+Business Media.

Derczynski, L., D. Maynard, G. R., M. v.E., and G. G. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing and Management 51*, 32–49.

Ilknur, C., A. F., and H. G. (2011). Learning semantic relations between entities in twitter. *Information Processing and Management 51*, 32–49.

Kumbla, S. (2016). Fast data: Powering real-time big data.

Lobzhanidze, A., W. Zeng, P. Gentry, and A. Taylor (2013). Mainstream media vs. social media for trending topic prediction - an experimental study. *Consumer Communications and Networking Conference (CNNC)*, 729–732.

Raimond, Y. and S. Abdallah (2007). The event ontology.

Valkanas, G. and D. Gunopulos (2013). Event detection from social media data.