The Workshops of the Tenth International AAAI Conference on Web and Social Media
Social Media in the Newroom: Technical Report WS-16-19

"Just the Facts" with PALOMAR: Detecting Protest Events in Media Outlets and Twitter

Konstantina Papanikolaou[1], Haris Papageorgiou[1], Nikos Papasarantopoulos[1], Theoni Stathopoulou[2], George Papastefanatos[3]

[1] Institute for Language and Speech Processing/Athena RC, Athens, Greece

[2] National Center of Social Research, Athens, Greece

[3] Athena RC, Athens, Greece
{konspap, xaris, npapasa}@ilsp.gr, theosta@ekke.gr, gpapas@imis.athena-innovation.gr

Abstract

The volume and velocity of available online sources have changed journalistic research in terms of cost and effort required for discovering stories. However, the heterogeneity and veracity of data sources pose further obstacles in knowledge extraction making it a hard task to handle. The purpose of this study is threefold. Firstly, we present a platform for automated data processing in the context of Computational Journalism. We then propose a general methodology for event extraction from different data sources. Finally, we conducted a pilot implementation of our methodology for protest events extraction from news and Twitter data. Evaluation showed promising results, indicating the feasibility of our approach.

1 Introduction

Event extraction has been a very critical yet challenging task for Information Extraction systems. It responds to needs that trigger interdisciplinary research with the goal of automating the collection and curation of knowledge that is related to certain domains. Nowadays, it is being extensively used by social and political sciences (Schrodt and Van Brackle 2013), as well as in the biomedical field (Ananiadou et al. 2010); at the same time events are the principal focus of journalists in their daily work. Many approaches have been reported, spanning machine learning, statistical, and knowledge-driven approaches which are mainly pattern-based. Despite the proliferation of research efforts, it remains a non-trivial task considering its possible applications and analysis possibilities, given the emergence of data sources requiring novel ways of curating.

Formerly, the main source for events discovery were the various news agencies. With the advent of Social Media (SM), the landscape has drastically changed reformulating the way people share information and communicate worldwide. SM data constitutes an intriguing source for journalists, as it can help them discover interesting topics for articles, check the validity and the veridicality through different sources, measure news stories ideological leanings, detect controversy, analyze event spread and impact or forecast civil unrest. However, the processing of this data is not easy to handle. Since natural language is involved, special tools are needed for event discovery. Furthermore, social media content is noisy, in terms of a great percentage of non-specific or trivial information posts.

The contribution of this paper is threefold: (a) to present the design and the big data architecture of PALOMAR, an automated Computational Journalism platform for scalable processing of data streams of news sources and social media, (b) to describe an innovative, semi-supervised methodology for the detection of certain event types together with their structural components. Here, we propose a data-driven yet linguistically based framework grounded on social and political sciences incorporating a human-in-the-loop. (c) To report on a pilot implementation of the methodology concerning a longitudinal study of protest events enabling the connection and analysis of protests in news and social media. The outcome of this analysis is subsequently visualized and linked in an intuitive way. Moreover, an internal and external evaluation of the pilot is also documented.

Thus, journalists can use PALOMAR to exploit heterogeneous sources and discover interesting topics for articles, examining possible bursts emerging in SM that are not recorded by news data or vice versa. This is facilitated by the multiple visualizations that the journalists themselves can produce in an interactive tool.

The paper is structured as follows: the system architecture is described in section 2, and the event extraction methodology in section 3. In section 4, details of the pilot implementation are provided. Finally, Related Work (section 5) and Conclusions (section 6) are discussed.

2 System Architecture

The objectives as described in the introductory section are traditionally approached via costly and non-reproducible coding of documents. However, this type of coding does not scale when confronted with the vast amount of textual material that is available in news archives or generated daily in the social media. The requirements for journalists who would code materials of such magnitude is prohibitive in terms of cost and size. We therefore develop PALOMAR in order to support these objectives by employing innovative technologies that will help journalists extract historical evidence from large textual collections and equip them with robust tools for providing explanations to questions of interest.

The **PALOMAR** Data Analysis and Modeling Platform is an innovative, automated Computational Journalism platform encompassing various scientific instruments, a wealth of datasets enriched with metadata automatically produced by reliable analytics workflows and key insights revealed by modeling and visualization tools. Figure 1 provides a conceptual overview of the architecture of the platform together with its key systems and components.

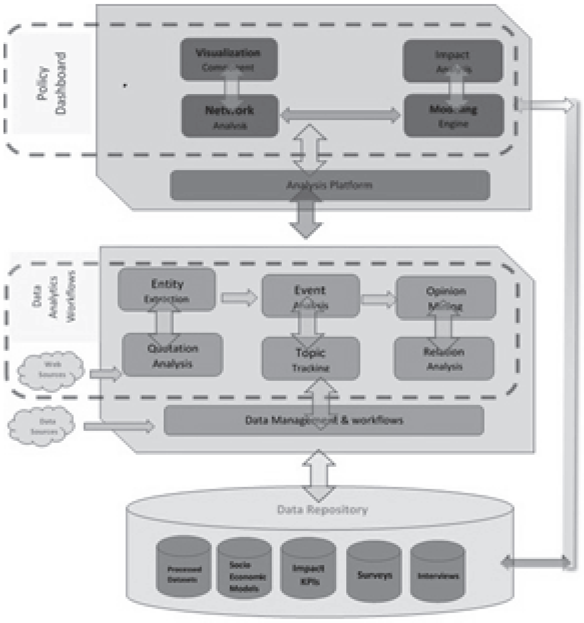

Figure 1: The PALOMAR architecture

PALOMAR is a **service-oriented** infrastructure offering an operational platform to enable the execution of text mining services in a seamless, interoperable way, through the use of existing cloud based infrastructure. On the **Data Repository** side, the PALOMAR infrastructure is powered by the ELK stack[1] (Elastic, Logstash, Kibana), a platform for indexing, searching and analyzing data[2]. In addition, extracted metadata and generated analysis results are also archived and curated. The Data Repository is initialized with private and public datasets. The **Data Analytics** Layer integrates Language Technology tools and workflows for processing the datasets curated in the Data Repository and generating a wealth of valuable metadata from these sources, including the extraction of specific **events, actors, topics, quotations** and **opinions**, according to coding frameworks like the PROMAP (Stathopoulou forthcoming) described in section 4.1. PALOMAR's third layer, the **Modeling Dashboard** is the place where content-based metadata and network analysis are synthesized in order to provide journalists with diverse ways of observing evolving patterns and trends. Finally, evaluation of the analysis is offered to journalists in the form of intuitive, exportable **visualizations** providing insights on the correlation between diverse variables of interest, the dynamics of events on social media streams, etc. The **visualization component**[3] includes timelines of configurable temporal windows, information maps, the ability to zoom in-out to specific textual collections or to filter results using sub-queries etc., which help journalists understand the dynamics of events at different dimensions.

3 Event Extraction

As mentioned above, the Data Analytics Layer activates scalable text analytics workflows in order to extract knowledge graphs including domain-specific events among others. The event extraction methodology we propose consists of the following phases:

At first, the coding framework covering a wide spectrum of event types is designed by news experts. The codebook we used for the pilot formalizes the topology of certain event types of interest and it is grounded on related work in the political and social sciences (cf. section 5). In parallel, the Data Repository caters for the collection and the curation of the data sources including data cleansing, preprocessing and preparation for further analysis. At the next phase, journalists interact with the pre-processed data through the Data Repository stack in order to explore the instantiation of the event types and their structural components, filter them according to the scope of their research,

[1] http://bit.ly/20VWsvi
[2] http://palomar.ilsp.gr/palomar
[3] http://bit.ly/1Q1WUVU

generate event-oriented data clusters, while at the same time enriching and improving the codebook in an iterative fashion. Based on the improved codebook and the event-oriented data clusters, a scalable, semi-supervised text workflow leverages linguistic structure in order to automatically extract event instances and create the event database. The PALOMAR modeling dashboard with its visualization component facilitates journalists in their exploration of the database, providing events in context and making them understandable and interpretable for their research. Finally, the automatically generated event database is further statistically analyzed and evaluated against a small testing set.

A pilot implementation of the above described methodology was conducted concerning protest events and aiming to evaluate the results so as to further extend it to other event types. This implementation took place in two stages. Initially, in collaboration with social scientists, protest events were extracted from news data using a linguistically based approach, which demonstrated significant evaluation results. This gave us the impetus to experiment with a completely different data source, namely Twitter, for the same purpose, extracting protest events. Our event extraction workflow consists of the following processes: (i) at a first stage, the codebook was developed, containing the protest events taxonomy and their complementary elements. (ii) Data were collected from different sources and ETL transformations turned them into a human readable corpus enabling researchers to efficiently and effectively investigate it. (iii) Event Extraction includes preprocessing and a cascade of transducers implemented as Gate JAPE[4] patterns which identify events and their constituents. (iv) Finally, results were visualized in the Palomar dashboard. This workflow is elaborated in the next section.

4 Pilot Implementation

The above proposed methodology, was implemented in the context of a project whose main goal was the mapping of protest events in Greece for the period 1996-2014. First, a Codebook (section 4.1) for this category of events was designed. Then, data were collected from newspapers and Twitter (section 4.2), filtered (section 4.3), and analyzed (section 4.4) accordingly. Finally, the generated Event database was internally and externally evaluated (section 4.6).

4.1 Coding Schema

A coding framework for protest events entitled PROMAP was designed by the National Center of Social Research. The PROMAP Codebook was primarily based on the codebook used for Europub project[5] on the transformation of political mobilization and communication in European public sphere (Koopmans 2002). According to the codebook, the analysis unit for mapping protest events, is the *Claim* which consists of six distinct elements: *Form*, A*ctor*, A*ddressee*, *Issue*, *Location* and *Time*. From an event extraction point of view, the *Claim* is an event tuple comprising six different information types. The essential element of each *Claim* is the *Form*, which represents a form of action, i.e. an event type. 19 *Forms* were examined, in particular: *Policing, Petitioning/Signature collection, Strike, Half-day strike, Work-to-rule, Abstention from duties, Withholding of labor, March/Demonstration/Public assembly, Motorized march, Sit-down, Suspension of service, Boycott, Disturbance of meetings, Blockade, Occupation/Sit-in, Hunger strike, Prison riot, Symbolic violence and Arson, bomb attacks, destruction of property*.

As far as the rest of the elements are concerned, *Actor* is classified into categories according to its role or status at the time the *Claim* is formulated, e.g. Government, Enterprises, Tertiary Trade Unions. *Addressee*, is categorized the same way as *Actor*. The *Issue*, which is the subject matter of protest, is recorded as found in the text, *Location*, refers to the city, or the exact place where an event took place, with the limitation of that being in Greece. Lastly, *Time*, is the event's date.

All of the above information types comprised the event template, namely a tuple of the following type:

<Form, Actor, Addressee, Issue, Time, Location>

However, not all of the elements were needed to record an event. The crucial element, as mentioned, was the *Form* and at least one of the *{Actor*, *Addressee*, *Issue}*. *Time* was by default the article's issue date and *Location* was optional information, recorded only when found in the text, otherwise *Greece* was used as the default value.

4.2 Data Collection

4.2.1 News Data

The news dataset consists of articles published in two Greek newspapers, *Kathimerini*[6] and *Avgi*[7]; specifically, all articles of the Wednesday and Friday issues, from 1996 to 2014. All the articles are in Greek and, for each one, along with the text, section labels, headlines and the names of the authors were collected. The dataset comprises 540.989 articles in total, 314.527 from *Kathimerini* and 226.462 from *Avgi*.

4.2.2 Twitter Data

The Twitter dataset consists of tweets from 2013 and 2014. All tweets posted in the Greek language were downloaded

[4] https://gate.ac.uk/sale/tao/splitch8.html#chap:jape

[5] http://europub.wzb.eu/

[6] http://www.kathimerini.gr/

[7] http://www.avgi.gr/

making use of the twitter API and they were stored and indexed together with their respective metadata (user information, retweet information etc.). The total number of tweets processed was 166.100.543.

4.3. Explorative Analysis

Data exploration is an integral part of the methodology, since our approach is data-driven and incorporates human-in-the-loop. Consequently, expert users explored the bulk of data gathered in order to determine the lexicalizations of the various event types and their components. This part was crucial for filtering the data and clustering them into targeted collections. This was an interactive, iterative procedure, since the exploration was driven by the Codebook (see Section 4.1), and the latter was modified according to the findings from the data exploration.

4.3.1 News Data

The main goal of the explorative analysis step was to better understand and obtain a broader view of the whole dataset. To achieve this, a full text search application was developed.

Sections of the newspapers articles that were considered irrelevant to the research scope were filtered out: the scope of the research was Greece, so international news were excluded. Moreover, sports or culture sections were considered out of scope as well. This procedure limited the number of articles from both newspapers to 362.884.

After that, analysts were able to make full-text queries, select articles, examine them and save their search. Later on, in the data analysis phase, the queries and the documents that were marked as relevant were retrieved to construct document clusters referring to one event type.

4.3.2 Twitter Data

The exploration step for the twitter dataset was similar to that of the news data. The PALOMAR platform was used in order to index the tweets and make the dataset available to the analysts. Again, the ability to save queries which were later used to make clusters of tweets on event types is the core functionality of the interface. In addition, it provides enhanced functionalities for storage, processing and visualization (see Section 2).

4.4. Event Extraction

The overall protest events extraction framework is data-driven. For the news data, the approach is mainly based on linguistic rules, while for Twitter data the special feature of hashtags is exploited. Both methods are described in the following sections.

4.4.1 News Data

In general, the rationale for extracting events from news data is bootstrapping, in terms of building over previously produced annotations. In brief, at the pre-processing stage, a basic natural language processing workflow was applied producing a set of annotations. After that, the Event Analysis workflow executes two tasks in sequence. First, the structural components of the event template are detected, and then linguistic rules that exploit all the previous annotations attempt to link the right components to create the event tuples with which the final Event Database is constructed. The system comprises a set of FST rules (Finite State Transducers) which exploit multiple linguistic annotation streams already produced by previous tools. The FST rules are ordered, forming a cascade, so that the output of a transducer is given as input to the next. The workflow of the event detection system is illustrated with the following indicative example:

“*Employees at the movement of goods at the premises of Cosco, in Piraeus, are on a strike since this morning, protesting about working conditions.*”

As already mentioned, in the first stage, raw text is the input to the NLP pre-processing tools (Prokopidis, Georgantopoulos and Papageorgiou 2011) which produce annotations for POS tags (e.g. noun), Lemmas, Chunks (e.g. prepositional phrases), Dependency relations (e.g. Subject) and Named Entities classified into four categories: *Person*, *Organization*, *Location* and *Facility*. For example, in the above excerpt, NERC would recognize *Cosco* and label it as *Organization*. The next module, Nominals Phase, exploits the POS tags, Lemmas and Chunks to annotate entities that are not named, yet their nominal expression is present in the text. In the example, *employees at the movement of goods* would be annotated as *Nominal*. These entities (Persons, Organizations and Nominals), are assigned the label *Candidate* as they all are entities that can potentially be *Actor* or *Addressee*. *Time* expressions and *Issue* are located next and annotated as such. It is noteworthy that *Issue*, namely the subject matter of the protest, is heavily dependent on semantics. Consequently, patterns containing trigger words along with their syntactic complements were used for its detection. In such a pattern, a trigger word is “protest” and its syntactic complement a prepositional phrase starting with “about”. Thus, in the example, *working conditions* would be recognized as an *Issue*.

Next, the element *Form* of action is detected. It is important to note that we address both verbal and nominal event extraction, so all possible lexicalizations are recorded. Furthermore, every such annotation is attributed with certain features, in terms of the semantic information it conveys. In our example *are on a strike* is annotated as *Form* and given the attribute *Strike* so that it can be distinguished and later used in linguistic patterns. Finally, a set of linguistic rules of shallow syntactic relations patterns are implemented. These rules utilize the semantic information of events and their structural components combined with syntactic information so that *Candidate* entities can be assigned an *Actor* or an *Addressee* label and link all the components to create a tuple at the sentence level. In the above example, the event tuple would be as follows:

<***Actor**: Employees at the movement of goods, **Form**: are on a strike, **Addressee**: Cosco, **Issue**: working conditions, **Time**: 07/18/2014, **Location**: Piraeus*>

It is important to note that a small, random sample of every dataset created in the phase of Data Exploration, was used as a development corpus, in terms of testing and improving the linguistic patterns and rules against it. Moreover, this corpus was annotated by humans – three social scientists – and used to evaluate and enhance the system’s performance before applying it to the whole news corpus.

Post-processing: At the post-processing phase, the processing window is expanded to the sentence where a *Form* was detected ±1. If a *Form* annotation has not been successfully linked to an *Actor* and/or *Addressee*, this module attempts to find a *Candidate* entity and assign one of these labels to it, using rules and restrictions related to the features of the *Candidate* entity. For example, an entity that is categorized as *Government*, corresponds to the government and its representatives and cannot be the *Actor* of a *Strike* event type.

4.4.2 Twitter Data

Twitter as a dataset exhibits some completely different features compared to news. One of the most important is the limit of 140 characters, which enforces users to condense the message they want to communicate. In order to do this, they use special elements such as hashtags, user mentions or URLs. Additionally, words that do not influence the general meaning (e.g. function words) are often omitted, making it hard for the standard NLP tools to process. Given that Twitter-specific NLP tools have not yet been fully developed for the Greek language, other ways of extracting information need to be used. Specifically, we followed a quite simple - yet effective- workflow of event extraction from Twitter that exploits a distinctive and commonly used feature, namely the hashtags.

The event type of S*trike* was the one we chose to experiment with in the Twitter dataset, due to its frequency in the news data. The method used for extracting such events from Twitter was semi-supervised. Our concern was to make news and Twitter results comparable, therefore we decided to extract similar event tuples, containing the same information types. Nonetheless, the *Issue* type was not addressed at this stage, because of the semantic and syntactic processing it requires. In any case, some first observations and hypotheses already made will be further explored in future work.

Similarly, to the news data event extraction, the goal here is to first detect the different information types and then link the constituents to create an event tuple. However, the implementation is different as follows: in a first stage, seed terms were used to filter out all the tweets referring to the specific event type, i.e. *Strike*. The next step was to detect the *Candidate* entities and *Location*, based on predefined word lists as well. Finally, Candidates were assigned the label *Actor* or *Addressee* with the same restrictions used for the news data (see Section 4.4.1). All the extracted components populated the event tuples recorded to the Twitter Event Database. For the *Time* element, the date the tweet was posted was used. This decision was made for two reasons: (a) due to the size limit of tweets, time expressions are not very often, and (b) tweets are messages commenting current events.

The seed terms for the detection of event instances consisted of the lexicalizations of strikes as derived from the news data analysis. In particular, the query terms from the Data Exploration, which were also an integral part of the Events Grammar, were converted into hashtags in three different ways: a) in **Greek**, as found in newspapers, e.g. #απεργία [aperjía] (=strike) b) **translated**, namely using the English word #strike, c) **transliterated**, that is conversing the text from the Greek script to Latin, i.e. #apergia. Moreover, wildcards were used in order to discover compound hashtags containing terms in question, such as #general_strike. In this way, the pool of tweets was filtered and only the ones referring to a strike were gathered into a collection consisting of 4002 tweets. Nevertheless, this was not sufficient for recording an event since – according to the Codebook – apart from the *Form* at least one of the other constituents (*Actor*, *Addressee*) is required for a *Claim* to be recorded.

In the next phase, the system detected the entities mentioned in the tweets. To do this, the NERC’s Organization and Person Gazetteers were enriched with nominal terms (see Section 4.4.1) that were used as search terms. Correspondingly to the *Form* annotation type, these lists were also translated and transliterated, so as to cover all the possible variations. Eventually, by using simple rules, the events that were linked to at least one entity, viz. the *Actor* or the *Addressee*, were recorded to the event database along with their structural constituents.

The Twitter workflow is illustrated by the next example tweet.

#24-hour_strike #adedu tomorrow Tuesday 14 May 2013 #athens http://t.co/zov0cl1ka0.

The extracted event tuple would be:

<***Form**: 24-hour_strike, **Actor**: adedu, **Time**: Tuesday 14 May 2013, **Location**: athens*>

4.5 Data Visualization

The last step in data processing concerned the actual data visualization. For that, we have employed the Socioscope platform[8], an interactive web-based visual analytics plat-

[8] www.socioscope.gr

form for social and political data that enables the analyst to explore and analyze social facts through a user-friendly visual interface. The Socioscope platform offers a variety of interactive visualizations for different types of data, such as charts and histograms, pies and stacked diagrams for numerical data, timelines for time series and choropleth and point maps for geographical data. It is based on a multidimensional modelling approach and offers several visual operations for data exploration and analysis, such as filtering through faceted browsing, hierarchical representation of coded lists in charts, free keyword search of literal values, and capabilities for combination of different datasets along common dimensions. Moreover, it enables the reusability of knowledge by making all data available for download in various formats as well as in the form of Linked Open Data, which is a standard means for data sharing on the web, enabling citation and unique referencing across sites.

4.6 Evaluation

An intrinsic and extrinsic evaluation of the events extraction system was conducted. In the intrinsic evaluation, the Strike event type and a specific month, viz. 2/2014 were selected. To measure the precision, we used the number of correct extracted instances. To calculate this number, we counted as false positives tuples where a wrong event was recognized (e.g. a hunger strike instead of a strike), or the *Actor*/*Addressee* was conceptually erroneous, or – for cases where the tuple comprised only the event and an Issue – the Issue was wrong or incomplete. To measure the recall, we manually annotated all of the articles from that period.

For Twitter, precision was estimated from the extracted events, while for measuring the recall only the tweets containing at least one of the seed terms were analyzed. The evaluation results for both approaches are presented in the table below.

	News	Twitter
Precision	90%	97.5%
Recall	93%	92%

Table 1: News and Twitter results evaluation

An extrinsic evaluation was also conducted for the news data. Precision was estimated by counting the correct records of the whole Event base for the event category “Strike”. For recall, on the other hand, GDELT was used as baseline. More specifically, we extracted an event database from GDELT using the following criteria: the event code 143 (*Strike* and *Boycott*), for the period 1996-2014, with the condition that at least one of the Event Location, Actor1 Country or Actor2 Country is Greece. This evaluation method resulted to a precision of 80%. GDELT follows a different coding framework from ours and matching distinct events was out of scope for this study.

A simple graph (Fig. 3) depicting the events reported in GDELT event database and the ones recorded in the PROMAP project, highlights a significant difference in coverage, favoring PROMAP. This is the case through the whole timeline, with an exception of the period 2010-2012.

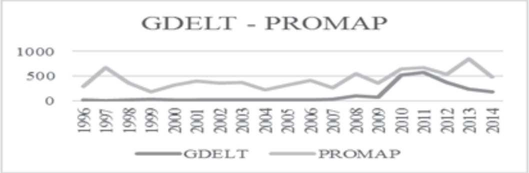

Figure 3: GDELT vs PROMAP results

4.7 Results

Regarding the events extracted from the newspapers, qualitative analysis allowed some interesting readings. Among them, it is worth mentioning a correlation between election years and number of events recorded. In particular, a notable decrease of the total number of claims is observed in election years. It would be interesting to verify if this is the case for Twitter mentions as well, yet Twitter is a newly emerged medium and there is no data before 2007.

One of our main goals was to correlate the events extracted from different sources, i.e. news and Twitter. For this purpose, we created a timeline for both results at a monthly level. Figure 4 represents the results for 2013 and Figure 5 for 2014.

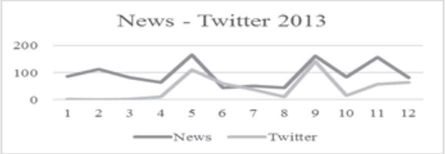

Figure 4: News vs Twitter timeline for strikes (2013)

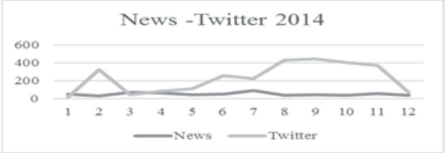

Figure 5: News vs Twitter timeline for strikes (2014)

Starting with the 2013 timelines, it was noted that event mentions follow a similar pattern. More interesting observations come from the 2014 comparison: first, there is a spike of strikes reported in Twitter in February, compared to newspapers; this holds almost throughout the year.

In an attempt to explain this inconsistency, and after taking a closer look to the results, we noted that more than 80% of the strike mentions in 2014 refer to one long-lasting strike, undertaken by the employees of Coca Cola, when the latter decided to shut down the factory in Thessaloniki. This strike lasted for more than a year and so do the mentions of it in Twitter. What would be more interesting is to look deeper into the users that post all these tweets and the ways they are connected to each other. A social network analysis could reveal the extent to which there is coordination of the users towards a certain end. In sum, comparing event mentions in news and Twitter data, is a useful tool for journalists, since they can discover interesting aspects of events to write for.

5 Related Work

Event extraction for political and social science has been a long-standing topic, dating back to hand coding data. Work on automatic annotation started within the KEDS/TABARI project (Shrodt, Shannon and Weddle 1994). Evaluations have shown that hand coded and automatic events coding show comparable performance (King and Lowe 2003). Several coding schemes have been developed since, including the IDEA (Bond et al. 2003) and ICEWS (O' Brien 2012). One of the most renown and influential frameworks for event extraction is CAMEO (Gerner et al. 2003), which is still used by the ongoing GDELT project[9] (Leetaru and Shrodt 2013). All of these efforts have focused on news data, that have traditionally been the main events' source. Our codebook, PROMAP, follows the same principles with a linguistically-driven implementation. Protest Events Analysis has been a central issue in the context of Political and Social sciences (Wueest, Rothenhäusler and Hutter 2013). Moreover, the use of event extraction for predictive analysis is a very challenging and far from trivial task, whose potential justifies its impact within the literature (Boschee, Natarajan and Weischedel 2013).

Social media, and especially Twitter, have also been extensively used for event extraction on many different topics and for various purposes such as, climate change (Olteanu et al. 2015). The authors in (Wang, Fink and Agichtein 2015) discover social events, like concerts or conferences and their structural components, i.e. location, time and title of the event, by linking the information from the tweet and the embedded URL. In Becker, Naaman and Gravano (2011), tweets are distinguished to event and non-event messages, with the first being clustered into topic categories. Ritter et al. (2012) extract event tuples from Twitter stream and classify them into topics, using an unsupervised approach. In Popescu, Pennacchiotti and Paranjpe (2011), events concerning specific known entities are discovered and structured, using a supervised method to decide for the relevance of tweets. Li, Sun and Datta (2012) and Qin et al. (2013) both rely on text segments in order to detect and classify events in Twitter, with the first making use of Wikipedia for the filtering of real events and the second implementing feature clustering. Temporal and spatial information has also been used for identifying and categorizing events in Twitter, as in Parikh and Karlapalem (2013) and Walther and Kaisser (2013). In Weng et al. (2011), signals are built for each word in a tweet and then correlated to form a distinct event. Finally, in the context of protest events extraction, Zachary et al. (2015) examine mass protest that can lead to political changes at a country level, using popular hashtags and measuring the extent to which Twitter users' coordination within social networks, can cause collective action. The different scope and evaluation methodology of the above systems make it difficult to compare their performance. However, most of them report a precision ranging between 70 – 85%.

All of the above mentioned projects implement methodologies varying from completely unsupervised, purely data-driven approaches, to knowledge-driven methods based on domain experts. Hybrid frameworks have also been used (Hogenboom et al. 2011). Our approach is a hybrid method with human-in-the-loop.

Moreover, other research efforts have focused on developing platforms for information extraction, gathering and visualizing, addressing mainly to journalists. For instance, in Marcus et al. (2011) TwitInfo is presented, viz. a system for identifying events in Twitter and visualizing them along with a sentiment aggregation of the corresponding tweets. SocialSensor (Diplaris et al. 2012) also aims to provide users a tool for discovering trends, events and other interesting information from online multimedia content. Diakopoulos, Choudhury and Naaman (2012) describe a framework for processing tweets according to user queries and detecting eyewitnesses, while TweetGathering (Zubiaga, Ji and Knight 2013) aspires to help journalists discover news stories in Twitter.

6 Conclusions

In this paper, an innovative platform for automated processing of data from different online sources is presented. The PALOMAR system allows journalists to analyze and visualize various information types. Event extraction is addressed first, for which a semi-supervised linguistically driven methodology is proposed. This methodology is implemented in a pilot task of extracting protest events along with their structural components.

As far as the pilot implementation on protest events is concerned, the evaluation showed significant results both for precision and recall. Also, results from the two different sources turned out to be comparable, in that protest events reports can be found both in newspapers – as expected – and in Twitter. This could be quite helpful for journalists in verifying an event through multiple sources. However, what is more interesting is the impact that certain events may have in social media. As mentioned Twitter revealed a disproportionally large number of mentions to a certain strike, which could be further examined in

[9] http://gdeltproject.org/

terms of users’ network and whether there is coordination and call for action. This analysis would be challenging, for example, in cases of terrorism, attacks or social revolutions. Future work includes expanding the functionality of the above presented platform. It is our intention to extend our analysis to other analytics tasks, such as topic and quotations analysis and thus enrich PALOMAR and equip journalists with an interactive, multifunctional toolkit for real-time analysis of multi-source data streams.

7 Acknowledgements

This work is partially funded by the projects EOX GR07/3712 and “Research Programs for Excellence 2014-2016 / CitySense-ATHENA R.I.C.”

8 References

Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D. B. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology* 28: 7, pp. 381–390.

Becker, H., Naaman, M., Gravano, L. 2011. Beyond Trending Topics: Real-World Event Identification on Twitter. In *Proceedings of the International Conference on Weblogs and Social Media*.

Bond, D., Bond, J., Oh, C., Jenkins, J., Taylor, C. 2003. Integrated Data for Events Analysis (IDEA): An Event Typology for automated events data development. *Journal of Peace Research* 40(6):733–745.

Boschee, E., Natarajan, P., Weischedel, R. 2013. Automatic Extraction of Events from Open Source Text for Predictive Forecasting. In Subrahmanian V. (ed) *Handbook on computational approaches to Counterterrorism*. New York: Springer.

Diakopoulos, N., Choudhury, M. De, Naaman, M. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2451–2460. New York, USA: ACM.

Diplaris, S., Papadopoulos, S., Kompatsiaris, I., Goker, A.S., MacFarlane, A., Spangenberg, J., Hacid, H., Maknavicius, L. & Klusch, M. 2012. SocialSensor: Sensing User Generated Input for Improved Media Discovery and Experience. In: *Proceedings of the 21st international conference companion on World Wide Web.*

Gerner, D., Schrodt, P., Yilmaz, O., Abu-Jabr, R. 2002. *Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions*.

Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F. 2011. An Overview of Event Extraction from Text. In M. van Erp, W. R. van Hage, L. Hollink, A. Jameson, and R. Troncy, (eds), *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web at Tenth International Semantic Web Conference*. 779, 48–57.

King, G., Lowe, W. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: a rare events evaluation design. *International Organization* 57(3):617-642.

Koopmans, R. 2002. *Codebook for the analysis of political mobilisation and communication in European public spheres*. Available from: <http://europub.wzb.eu/codebooks.en.htm>.

Leetaru, K., Shrodt, P. 2013. *GDELT: Global Data on Events, Language and Tone*, 1979–2012.

Li, C., Sun, A., Datta, A. 2012. Twevent: Segment-based event detection from tweets. *Proceedings of the 21st ACM international conference on Information and knowledge management*.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S. and Miller, R. C. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*.

O’ Brien, S. 2012. A multi-method approach for near real time conflict and crisis early warning. In Subrahmanian V. (ed) *Handbook on computational approaches to Counterterrorism*.

Olteanu, A., Castillo, C., Diakopoulos, N., Aberer, K. 2015. Comparing Events Coverage in Online News and Social Media: The Case of Climate Change. In *Proceedings of the International Conference on Weblogs and Social Media*.

Parikh, R., Karlapalem, K., 2013. ET: Events from Tweets. In *Proceedings of International Conference on World Wide Web (WWW)* (Companion Volume). pp. 613–620.

Popescu, A. M., Pennacchiotti, M., Paranjpe, D. A. 2011. Extracting Events and Event Descriptions from Twitter. In *Proceedings of International Conference on World Wide Web.*

Prokopidis, P., Georgantopoulos, B. & Papageorgiou, H. 2011. A suite of NLP tools for Greek. In *The 10th International Conference of Greek Linguistics*. Komotini, Greece.

PROMAP Codebook. Forthcoming. In Stathopoulou T. (ed.) *Trasformations of protest in Greece.* (provisional title). Papazisis publishers, Athens.

Qin, Y., Zhang, Y., Zhang, M., Zheng, D. 2013. Feature-Rich Segment-Based News Event Detection on Twitter. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

Ritter, A., Mausam, Etzioni, O., Clark, S. 2012. Open Domain Event Extraction from Twitter. In *Proceedings of the international conference on Knowledge Discovery and Data mining*.

Schrodt, P. and Van Brackle, D. 2013. Automated Coding of Political Event Data. In Subrahmanian V. (ed) *Handbook on computational approaches to Counterterrorism*. New York: Springer.

Shrodt, P., Shannon, D., Weddle, J. 1994. Political Science: KEDS-A Program for the Machine Coding of Event Data. In *Social Science Computer Review*.

Walther, M., and Kaisser, M. 2013. Geo-spatial event detection in the twitter stream. In *Advances in Information Retrieval*. 356–367. Springer Berlin Heidelberg.

Wang, Y., Fink, D., Agichtein, E. 2015. SEEFT: Planned Social Event Discovery and Attribute Extraction by Fusing Twitter and Web Content. In *Proceedings of the International Conference on Weblogs and Social Media*.

Weng, J., Yao, Y., Leonardi, E., Lee, F. *Event Detection in Twitter*. HP Laboratories. HPL-2011-98.

Wueest, B., Rothenhäusler, K., Hutter. S. 2013. Using Computational Linguistics to Enhance Protest Event Analysis. In *SSRN Electronic Journal*.

Zachary, C. S., Mocanu, D., Vespignani, A., Fowler, J. 2015. Online Social Networks and Offline Protest. *EPG Data Science* 4:19.

Zubiaga, A., Ji, H. and Knight, K. 2013. Curating and contextualizing twitter stories to assist with social newsgathering. In *Proceedings of the 2013 international conference on Intelligent user interfaces*.