Towards Social Event Detection and Contextualisation for Journalists

Prashant Khare
Insight Centre for Data Analytics
National University of Ireland,
Galway, Ireland

prashant.khare@insight-centre.org

Bahareh Rahmanzadeh Heravi
Insight Centre for Data Analytics
National University of Ireland,
Galway, Ireland

bahareh.heravi@insight-centre.org

Abstract

Social media platforms have become an important source of information in course of a breaking news event, such as natural calamity, political uproar, etc. News organisations and journalists are increasingly realising the value of information being propagated via social media. However, the sheer volume of the data produced on social media is overwhelming and manual inspection of this streaming data for finding, aggregation, and contextualising emerging event in a short time span is a day-to-day challenge by journalists and media organisations. It highlights the need for better tools and methods to help them utilise this user generated information for news production. This paper addresses the above problem for journalists by proposing an event detection and contextualisation framework that receives an input stream of social media data and generates the likely events in the form of clusters along with a certain context.

1 Introduction

Social media platforms have evolved to being more than just a user-to-user interaction channel, and play a prominent role in real-time information sharing. In many cases the real life 'events' are now shared and broadcast on the social media platforms, by normal citizens, and not professional journalists. This has turned the former consumer [only] of the news into [also] a broadcaster of the news, and thus the social media platforms into an invaluable source of newsworthy information. The news organisations are now more and more interested in gathering real-time information (such as breaking news, images, videos) by means of monitoring and harvesting the user-generated content (UGC). Survey results reveal that journalists are increasingly using social media platforms for their professional activities. For example surveys reveal that 96% of journalists in the UK and use feeds from social media in their work on a daily basis (Cision, 2013), 99% of Irish journalists use social media as a source of information in their work (Heravi et al., 2014), and 51% of journalists globally leverage microblogs to consume feeds for news and stories (Oriella, 2013). With the increasing usage of social media in the journalistic processes, it is critical for journalists to be able to filter the social streams to discover breaking news, and then analyse, aggregate, contextualise, and verify them in timely manner.

The concept of Social Semantic Journalism, introduced by Heravi et al. (2012), targets the above problems encountered by the media organisations. The Social Semantic Journalism framework (Heravi and McGinnis, 2013) utilises the social and semantic web technologies, and provides an integrated view for enhancing newsworthy information discovery, filtering, aggregation, verification and publication. While there is considerable work done to retrieve information from various sources of data (such as text) by various means, there is a paucity of tools available for detecting events from social media data and extracting relevant information about such events in the real time. Building upon the ideas of Social Semantic Journalism, to aid journalists in utilising UGC in an efficient manner, this paper proposes a framework that implements an event detection pipeline, which *clusters the data into different events,* and *determines the context of the events based on entities* (mentions particular to any person, place, event, or thing) related to the events. The information that flows on the social media is often via textual medium, and therefore in this proposed framework, we leverage text mining and Natural Language Processing (NLP) technologies to extract the information.

The remainder of this paper is organised as follows. Section 2 provides a background to the problem and briefly reviews related work. Section 3 presents our proposed Event Detection and Contextualisation framework and gives a detailed overview of its components and phases. Section 4 concludes the paper and discusses directions for future research.

2 Background and Related Work

Identifying new events, in the form of news from the data, is an area of interest for researchers for a long time. Topic detection and tracking (TDT) (Allan, 2002) focuses on breaking down a streaming text from newswire into smaller cohesive news pieces and determining if something has not been earlier reported. An event detection cycle is seen as a subtask within TDT (Allan, 2002). The data from social media platforms, such as Twitter, is quite voluminous and the streaming nature of this data warrants the usage of streaming algorithm models, where the data arrives in a chronological order (Muthukrishnan, 2005). The social media data is further processed in a bounded space and time, i.e. as every entry arrives it gets processed. Traditional approaches for identifying new information (an event) were to compare each new entry in the data with the previously arrived entries. Petrovic et al. (2010) investigated ways to identify tweets that first report the occurrence of an event by clustering mechanism to identify nearest neighbours in the textual data. This work has motivated many other contemporary research works to head in a related direction

Osborne et al. (2012) used the approach by Petrovic et al. (2010) as a baseline and investigated the ways to improve the event detection mechanism on Twitter data, by matching the frequency of newly occurring events from tweets with the activity (number of visits on a page) of the corresponding pages of entities from Wikipedia and analysed if there was a similar pattern observed while determining an event. Parikh and Karlapalem (2013), also considering frequency based analysis, developed an event detection system that extracts events from tweets by examining frequencies in the temporal blocks of streaming data.

Natural Language Processing (NLP) techniques can be leveraged in detecting events from voluminous social media data. Events are associated with entities and NLP techniques can be applied to extract the entities that are mentioned in the text that defines an event. To perform Named Entity Recognition (NER) on tweets Ritter et. al. (2011) redeveloped the taggers and segmenters of Stanford NLP library1. Ritter et al. (2012) extending the above work created an application Twical, that extracted an open domain calendar for events that were shared on Twitter.

For an event detection system, it is also crucial to determine the context of a piece of text/information. The contextualisation is answering the question ‘what is this about?’ and one of the ways to answer it could be by aggregating information from knowledge base such as Wikipedia (SanJuan et al., 2012). A potential context of the content can likely be inferred by extracting set of topics that bound the text. Hulpus et al. (2013) proposed an approach by linking the topics inherent to a text with the concepts in DBpedia[2] and thereby automatically extracting the topic labels from the corpus. Meij et al. (2012) extracted underlying concepts of a text from a large knowledge base of Wikipedia articles by applying a supervised learning using a Naive Bayes (NB), Support Vector Machines (SVM), and a C4.5 decision tree classifier. Large knowledge bases, such as YAGO[3], are also used (Hoffart et al., 2013) to explore the inherent relationship between entities and disambiguate them to derive the context. Taking insights from various research works briefed above, we aim to construct a framework that is inspired by ideas from different works in the next section.

3 The Event Detection & Contextualisation Framework

There are various approaches to extract the information from data, by means of clustering, entity extraction and contextualisation, yet there is no observed pipeline that incorporates different methods and brings them under one framework so as to to generate insights from streaming social media data.

[1] http://nlp.stanford.edu/software/corenlp.shtml
[2] http://dbpedia.org/About
[3] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

We aim to address this gap, by proposing a framework that performs the aforementioned functionalities under one system. A complete illustration of the framework is visualised in the Figure 1 (further explained in detail). It is a pipeline that incorporates several components, each followed by another phase that uses the output from the previous one. The data could potentially come from various social media APIs; however we have focused on data collected from Twitter streaming API[4]. Following sections explain the different phases, in order, that process the input streaming data: *indexing and clustering*, *entity recognition*, and *entity disambiguation and contextualisation*.

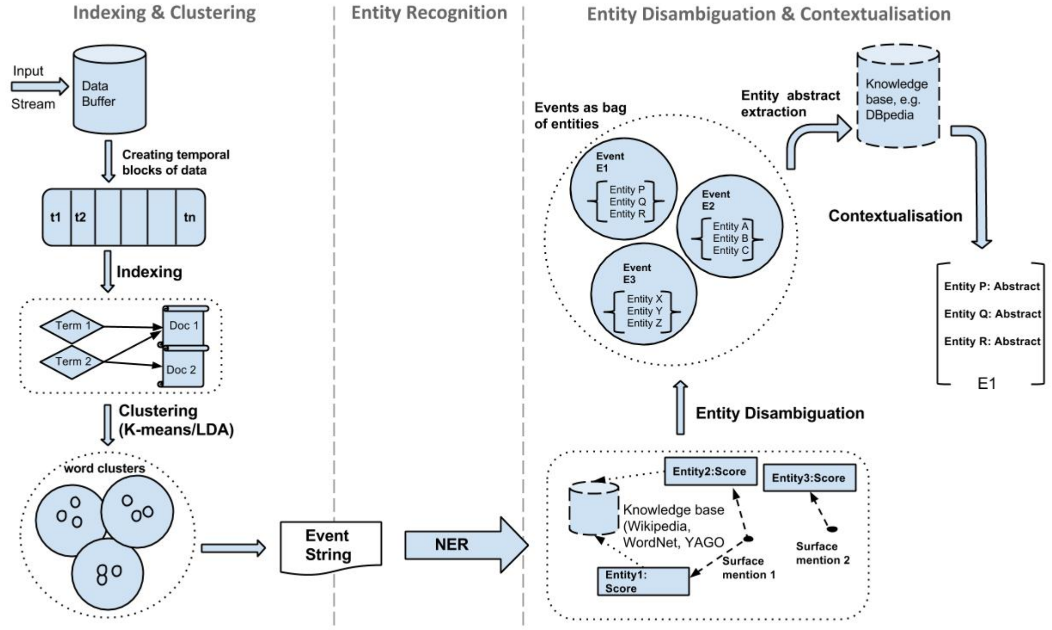

Figure 1. Event detection and Contextualisation framework

3.1 Indexing and Clustering

This phase is aimed at pre-processing the data, breaking the data into set of keywords and generating an index that maps words against their corresponding document. Once an index is created, the data is clustered as sets of word vectors occurring together prominently. These clusters tend to represent the events that exist in the data.

Indexing: The data is indexed in this sub-phase. The incoming data stream is stored and then the divided in slabs of time windows (say of 10 minutes each). This is done to analyse the data based on regular time intervals, which may result in inferring only the highly dominating events/clusters present in the data. An index, between terms and corresponding documents (that initially contained those terms), is generated for this slab of data using standard libraries such as Lucene[5] and Solr[6] (built over Lucene).

Clustering: In this sub-phase we derive the preliminary clusters of the data, which are likely to reflect the most related content within the data slab that was earlier created. Examples of the clustering algorithms that can be hired to cluster the data are k-means, PLSA (Hofmann, 1999) and LDA (Blei et al., 2003). After the clusters are formed, the terms with high weight in the clusters are taken to query the index for retrieving the most relevant documents based threshold relevance score. The relevance score is derived from term frequency and inverse document frequency (*tf-idf*) (Manning et al., 2008) value and accordingly the documents are retrieved. The text from those top scored documents can now be extracted and merged into one string, hereafter called *event string*, which tends to represent the infor-

[4] https://dev.twitter.com/docs/api/streaming
[5] http://lucene.apache.org/
[6] http://lucene.apache.org/solr/

mation stored against a particular cluster or event. This *event string* is further used for NER and disambiguation.

3.2 Entity Recognition

The *event string,* derived above, is further annotated for its entities by applying Named Entity Recognition techniques. NER is an information extraction task to extract key elements, hereafter referred to as *Entities,* from a text and categorise them into person, location, organisation, etc. In this work we rely on libraries such as Stanford NER (Finkel et al., 2005) or other wrappers to this library, which implement it to extract named entities. However, there are other libraries available for this purpose, for instance, Open NLP[7], Open Calais[8], etc. A detailed explanation of the NER models is given in the research work by Sang and Meulder (2003) and Finkel et al. (2005).

For each mention in the string there can be multiple candidate entities which further need to be disambiguated. An explanation of it could be given with an example, such as in "David was playing for Manchester United when Victoria gave her auditions. Victoria later became part of band Spice Girls": how could it be determined whether Victoria is a person (particularly Victoria Beckham) and not Victoria- a place or Queen Victoria, and David implies David Beckham and not David - a figure in religious text/history. Establishing such a mapping between mention and most relevant entity is termed as named entity disambiguation process.

3.3 Entity Disambiguation and Contextualisation

In the entity disambiguation and contextualisation phase, initially an input text (web page, language paragraph, sentence, article) is resolved into various mentions of entities (surface mentions- that means its just a mention with no associated knowledge) by matching all the potential candidate entities with the surface mentions. For this purpose Stanford NER tagger is used. For each mention (a potential entity) knowledge sources such as DBpedia and/or Yago (Hoffart et al., 2013; Hoffart et al., 2011) are harvested to extract potential entity mentions. Each mention will then be mapped for numerous potential entity candidates. After extracting the candidate entities, a relevance score can be assigned to each based on features such as *a prior for candidate entity popularity*, *mutual information* (similarity between key-phrase or query string and description of the entity), *syntax based similarity* (Thater et al., 2010), *entity-entity coherence* (quantifying the number of similar incoming links on a knowledge base as Wikipedia). Milne and Witten (2008) extended few similarity measures defined by Bunescu and Pasca (2006), which compared the context of a given text to the entities mention in Wikipedia.

Considering the above features, a graph of mentions and candidate entities, with the edges as weights, can be generated. Each node will have a certain weight on its edge, a greedy approach can be adopted to iteratively remove the low weight nodes to disambiguate the entities (Hoffart et al., 2011). This approach will result in disambiguated entities (to a high degree) for each surface mentions of the input text and represent entities according to the context of the input text. After the disambiguation of the entities, a knowledge resource can be hired to query for generating a brief description about the prominent entities (such as their *abstract*/*description* and *type*), and thereby contextualising the whole input text with a bag of entities and their corresponding description.

The overall framework describes a mechanism to design a tool that can process input streaming data into set of clusters that reflect events and assists in visualising the context of those events. This framework is considered to enhance event detection approaches by enriching the events with their relevant information being extracted from knowledge resources. While some of the state of the art techniques and tools incorporated in this framework have been proposed and/or utilised in other domains, the proposed framework is a novel end-to-end pipeline specifically designed for the news industry and for breaking event detection and contextualisation.

4 Conclusion and Future Work

This paper presents a framework, which aims at assisting journalists in dealing with the ever- flooding UGC to detect the upcoming/breaking events. Various surveys (Oriella, 2013; Cision, 2013; Heravi et.

[7] https://opennlp.apache.org/

[8] http://www.opencalais.com/

al., 2014) highlight the growing need for specialised tools to allow journalists utilise the user-generated for news production and storytelling. The proposed framework is believed to be an important step forward in addressing the challenges encountered by journalists in leveraging the social media content for emerging event detection and event contextualisation in the process of news production. The emerging events can now be visualised without needing to manually assess the frequency of any particular information propagation on social media and also generate the context of the information at the same time.

An early phase test was performed on the proposed pipeline so as to assess the viability of the framework. The framework was simulated with a sample data constituting of tweets from three different known events and it reflected encouraging results with respect to the viability of the underlying processes and the framework as a whole. The framework successfully clustered the sample data, using *k-means* algorithm, into unique clusters and the entity disambiguation phase, implemented using AIDA framework (Hoffart et al., 2011), yielded relevant entities. An end-to-end evaluation of the pipeline, however, is yet to be performed to analyse the results of every phase, and the pipeline as a whole.

There are foreseen challenges such as noise filtering, the non-lexical nature of the data, and the verity of the content. The data from social media contains an enormous amount of noise (such as random tweets posted by users which do not have a relevance with the event and may yet contain the filtering keywords) in exhaustive social media streams when it comes to filtering the content specific to certain events/topics and that could certainly affect the outcome of the event clusters. Apart from noise, often the language used on social media is non-lexical and non-syntactic in nature because users compromise with the language rules to share more information in limited space (e.g. Twitter allows only 140 characters) hence leveraging the NLP techniques may not result in most efficient results.

The above challenges require a thorough investigation of the current state of the research and as a future work we aim to address 1) perform an end-to-end evaluation on the pipeline and 2) address the above challenges by exploring how information extraction techniques can be customised for syntactically and lexically inefficient data and thereby refine the information gathering processes for journalists.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers 12/TIDA/I2389, SFI/12/RC/2289 and 13/TIDA/I2743.

References

Allan, J. 2002. Introduction to topic detection and tracking. In Topic detection and tracking (pp. 1-16). Springer US.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Bunescu, R. C., & Pasca, M. 2006. Using Encyclopedic Knowledge for Named entity Disambiguation. In EACL (Vol. 6, pp. 9-16).

Cision. 2013. Social Journalism Study 2013. Report by Cision & Canterbury Christ Church University (UK). http://www.cision.com/uk/wp-content/uploads/2014/05/Social-Journalism-Study-2013.pdf visited July 13th,2014

Finkel, J. R., Grenager, T., & Manning, C. 2005, June. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 363-370). Association for Computational Linguistics.

Heravi, B. R., Boran, M., & Breslin, J. 2012. Towards Social Semantic Journalism. In Sixth International AAAI Conference on Weblogs and Social Media.

Heravi, B. R., & McGinnis, J. 2013. A Framework for Social Semantic Journalism. In First International IFIP Working Conference on Value-Driven Social & Semantic Collective Intelligence (VaSCo), at ACM Web Science.

Heravi, B. R., Harrower, N., Boran, M. 2014. Social Journalism Survey: First National Study on Irish Journalists' use of Social Media. HuJo, Insight Centre for Data Analytics, National University of Ireland, Galway (forthcoming 20 July 2014).

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G. 2011. Robust disambiguation of named entities in text. InProceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 782-792). Association for Computational Linguistics.

Hofmann, T. 1999. Probabilistic latent semantic analysis. InProceedings of the Fifteenth conference on Uncertainty in artificial intelligence(pp. 289-296). Morgan Kaufmann Publishers Inc.

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 194, 28-61.

Hulpus, I., Hayes, C., Karnstedt, M., & Greene, D. 2013. Unsupervised graph-based topic labelling using DBpedia. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 465-474). ACM.

Manning, C. D., Raghavan, P., & Schütze, H. 2008. Introduction to information retrieval (Vol. 1, p. 6). Cambridge: Cambridge university press.

Meij, E., Weerkamp, W., & de Rijke, M. 2012. Adding semantics to microblog posts. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 563-572). ACM.

Milne, D., & Witten, I. H. 2008. Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 509-518). ACM.

Muthukrishnan, S. 2005. Data streams: Algorithms and applications. Now Publishers Inc.

Oriella. 2013. The New Normal for News: Have global media Changed forever Oriella PR Network Global Digital Journalism Study 2013. Available from http://www.oriellaprnetwork.com/sites/default/files/research/Brands2Life_ODJS_v4.pdf visited July 13th,2014

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., & Ounis, I. 2012. Bieber no more: First story detection using Twitter and Wikipedia. InProceedings of the Workshop on Time-aware Information Access. TAIA (Vol. 12).

Parikh, R., & Karlapalem, K. 2013. Et: events from tweets. InProceedings of the 22nd international conference on World Wide Web companion (pp. 613-620). International World Wide Web Conferences Steering Committee.

Petrović, S., Osborne, M., & Lavrenko, V. 2010. Streaming first story detection with application to twitter. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 181-189). Association for Computational Linguistics.

Ritter, A., Clark, S., & Etzioni, O. 2011. Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1524-1534). Association for Computational Linguistics.

Ritter, A., Etzioni, O., & Clark, S. 2012. Open domain event extraction from twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1104-1112). ACM.

Thater, S., Fürstenau, H., & Pinkal, M. 2010. Contextualizing semantic representations using syntactically enriched vector models. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 948-957). Association for Computational Linguistics.

Tjong Kim Sang, E. F., & De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: Language independent named entity recognition. InProceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 142-147). Association for Computational Linguistics.

Zaki, M. J., & Meira Jr, W. 2011. Fundamentals of data mining algorithms.