International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869, Volume-2, Issue-9, September 2014

Humanitarian Applications of Big Data

Prof. (Mrs.) Sindhu Nair, Mr. Neel Shah, Mr. Pinank Shah

Abstract— **Archives of human rights violation reports obscure fine grain analysis of event patterns due to their poor metadata and basis in natural language. Analysing a huge web of data obtained from different sources is challenging. The transition of these small-scale data to big data analysis is crucial. It is very difficult to conduct cross-document analysis of all these events and come to the right conclusion about the reality of these events. This paper discusses these issues and proposes a framework to address these challenges. Though the method we have studied is used for analyzing human rights violations, it can be used for other big data problems in humanities as well.**

***Index Terms*—Digital Humanities, Big data, Storygram**

I. INTRODUCTION

Records pertaining to human rights violation are very difficult to analyse as the amount of data is huge and there is no hundred per cent certainty in any of the sources [1]. The records are of heterogeneous origins, collecting material produced by civilians, NGOs and governments, and by observers, victims and perpetrators. They are produced both during the event and after an event when the witnesses come out to speak out against violators. The records may contain interrogation transcripts, observation reports by professional observers, information about various geospatial sites etc.

The data from the above mentioned factors vary and thus it is difficult to reach to a final conclusion. How does one create meaning from these records collections of a huge scale? Although human rights corpus are smaller than the datasets addressed as big data [2], various features of human rights like the high dimensionality of this information, the heterogeneity, the requirement of real-time analysis etc. make this a big data problem.

II. METHOD

Our framework is designed to process large numbers of narratives describing time, location, and person, by using big data analysis. This process can also identify perpetrators and victims on the basis of their linguistic and narrative structures. We propose a two tiered framework consisting of Data and Presentation layers. The task of Data Layer is responsible for data extraction, parsing and running NLP modules to get the entities and events. The Presentation Layer handles the visualization of data and the feedback loop which modifies the data based on the user feedbacks.

Prof. (Mrs.) Sindhu Nair, Mr. Neel Shah, Mr. Pinank Shah, D.J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

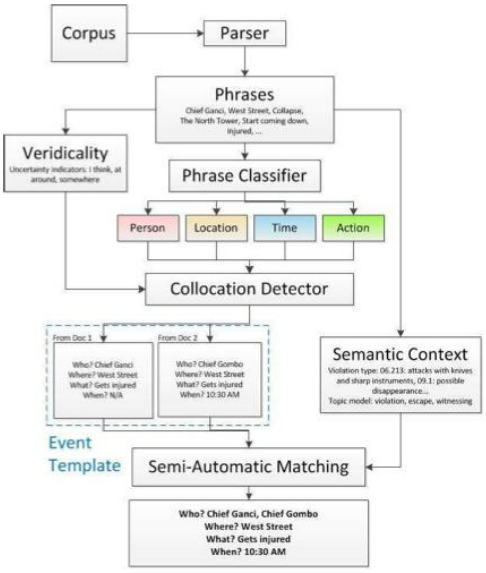

Figure 1. Model used in the NLP module for cross-referencing the entitites.

A. DATA LAYER

This layer focuses on how the data is extracted, processed, and stored in the backend, supporting the presentation layer. The extraction module is responsible for digitizing and parsing the documents. A large number of human rights documents that we obtain are usually stored in hard copy format and, in many cases, handwritten. OCR tools like OmniPage are used to digitize the documents. After digitizing, the document is parsed and the output is sent to the NLP module. The goal of the NLP module is to facilitate cross-document co-reference of entities in collections of witness statements and interviews. This module decides whether or not two mentions of entities refer to the same person, by considering exophora and relies on placing pronominal entities within a high-order Event Storygram of time [3], name, location and semantic context.

i.) Event summarization based on matching process:

Here the main is to situate entities in the series of events that define their appearances. Who, What, When, Where and Why, are the phrases useful for this process [4], [5]. The goal is to frame a system that can automatically extract these important entities as phrases and recognize the duplicate entities across documents based on these extract. Then a phrase classifier is used to categorize these phrases such as Geographic Location/Date/Time/Person Names. Then a module named, Collocation Detector, detects which of the entities are described in the same context within a passage in the corpus. A collocation is only true when multiple elements correlate. After this the collocated phrases are placed into the event template and fed into a visualization engine. Human

observes then decide which cross-document entities are identical. Scores are then assigned to enable observers on which events and entities should be merged.

ii.) Phrase extraction and classification:

The first step of phrase extraction involves running a full parser on each document and then extracting all the retrieved noun phrases and verb phrases from the parse tree. The parser we plan to use is Stanford parser [6]. Important phrases are then labeled for event extraction. The phrases are classified into 8 categories: Organization, Person, Location, Date, title, Time, Event, Miscellaneous, and the background category of Unimportant. Phrases such as "the pedestrian bridge", or "the ferry" which are not identifiable as a particular named entity but might be crucial in depicting the event are classified as Miscellaneous. For this reason the phrases need to be ranked in a particular order based on its frequency in the data set.

$$S_c(\text{phrase}) = \log P(\text{phrase}) - \max(\log P_{bg}(\text{phrase}) - \log P(\text{phrase}), 0)$$

In this model the probability of a phrase is only discounted if $P_{bg}(\text{phrase}) > P(\text{phrase})$.

Eventually for the collocation of the phrases we use a simple metric: A Gaussian kernel on the distance between mention of different phrases. It is defined as:

$$P(p_1|p_2,\ldots,p_k) = exp(-\beta \sum_i (S(p_1) - S(p_2))^2)$$

where $S(p_i)$ is the sentence number where pi occurred. Given the defined conditionals, one can compute the joint probability P(p1,p2,p3,...,pk) and use a threshold to determine which phrase set goes to an event template.

B. PRESENTATION LAYER

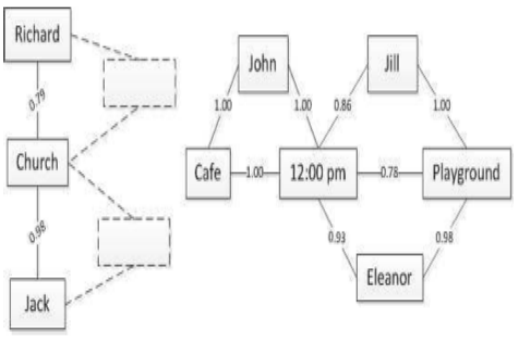

Figure 2. A model of Storygram showing events and uncertainty within the entities of the event.

This layer consists of the visualization module and the feedback module. The visualization module consists of Storygraphs [7] and Storygrams. Storygraph is a 2D visualization technique for presenting time and location on the same chart. It consists of two parallel vertical axes, which are used for latitude and longitude, and an orthogonal horizontal axis, which is used for time. An event, E (lat, lng, time) is mapped in into the Storygraph by first drawing a line segment connecting the corresponding latitude and longitude in two vertical axes. A marker is then placed on the line above the corresponding time of the event. Any additional attributes like the type of event can be shown by changing the shape, size and the colour of the marker. Various uncertainties include locative like "I guess that would be South End Avenue", temporal like "at that time I noticed" and entity like "At this point I had my five guys", are presented by using Storygrams.

A Storygram is a 2D-planar diagram consisting of events as its building blocks. An event in this case is a 3-tuple consisting of time, location and entity. Storygrams are represented as triangles using the entities as vertices connected with weighted edges. The weights represent the confidence value in the relationship obtained from the NLP module and this value represents the connection between two elements. Figure 2 shows a trigram with confidence values and elements.

Iii.) Future Scope

1.) This above method can further have a host of other applications. It can be used in solving crimes and also in anti-terrorism activities. For example, data mining is used by the NATO (North Atlantic Treaty Organization) [8] for counter terrorism. The project seeks to provide tools for data preparation as well as importance ranking of data elements based on an original model of information value.[1] Thus, this model that we have analyzed, if used, can be beneficial for not just humanitarian applications but in a plethora of domains.

2.) Novel Intelligence from Massive Data (NIMD) [8] is a system incorporated by the United States government for anti-terrorism activities. Agencies like TIA, NIMD seeks both to bring together information from a variety of data sources and to assist human analysts in overcoming natural limits and failures in human cognition so that they may recognize the significance of intelligence data and evaluate it properly. The above method if used for NIMD can improve its efficiency as well.

IV. CONCLUSION

In this paper we explored the challenges faced by researchers and analysts trying to study human rights violations and also studied a model to solve the same. This framework included elements for processing the textual data, visualizing that data, and feeding judgments made by users of the framework back into the processing layer.

REFERENCES

[1] M. L. Best, W. J. Long, J. Etherton, and T. Smyth, "Rich digital Media as a tool in post-conflict truth and reconciliation," *Media, War & Conflict*, vol. 4, no. 3,pp. 231-249, 2011.

[2] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a culture, technological, and scholarly phenomenon," *Information, Communication & Society* , vol. 15, no. 5,pp. 662-679, 2012.

[3] C. Northwood, "Ternip: temporal expression recognition and normalization in python," Ph.D. dissertation, Masters thesis, University of Sheffield, 2010.
[4] P. Ball, J. Asher, D. Sulmont, and D. Manrique, "How many Peruvians have died?" American Association for the Advancement of Science, Tech. Rep., 2003.
[5] P. Ball, E. Tabeau, and P. Verwimp, "The Bosnian book of dead: Assessment of the database (full report)," Households in Conflict Network, Tech, Rep., 2007.
[6] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423-430.
[7] A. Shrestha, Y. Zhu, B. Miller, and Y. Zhao, "Storygraph: Telling stories from spatio-temporal data," in *Lecture Notes in Computer Science*, vol. 8034. Springer, 2013, pp. 693-703.
[8] Dr. Daniel Moeckli, University of Zurich; James Thurman, University of Zurich: Seventh Framework Programme: Survey of Counter-Terrorism Data Mining and Related Programmes

Prof (Mrs.) Sindhu Nair- Assistant Professor at DJ Sanghvi College of Engineering. Prof. Nair has pursue her BE and ME in computers from Mumbai University. She has a teaching experience of 10 years. And an experience of 4 years as a software developer. Prof Nair has authored 3 national and 3 international papers. She is a member of IETE.

Neel Shah is currently pursuing his final year of computer Engineering at DJ Sanghvi College of Engineering.

Pinank Shah is currently pursuing his final year of computer Engineering at DJ Sanghvi College of Engineering