# Semantic Concept Discovery Over Event Data

Oktie Hassanzadeh, Shari Trewin, and Alfio Gliozzo

IBM Research, USA

Preparing a *comprehensive*, *accurate*, and *unbiased* report on a given topic or question is a challenging task. The first step is often a daunting discovery task that requires searching through an overwhelming number of information sources without introducing bias from the analyst's current knowledge or limitations of the information sources. A common requirement for many analysis reports is a deep understanding of various kinds of historical and ongoing *events* that are reported in the media. To enable better analysis based on events, there exist several *event databases* containing structured representations of events extracted from news articles. Examples include GDELT [4], ICEWS [1], and EventRegistry [3]. These event databases have been successfully used to perform various kinds of analysis tasks, e.g., forecasting societal events [6]. However, there has been little work on the discovery aspect of the analysis, that results in a gap between the information requirements and the available data, and potentially a biased view of the available information.

In this presentation, we describe a framework for concept discovery over event databases using semantic technologies. Unlike existing concept discovery solutions that perform discovery over text documents and in isolation from the remaining data analysis tasks [5, 8], our goal is providing a unified solution that allows deep understanding of the same data that will be used to perform other analysis tasks (e.g., hypothesis generation [7] or building models for forecasting [2]). Figure 1 shows the architecture of our system. The system takes in as input a set of event databases and RDF knowledge bases and provides as output a set of APIs that provide a unified retrieval mechanism over input data and knowledge bases, and an interface to a number of concept discovery algorithms. Figures 2 shows different portions of our system's UI that is built using our concept discovery framework APIs. The analyst can enter a natural language question or a set of concepts, and retrieve collections of relevant concepts identified and ranked using different concept discovery algorithms. A key aspect of our framework is the use of semantic technologies. In particular:

– A unified view over multiple event databases and a background RDF knowledge base is achieved through semantic link discovery and annotation.
– Natural language or keyword query understanding is performed through mapping of input terms to the concepts in the background knowledge base.
– Concept discovery and ranking is performed through neural network based semantic term embeddings.

We will present the results of our detailed evaluation of our proposed concept discovery techniques. We prepared a ground truth from reports on specific topics written by human experts, including reports from the Human Rights Watch or-

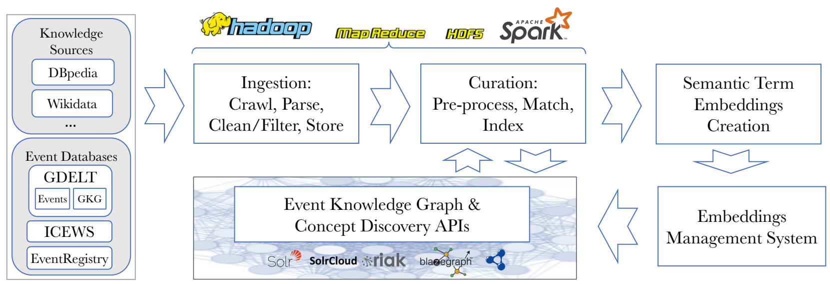


Fig. 1: System Architecture

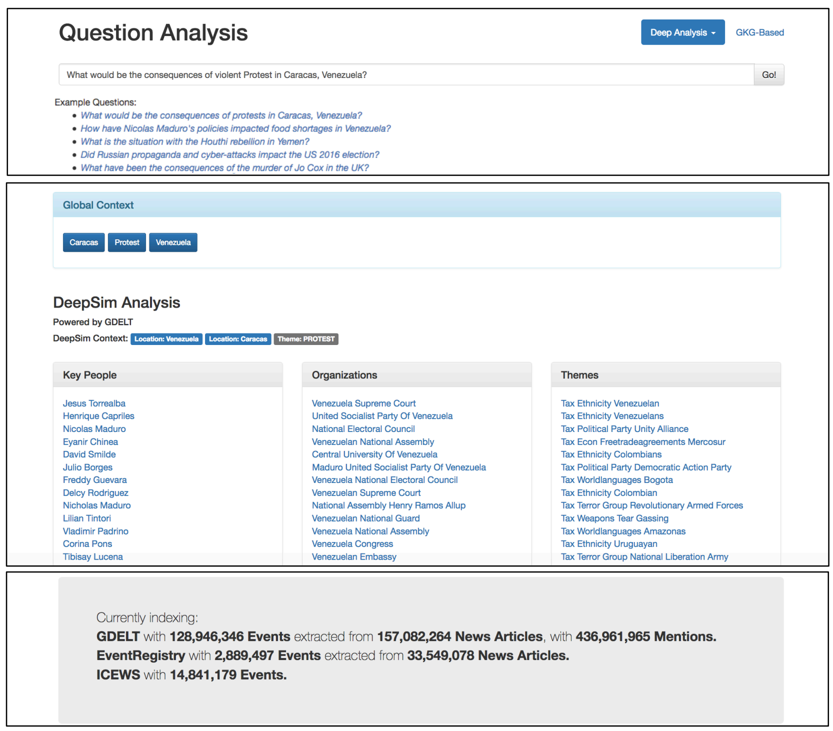


Fig. 2: Views from the Question Analysis UI

ganization, and Wikipedia pages on people and events. The ground truth queries included hand-built test queries on various topics, and an automatically generated set of queries based on the title of the reports. Given only these query terms, we measure the ability of different algorithms to find the concepts mentioned in the original reports. Our study finds that combining our neural network based semantic term embeddings over structured data with an index-based method can significantly outperform either method alone.

## References


1. Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., Ward, M.: ICEWS Coded Event Data (2017), http://dx.doi.org/10.7910/DVN/28075
2. Korkmaz, G., Cadena, J., Kuhlman, C.J., Marathe, A., Vullikanti, A., Ramakrishnan, N.: Combining heterogeneous data sources for civil unrest forecasting. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. pp. 258–265. ASONAM '15 (2015), http://doi.acm.org/10.1145/2808797.2808847
3. Leban, G., Fortuna, B., Brank, J., Grobelnik, M.: Event Registry: Learning About World Events from News. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 107–110. WWW '14 Companion (2014), http://doi.acm.org/10.1145/2567948.2577024
4. Leetaru, K., Schrodt, P.A.: GDELT: Global data on events, location, and tone, 1979–2012. In: ISA Annual Convention (2013)
5. Lin, D., Pantel, P.: Concept Discovery from Text. In: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1. pp. 1–7. COLING '02 (2002), http://dx.doi.org/10.3115/1072228.1072372
6. Muthiah, S., Butler, P., Khandpur, R.P., Saraf, P., Self, N., Rozovskaya, A., Zhao, L., Cadena, J., Lu, C.T., Vullikanti, A., Marathe, A., Summers, K., Katz, G., Doyle, A., Arredondo, J., Gupta, D.K., Mares, D., Ramakrishnan, N.: Embers at 4 years: Experiences operating an open source indicators forecasting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 205–214. KDD '16 (2016), http://doi.acm.org/10.1145/2939672.2939709
7. Sohrabi, S., Udrea, O., Riabov, A.V., Hassanzadeh, O.: Interactive Planning-Based Hypothesis Generation with LTS++. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. pp. 4268–4269 (2016), http://www.ijcai.org/Abstract/16/654
8. Tan, A.h.: Text Mining: The state of the art and the challenges. In: In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. pp. 65–70 (1999), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.6973