THOR: Text-enabled Analytics for Humanitarian Operations

Mayank Kejriwal

USC Information Sciences Institute
Marina Del Rey, California
kejriwal@isi.edu

Pedro Szekely

USC Information Sciences Institute
Marina Del Rey, California
pszekely@isi.edu

Daniel Gilley

Next Century Corporation
Annapolis Junction, Maryland
daniel.gilley@nextcentury.com

Jill Crisman

Next Century Corporation
Annapolis Junction, Maryland
jill.crisman@nextcentury.com

ABSTRACT

In this demonstration, we present the Text-enabled Humanitarian Operations in Real-time (THOR) framework, which is being prototyped to provide visual and analytical *situational awareness* to humanitarian and disaster relief (HADR) planners. THOR is a collaborative effort between industrial and university research laboratories, designed with an intent to support both military and civilian HADR operations. At its core, THOR is powered by a domain-specific knowledge graph, which is derived from natural language outputs and is amenable to real-time analytics. THOR is designed to operate in low-resource linguistic environments, process heterogeneous data, including news and social media, reason about arbitrary disasters not knowable in advance, and provide advanced graphical interaction capabilities. We will demo the latest prototype of THOR using an interactive case study situation.

KEYWORDS

Low-resource languages; humanitarian operations; entity resolution; knowledge graphs; visualization; situational awareness

ACM Reference Format:

Mayank Kejriwal, Daniel Gilley, Pedro Szekely, and Jill Crisman. 2018. THOR: Text-enabled Analytics for Humanitarian Operations. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3184558.3186965>

1 INTRODUCTION

The United Nations Office for the Coordination of Human Affairs (OCHA) reported¹ that in 2016, more than a hundred million people were affected by natural disasters alone, while over sixty million people were forcibly displaced by violence and conflict. Using interdisciplinary research from the data sciences to build a visualization and analytics platform for US civilian and military rescue personnel, as well as the intelligence community, is an important problem with the potential for widespread long-lasting social impact.

¹<http://www.unocha.org/datatrends2016/WHDT2016.pdf>

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '18 Companion, April 23–27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5640-4/18/04.

<https://doi.org/10.1145/3184558.3186965>

The DARPA LORELEI program² was established with the explicit agenda of providing *situational awareness* for *emergent incidents*, under the assumption that the emergent incident occurs in a region of the world where the predominant language is computationally *low-resource* [15]. An example is Uyghur, a Turkic language spoken by about 10-25 million people in Western China, for which few automated technology capabilities currently exist [2]. The technologies resulting from LORELEI research will be capable of supporting situational awareness based on low-resource foreign language sources within an extremely short time frame (about 24 hours after a new language requirement emerges).

The notion of situational awareness involves far more than *machine translation* of the low-resource language into English [19]. In addition to *natural language processing* on the translated text (e.g. named entity recognition, sentiment analysis [22], [23]), solutions to problems such as *entity resolution* and *provenance reasoning* are required to support detailed analytics that are ultimately visualized on a graphical user interface (GUI) [11], [18].

We demonstrate the *text-enabled humanitarian operations in real-time* (THOR) framework that is designed to tackle a subset of the challenges proposed by the LORELEI program. Figure 1 provides a contextualized illustration of THOR as part of a broader solution currently under development. The input to THOR is a *streaming* corpus of raw documents and a set of natural language processing (NLP) modules collectively denoted as the *Language Technology Development Environment* (LTDE) that includes state-of-the-art implementations for NLP services such as machine translation and named entity recognition [19], [22]. Note that LTDE is not a part of THOR. The LTDE outputs are used to structure *each* raw document in the stream into a *situation frame*, serialized as a semi-structured JSON document that contains a combination of unstructured and structured data, and metadata. These JSON documents are then visualized in an interactive GUI to support both search and analytics.

Figure 2 presents the tile-based dashboard. Among the analytics that are rendered in the THOR GUI are *situation frames*, which classify each JSON document as expressing one or more needs in pre-set categories such as food or water. Along with facilities such as spatio-temporal rendering, these ‘situations’ are ultimately designed to help field analysts decide *where*, *when* and *how* to allocate resources to meet current situational needs.

²<http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents>

Figure 1: The overall architecture in support of LORELEI goals, of which THOR (expanded view) and LTDE are both core components. THOR supports both graphical and text outputs; however, we focus on the GUI in this demonstration.

Figure 2: The THOR dashboard, with the visualization describing a real-world Haiti dataset. In the demonstration, we will take the user through the navigation and analytical facilities, using resource allocation as the target problem.

THOR is expected to support two different user-facing services. The first service, which will not be covered by this demo, is a direct product of the analyses tools and comprise machine generated reports such as mission plans and status reports. The second service is an advanced graphical user interface that is expected to support personnel to visually navigate diverse, spatio-temporally evolving crisis scenarios in low resource regions. An anticipated demonstration workflow is described in Section 2.

Innovations. Specific innovations in THOR are as follows. First, THOR is an integrative solution that unifies many different aspects of the crisis informatics problem in a single *knowledge graph*-centric framework. Second, THOR is meant to serve a variety of users in the real-world, ingests documents from multiple Web (and also non-Web) sources such as Twitter, news, text messages and Ushahidi, and is the result of a multi-organizational collaboration. Third, the THOR GUI is multi-faceted and customizable, including support for maps, charts, networks, and word clouds, along with providing both analytical and entity-centric dashboards. The GUI will be a central focus of the demonstration. Fourth, THOR expresses an innovative and visual way of rendering (potentially noisy) NLP outputs on a screen to users who do not necessarily know about the inner workings of such technology, but can still use the outputs to accomplish useful tasks (such as deciding when and where to allocate aid, in a foreign, low-resource environment).

Significance to the Web. *Crisis informatics* is an emerging field with a significant intersection with the Web, due to the proliferation of social media and real-time information channels [25], [5], [9], [4], [31], [26], [14], [27], [12]. Although multiple computational approaches can be used for gaining better situational awareness in crisis situations, practitioners still tend to rely on methods and tools that are either *manual* or *non-adaptive* (e.g. keyword search, copy-pasting documents), or are too general for the domain-specific needs of crisis situations. THOR presents an important advance in this respect, and is also an example of using Web data and advancements for powering AI systems for social good.

2 DEMONSTRATION WORKFLOW

Earlier, we presented an illustration of the THOR GUI and dashboard in Figure 2. In the actual demonstration, we will use the problem of *resource allocation* as our contextual backdrop. Namely, we will assume that we are in the still-evolving aftermath of a crisis situation, like an earthquake in the Haiti region. With text messages, and news, communications and intelligence fragments emerging from multiple sources in the aftermath of the earthquake, an aid provider seeks rapid situational awareness and is looking to answer questions such as *What are the focal (i.e. most affected) and hot-spot regions? What are the resources (e.g., food, water, shelter etc.) required in such regions, and how should they be allocated? Who are the primary political entities involved in the region?*

We will also use a second dataset involving earthquakes in the Turkic region of China to illustrate the problems that can arise due to automatically translating a challenging, computationally under-studied (i.e. low-resource) language like Uyghur (Figure 3). For such languages, not only is translation difficult, but other NLP processes like entity and relationship extraction and recognition are also difficult and prone to noise. For such situations, visualization

Figure 3: The THOR dashboard for the Uyghur dataset.

is particularly important as the user is able to abstract away the effects of the noise and still make useful decisions.

Finally, we will demonstrate how THOR seamlessly integrates with multiple LTDEs. This is facilitated by the API- and knowledge graph-centric architecture of THOR. Any LTDE that delivers its output using JSON and a simple pre-defined schema can be ingested and visualized by THOR. This implies that an LTDE does not need to be ‘frozen’ in order to be processed by THOR, but can continue to undergo improvements and organic development. Many of our partner collaborations in the LORELEI program expose their LTDE ‘pipelines’ as APIs on the Web, which means that the code does not have to be released or integrated either. Thus, the various components of THOR can fully interoperate in classified and proprietary environments, which can be an important requirement in sensitive regions of the world.

Infrastructure and Materials. The THOR system (including backend and frontend services) are currently running on a program-approved facility in the United States. The GUI itself can be rendered on a browser, and will only require us to use our computers, with an Internet connection. For backup purposes (in case the Internet connection is lost or there is high latency), we will also prepare a video walkthrough of THOR and how it is expected to be used, but the video will not be publicly released. Unfortunately, due to the sensitivity of the data that is currently being ingested in THOR, a publicly viewable screencast or video cannot be released at this time.

3 RELATED WORK

The design and architecture of THOR draws on many areas in Web-centric data management and Artificial Intelligence, including real-time processing of data streams [28], online natural language processing [16], especially named entity recognition [22], machine learning problems such as topic modeling and sentiment analysis [7], [23], knowledge graphs [29], [32], network analysis [3], entity resolution and linking [11], [6], [33], [20], provenance reasoning [18], route planning and graphical user interfaces [10]. In the literature, there have been some examples of systems that act as

‘social sensors’ for the purpose of detecting disasters in real-time from social media streams, influential examples (for earthquake detection from social media) being [25], [5] and [9]. In principle, THOR resembles such systems; however, THOR includes innovations and capabilities not encompassed by any of those systems. For example, THOR has to take NLP outputs (especially machine translation), themselves noisy, as inputs, and ultimately outputs interactive user-driven elements. Also, THOR is designed to handle *emergent* disasters and incidents that may not have been seen before, unlike a particular disaster (e.g. earthquake). *Real-time scalability* is important in THOR.

Visualization is an important part of any human-centric system that is attempting to make sense of a large amount of information. Several good crisis informatics platforms that provide visualizations include Ushahidi [21], Twitris [17], Twitcident [1], AIDR [13], CrisisTracker [24], and several others [8], [30].

4 CONCLUSION

The Text-enabled Humanitarian Operations in Real-time (THOR) framework is a multi-organization collaboration effort that is being implemented to provide humanitarian operators superior situation awareness for emergent incidents in low-resource environments. THOR involves many different components, and serves as the gateway between a large corpus of streaming documents, usually available over Web channels, natural language processing modules, and a broad category of users (military, civilian and intelligence personnel). THOR is under active development, and presents a variety of challenges that require interdisciplinary research at the frontiers of Web-centric crisis informatics. It also serves as an example of using Web and AI technology for social good.

Acknowledgements. The authors gratefully acknowledge the ongoing support and funding of the DARPA LORELEI program, and the aid of our partner collaborators and users in providing detailed analysis. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

REFERENCES

- [1] Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. 2012. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 305–308.
- [2] Yusup Aibaidulla and Kim-Teng Lua. 2003. The development of tagged Uyghur corpus. *Proceedings of PACLIC17* (2003), 1–3.
- [3] Brian Anderson and Sumeth Vongpanitlerd. 2006. *Network analysis and synthesis*. Dover.
- [4] Rudy Arthur, Chris A Boulton, Humphrey Shotton, and Hywel TP Williams. 2017. Social Sensing of Floods in the UK. *arXiv preprint arXiv:1711.04695* (2017).
- [5] Marco Avvenuti, Stefano Cresci, Marianonietta N La Polla, Andrea Marchetti, and Maurizio Tesconi. 2014. Earthquake emergency management by social sensing. In *Pervasive computing and communications workshops (PERCOM workshops), 2014 IEEE international conference on*. IEEE, 587–592.
- [6] Cosmin Adrian Bejan and Sanda Harabagiu. 2014. Unsupervised event coreference resolution. *Computational Linguistics* 40, 2 (2014), 311–347.
- [7] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [8] Seonhwa Choi and Byunggul Bae. 2015. The real-time monitoring system of social big data for disaster management. In *Computer Science and its Applications*. Springer, 809–815.
- [9] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. 2013. # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17, 1 (2013), 124–147.
- [10] Daniel Delling, Peter Sanders, Dominik Schultes, and Dorothea Wagner. 2009. Engineering route planning algorithms. In *Algorithmics of large and complex networks*. Springer, 117–139.
- [11] Lise Getoor and Ashwin Machanavajjhala. 2013. Entity resolution for big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1527–1527.
- [12] Xingsheng He, Di Lu, Drew Margolin, Mengdi Wang, Salma El Idrissi, and Yu-Ru Lin. 2017. The signals and noise: Actionable information in improvised social media channels during a disaster. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 33–42.
- [13] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 159–162.
- [14] Muhammad Imran, Shady Elbassouni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Extracting information nuggets from disaster-Related messages in social media.. In *ISCRAM*.
- [15] Ann Irvine and Alexandre Klementiev. 2010. Using Mechanical Turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Association for Computational Linguistics, 108–113.
- [16] Peter Jackson and Isabelle Moulinier. 2007. *Natural language processing for online applications: Text retrieval, extraction and categorization*. Vol. 5. John Benjamins Publishing.
- [17] Ashutosh Sopan Jadhav, Hemant Purohit, Pavan Kapanipathi, Pramod Anantharam, Ajith H Ranabahu, Vinh Nguyen, Pablo N Mendes, Alan Gary Smith, Michael Cooney, and Amit P Sheth. 2010. Twitris 2.0: Semantically empowered system for understanding perceptions from social data. (2010).
- [18] Grigoris Karvounarakis, Zachary G Ives, and Val Tannen. 2010. Querying data provenance. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 951–962.
- [19] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, 177–180.
- [20] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics* 2 (2014), 231–244.
- [21] Nathan Morrow, Nancy Mock, Adam Papendieck, and Nicholas Kocmich. 2011. Independent evaluation of the Ushahidi Haiti project. *Development Information Systems International* 8 (2011), 2011.
- [22] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30, 1 (2007), 3–26.
- [23] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.
- [24] Jakob Rogstadius, Maja Vukovic, CA Teixeira, Vasilis Kostakos, Evangelos Karapanos, and Jim Alain Laredo. 2013. CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development* 57, 5 (2013), 4–1.
- [25] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.
- [26] Axel Schulz, Petar Ristoski, and Heiko Paulheim. 2013. I see a car crash: Real-time detection of small scale incidents in microblogs. In *Extended Semantic Web Conference*. Springer, 22–33.
- [27] Tomer Simon, Avishay Goldberg, and Bruria Adini. 2015. Socializing in emergencies—A review of the use of social media in emergency situations. *International Journal of Information Management* 35, 5 (2015), 609–619.
- [28] Michael Stonebraker, Uğur Çetintemel, and Stan Zdonik. 2005. The 8 requirements of real-time stream processing. *ACM SIGMOD Record* 34, 4 (2005), 42–47.
- [29] Pedro Szekely, Craig A Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, et al. 2015. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference*. Springer, 205–221.
- [30] Dennis Thom, Robert Krüger, Thomas Ertl, Ulrike Bechstedt, Axel Platz, Julia Zisgen, and Bernd Volland. 2015. Can twitter really save your life? A case study of visual social media analytics for situation awareness. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific*. IEEE, 183–190.
- [31] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1079–1088.
- [32] Quan Wang, Bin Wang, and Li Guo. 2015. Knowledge Base Completion Using Embeddings and Rules.. In *IJCAI* 1859–1866.
- [33] Linhong Zhu, Majid Ghasemi-Gol, Pedro Szekely, Aram Galstyan, and Craig A Knoblock. 2016. Unsupervised Entity Resolution on Multi-type Graphs. In *International Semantic Web Conference*. Springer, 649–667.