Data to Knowledge: Modernizing Political Event Data for Big Data Social Science Research

Latifur Khan, Patrick T. Brandt, Jennifer Holmes, Vincent Ng, Javier Osorio
The University of Texas at Dallas, University of Arizona

eventdata.utdallas.edu

Motivation:

• Provide accurate structured dataset on global political and social events, with historical coverage, geographic-location, drawn from multiple languages, and freely available in near-real time.
• Modernize event coding by moving beyond the specialized skills required for generating, visualizing, and analyzing event data.
• Build a platform to code across multiple languages, topics, and issue areas.

Key Problems

- Unavailability of real-time updated events data in structured format
- Lack of system handling large number of news articles for event coding
- Missing of dynamic ontology extansion in Rule-Based event coding system
- Identification and disambiguation of location mentioned in text documents and related to particular events.
- Lack of real-time systems that can provide query based datasets for visualization and further analysis

Scientific Impact:

Current system works as Information Retrieval tool for political science domain. With some modification the system can capture

- Events in different domain with the appropriate rules and ontology.
- Multilingual extension to capture events from different languages
- Distributed processing of large amount of text
- Comapartive study on Human coding vs Machine Coding
- Application of text minning techniques for knowledge base extension

Solutions:

- Event coding software PETRARCH and it's Multilingual Extension UD-PETRARCH works on capturing events in *who-did-what-to-whom* format
- Real-time political event coding framework based on Apache Spark Streaming for processing metadata extraction step using Stanford CoreNLP/UDPipe
- RePAIR system for actor recommendation based on news media presence and minning of verp patterns from text using different text minning techniques involving word embeddings, Association Rule Mining, etc. (ongoing)
- PRoFILE, a multilingual geo-location tool that can identify focus location based on text, uses standard text mining approaches along with sophisticated ones like Kernel Mean Matching techniques to provide multilingual location extraction where labeled data is not available or inadequate for learning purposes
- Event data API for serving data in realtime for analytical and visualization purposes (i.e. TwoRaven tool).
- Verb Translation App (VTA) to translate ontology rules defined in English to other languages including Spanish. (ongoing)

Broader Impact on Society:

- Creates a workforce that is able to work in both science, engineering, national security, and intelligence
- Continue our previous NSF RIDIR funded outreach where over 30% of all direct project participants were women and 20% were minorities.

Broader Impact on Education and outreach

- Scaling the related software and data infrastructure aids the political science, national security and big data research communities.
- Courses form CS/Political Science dept. leverages the content of the project for designing class projects.

Broader Impact (quantify potential impact)

- Text based focus location detection is adaptive for other domains
- Novel Techniques (i.e. RePAIR, VTA) for ontology translation/extension and related tools will be helpfull for foreign language/other domains.