RAPID ETHNOGRAPHIC ASSESSMENT: DATA-TO-MODEL

Kathleen M. Carley[1], Michael W. Bigrigg, David Garlan, Jeff Johnson*, Frank Kunkel, Michael Lanham, Michael Martin, Geoff Morgon, Bradley Schmerl, Tracy van Holt*
Carnegie Mellon University
*East Carolina University

ABSTRACT
When faced with understanding what is going on in a contested region or a military experiment it is often necessary to rapidly process vast quantities of textual data, assess the underlying situation, and identify changes early on. A human-in-the-loop approach for rapidly extracting the social and organizational networks, key activities, issues and sentiment is described. Using a combination of pre-defined ontological categorization schemes, thesauri, machine learning and topic-modeling information on people, groups/organizations, issues, activities, and locations are extracted as are relations among them. The value of this data-to-model methodology, its re-usability across contexts, and key challenges are demonstrated using data from three diverse sources: a scenario driven deterrence assessment, a military multi-actor experiment, and open-source information on the Sudan. We find that segmentation of people, groups, and locations into generic and specific, syntax hierarchies for identification, and re-usable thesauri enable rapid and meaningful meta-network extraction. Sentiment and events can often be post-inferred. Forecasted change in these networks and the dispersion of sentiment can then be simulated. The data-to-model system enables improved scalability, early results are 5x greater, and decreased processing time, early results are over 500% faster than without process, with fidelity similar to that by ethnographers.

PRIMARY TRACK
Socio-Cultural Data

SECONDARY TRACK
Analytic Methods Science and Technology (S&T)

DESCRIPTION
The soldier encountering a new area of operations needs to be able to rapidly assess and understand key aspects of the underlying culture, identify key actors, hot spots of activity, and assess who can be relied on for what type of resources of activities. Ethnographers, field workers, and anthropologists often collect relevant data through painstaking and lengthy interviews and observations. The goal of this project is to provide the soldier with a rapid assessment of the socio-cultural situation using open-source data by providing a simple data-to-model procedure that moves directly from the identification of relevant texts to simulation models for forecasting. We describe this multi-modeling approach which can be supported with a SORASCS backend [6], or directly through a script-runner on local machines.

[1] Dr. Carley, due to prior engagement, is only available on the 8th. If the presentation is on a different day another member of the team will present.

The data-to-model process is focused on a meta-network representation of the data [4][5][7]. The data-to-model process involves, at a high level, these steps: 1) collect text-based data; 2) clean the text corpus; 3) extract meta-networks from texts; 4) define and revise context specific thesauri; 5) extract final meta-networks from texts; 6) post-process to add geo-spatial information; 7) analyze meta-networks to identify key actors, resources, locations etc. and change in these; 8) identify potential interventions; 9) run simulations to forecast change and the impact of these interventions.

At this point we have tested the data-to-model process using corpi from the following cases:

- Indo_Pak: A scenario driven deterrence assessment: Uses a corpus of 27,000 text files from news sources and government websites. The project-based thesauri added only an additional 962 entries.
- Singapore: A military multi-actor experiment. Uses a corpus of 3,100 text files from news sources, web sites, and communication logs. The project-based thesauri added only an addition 500 entries.
- Sudan: Open-source information on the Sudan. Uses a corpus of 71,000 text files from news sources, web sites, books, as well as additional information from a wide variety of collected information by scholar experts. The project-based thesauri includes 38,552 location listings extracted from a gazateer leaving 16,001 unique entries.

Step 1: The data for these cases was collected via web-scrapers and/or was provided by external sources.

Step 2: For each of these cases, the corpi was cleaned. Initial cleaning of the texts involves reformatting as well as a generic cleaning. The generic activities include preprocessing used to correct the text. Examples include: meta-data removal, deduplication, typo correction, the expansion of contractions and abbreviations. Pronoun resolution should be done and unidentified pronouns removed. Identification of compound concepts is done by applying a list of concept-changing n-grams. While typically the use of an n-gram is to identify words that are most commonly used together, in this context an n-gram is a multi-word concept whose definition changes when the concepts are reviewed individually versus as a single compound entity. Examples of concept-changing n-grams are "first aid" and "black market". Sub-tools for this processing exist and are correctly ordered in a reusable workflow.

Step 3: The meta-network ontology used includes agents, organizations, locations, events, knowledge, resources, beliefs and tasks. Agents, organizations and locations are further segmented into specific and generic. The specific concepts, a.k.a. named entities, identify instances of items, such as George W. Bush for agent, UNICEF for organization, and Pittsburgh for location. Some specific entities can be pre-established from existing lists such as a list of all countries or major cities, or a list of world leaders. A base thesauri is formed from the pre-existing ontologically categorized concepts and augmented with project-based material. The current pre-existing generics material consists of 22,455 entries. There are N entries in the "noise" list for concepts to be deleted when cleaning the text. The current pre-existing base material (including general and specific entities) consist of 150,749 entries including general entities 784 agents, 321 organizations, 708 resources, 1906 tasks, 1064 knowledge, 1274 locations, and 63 events. In addition, CRF and other advanced text-mining tools are used to extract these. Generic concepts include soldier (agent), tank (resource), and base (location). Many of the generic concepts in the ontology can be pre-established. All of this processing is available within AutoMap [2] and can be run from SORASCS [6].

Step 4: Some minor adaptation needs to be done based on the context as "front" is different for a military context and a weather forecasting context. Context specific information is used to refine the process through the creation of specialized thesauri and delete lists using a computationally supported human-in-the-loop approach. Context specific entities can be found by reviewing all proper nouns identified in the corpus by applying part-of-speech analysis. For convenience, proper nouns adjacent to one another can be listed in an n-gram form as many project-based specifics are compound concepts. Alternatively, the approach to finding specific compound concepts would involve the generation of all n-gram possibilities, which is prohibitively large for human review beyond bigrams (n-grams with N=2). The list of possible concepts of interest can be culled by removing all concepts already placed in an ontological category. Using the pre-established material significantly reduces the amount of human involvement by at least 500%. The use of pre-existing thesauri reduced significantly the amount of work needed to extract a meaningful meta-network. For example, in a previous project the data-to-model process without this new strategy for a corpus of 1109 text files previously took over 1000 man hours. The scenario driven deterrence assessment described above took 160 man hours, an order of magnitude improvement even with a larger corpus.

Step 5: Repeats the processing step3 using the new thesauri and delete lists of step 4. This sequence is repeated as needed until the analyst is satisfied. Typically 3-5 processing rounds are needed.

Step 6: Using post-processing geo-information is added. This typically involves adding the latitude and longitude to facility geo-based analysis and visualization. For this purpose we used GeoNames (http://www.geonames.org), which is an open source gazetteer site. An alternative site that has been used in the past is NGIA GeoNET(http://earth-info.nga.mil/gns/html/) . Each of these sites contains geospatial information, along with other important information, such as population, type of location (city, village, landmark, etc). Any of this information can be used and added as attributed through a human-in-the-loop process.

Step 7: The resulting files, which are in DyNetML can then be processed using network analysis routines. Table 1 shows the size of the processed data and illustrative results. We used ORA for our processing [1].

Table 1. Processed Data

Factor	Indo_Pak	Singapore	Sudan
Agents -Total	119	98	996
Specific	119	98	585
Generic	0	0	411
Locations - Total	3306	175	2671
Specific	900	175	2627
Generic	2406	0	44
Organizations - Total	326	142	487
Specific	326	142	326
Generic	0	0	161
Task	428	446	1231
Resources	114	202	310
Event	22	4	52
Knowledge	140	99	991
Beliefs	42	1	1
Concepts	4497	1167	6739

Step 8: A type of intervention is to isolate key actors. Another is to send a message via some media. We automatically run the ORA key-entity report to identify key actors and resources and the hot-topic report to identify potential message topics. This defines a set of possible interventions by characterizing whom we might want to isolate or what topics are "hot" around which beliefs may be coalescing. Illustrative key actors and beliefs are: Indo Pak: Mahmud Ahmad, nuclear action needed; Singapore: engineer, commander's intent met; Sudan: Minnawhi, separation of S. Sudan.

Step 9: The human analyst selects the interventions of interest and then using the experimental design system builds a virtual experiment to examine the impacts. We used Construct [3] for this process. Simulations suggest strengthening of beliefs.

Summary: Several challenges exist, whose resolution will further speed and enhance this process. These include, but are not limited to, need to improve geo-information extraction, need to handle embedded concepts that are not English, automated re-merging of email/chat content and header data, improved automated removal of meta-data from texts, utilization of auto-topic identification to supplement and direct user defined topic identification, improved post-processing for event and sentiment extractions, semi-automated construction of imports for other simulators, and improved support for constructing virtual experiments.

The cleaning of the text, extraction of proper nouns and the subsequent removal of pre-existing items for human review, and the generation of the meta-network using pre-existing and project-based thesauri, are all examples of workflows. The workflows are a sequence of common steps used to perform a task, often using different inputs or different thesauri. The workflows automate the task management allowing the analyst to focus on the domain and not on keeping track of the individual steps to be taken. By sharing workflows, a consistent approach for data-to-model is established. When advances are made, the workflow can be easily adapted to the new capability and the data-to-model processing re-run. This data-to-model process can reduce analysis time from years to less than a month. In some cases we have operated the results in less than two weeks when parallel processing was used.

ACKNOWLEDGEMENTS

This work was supported in part by the Office of Naval Research - ONR -N000140811223 (SORASCS), N000140811186 (Ethnographic), MURI with GMU Cultural Modeling of the Adversary, 600322 and W15P7T-09-C-8324 awarded by CERDEC-C2D under the THINK ATO. Additional support was provided by the center for Computational Analysis of Social and Organizational Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, or the U.S. government.

BIOGRAPHY

Kathleen M. Carley is a Professor of Computation, Organizations and Society and the center director for Computational Analysis of Social and Organizational Systems (CASOS) in the Institute for Software Research, School of Computer Science, at Carnegie Mellon University. Her research combines cognitive science, social networks, agent-based modeling and computer science to address complex social and organizational problems. Specific research areas include dynamic network analysis, computational social and organization theory, adaptation and evolution, text mining, information diffusion and belief dispersion, disease contagion, command control, and disaster response. She and members of her center have developed and deployed

multiple tools. She founded the journal Computational and Mathematical Organization Theory, co-edited several books and has over 250 publications in the computational organizations and dynamic network area. In 2001 she received a lifetime achievement award in sociology and computers and is the 2010 winner of the Social Network Simmel award.

REFERENCES

[1] Carley, K.M., Reminga J., Storrick J., and Columbus D., 2010, "ORA User's Guide 2010," Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR-10-120.

[2] Carley, K.M., Columbus D., Bigrigg M. and Kunkel F., 2010 "AutoMap User's Guide 2010," Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR-10-121.

[3] Carley, K.M., Martin M.K. and Hirshman B., 2009, "The Etiology of Social Change," Topics in Cognitive Science, 1(4):621-650.

[4] Carley, K.M. and Hill, V., 2001, Structural change and learning within organizations. In A. Lomi & E.R. Larsen (Eds.), Dynamics of organizations: Computational modeling and organizational theories (pp. 63-92). Live Oak, CA: MIT Press/AAAI Press.

[5] Carley, K.M., 2002, Smart agents and organizations of the future. In L. Lievrouw & S. Livingstone (Eds.), The handbook of new media (pp. 206-220). Thousand Oaks, CA: Sage.

[6] Garlan, D., Carley, K.M., Schmerl, B., Bigrigg, M., and Celiku, O. Using Service-Oriented Architectures for Socio-Cultural Analysis. In *Proceedings of the 21st International Conference on Software Engineering and Knowledge Engineering (SEKE2009)*, Boston, USA, 2009.

[7] Krackhardt, D. and Carley, K.M., 1998, A PCANS model of structure in organization. In Proceedings of the 1998 International Symposium on Command and Control Research and Technology Evidence Based Research, (pp. 113-119). Vienna, VA.