OVERCOMING INFORMATION CHALLENGE IN FEDERATED ANALYSIS: FROM CONCEPTS TO PRACTICE

WORKSHOP REPORT

27-28 August 2008

SPONSORED BY:

United States Department of Defense

Office of the Secretary of Defense (OSD), Director of Defense Research and Engineering (DDR&E), Rapid Reaction Technology Office (RRTO)
Joint Staff (J3)
US Strategic Command (STRATCOM), Global Innovation and Strategy Center (GISC)
US Special Operations Command (SOCOM)
Institute for Defense Analyses (IDA)

PREPARED BY:

National Security Innovations (NSI), Inc.
Sarah Canna
301.466.2265
scanna@natlsec.com

OVERCOMING INFORMATION CHALLENGE IN FEDERATED ANALYSIS

EXECUTIVE SUMMARY

The *Overcoming Information Challenge in Federated Analysis: from Concepts to Practice* workshop, held on 27-28 August 2008, was convened in response to requests from US Special Operations Command and US Strategic Command, Lieutenant General Carter F. Ham, Director of Operations, J-3, Joint Staff. The workshop is part of an effort to develop national-level capability to identify and anticipate the action of Violent Non-State Actors (VNSA) with regard to their intent and capability to acquire, build, store, employ, deploy, or supply weapons of mass destruction (WMD). In support of this effort, the workshop was sponsored by a multiagency team and hosted at the Directed Technologies Incorporated (DTI) facility in Arlington, Virginia.

The purpose of the workshop was to identify analytical methods and tools that could (1) aid in the sorting of large volumes of unstructured, multilingual data and to (2) help make sense of the output. Concerted novel approaches to data trolling, ingest, characterization, processing, extraction, coding, and visualization may lead to great potential for proactive rather than reactive capabilities, not only within the military context, but within all components of national power.

Focusing on the front-end issues of the data analysis process will significantly improve our nation's and its allies' capabilities to proactively uncover global WMD proliferation. The workshop brought together a diverse group of physical and social scientists, academics, analysts, intelligence collection managers, operational planners, and decision makers to demonstrate and discuss usable techniques and methods for a more effective nexus of the federated informational and analytical processes.

This workshop examined technologies that could play a role in addressing the problems faced by the Strategic Multi-layer Analysis (SMA) team in marshalling data and putting it in a form that could be readily analyzed to perform quantitative analysis. The purpose of the workshop was to expose problems and offer potential solutions to the problem of ingesting large quantities of data, particularly data from multilingual text and coding it for use in analytical models. The discussion of the Weapons of Mass Destruction – Terrorism (WMD-T) Joint Intelligence Preparation of the Operational Environment (JIPOE) problem space and technical approach are included to give the reader an understanding of why the technologies being explored will be needed in the long term by any group seeking to look globally at non-state actors who may choose to use weapons of mass destruction.

JIPOE is a proven doctrinal process typically applied to a bounded geographic space.[1] While it has rarely been applied to the government as a whole, the process can be used to assess the information challenge problem and the federated analysis solution. While the JIPOE process typically focuses on the commander, the process applies to any federal decision maker. The combined forces of the nature of globalization and increasing interdependency generated by the informational revolution have ushered in a new imperative for a more holistic understanding of the operation environment. This workshop addressed these issues and looked for solutions to automated multi-data lingual processing in a federated environment.

The objective of the WMD-T JIPOE effort is to develop a sustained, national-level means for combating WMD-Terrorism with a forecasting, anticipating, or inferential strategic environmental level assessment capability covering the full threat spectrum from intent to act, to acquisition and attack preparation, and all

[1] This section is taken in part from WMD-T JIPOE Interim Report.

the way downstream to deployment by non-state actors (i.e., from far left to just left of "boom"). The ability of the government to process data, identify indicators of terrorism, and share those findings in a truly federate manner needs to be enhanced. As a consequence, the only way to fight a dynamic and adaptive threat network is a similarly dynamic and adaptive USG federated forecasting and anticipating assessment network.

The notion of a sustained and repeatable capability requires that the adopted methods and processes be founded on robust scientific methods and theory. The nature of global reach forces the intelligence and defense communities to continue to enhance their efforts to exploit open source information and fuse it, as appropriate, with all the source information currently collected. The volume of that information demands that the effort use reliable automation technology to the fullest extent, recognizing that multilingual data mining is not mature for all languages. Even with the best data mining techniques, new methods for sorting the data and targeting the factors that will most reliably detect intent to acquire and use WMD must be established to avoid increasing the workload of already over-burdened analysts. Thus from the front end extraction processes to the back end analysis methods, existing tools must be augmented by emerging capabilities that keep analytical decisions properly centered on the human analysts while enhancing their performance.

Pieces of the WMD-T JIPOE capability already exist in various places among a number of groups within and outside the government. Thus the WMD-T JIPOE will work to take full advantage of these unique and disparate efforts in such a way that each participating group is capable of sustaining its own efforts while collaborating with and benefiting from the work of the other participants. The Information Challenge workshop highlighted pieces of the puzzle to join the research and the operation communities together to work toward a unified effort. The ultimate vision of the WMD-T JIPOE enterprise is a federation in which the WMD-T JIPOE cell is the core and integrator of data and where methods and tradecraft are shared among collaborating groups.

The workshop was designed to showcase tools and technologies that could be brought to bear on the information challenge problem. The problems and potential solutions explored in this workshop are common to challenges that rely on the ability to ingest large amounts of multilingual text information and extract from it data that can be used in analytical models.

TABLE OF CONTENTS

Executive Summary 2

Table of Contents 4

Abbreviations 6

Addressing Multilingual Information Overload Problems 7

Introduction to the Flow of the Workshop 7

Addressing Multilingual Information Challenge Overload Problems 8

Orientation/ Overview to WMD-T JIPOE 8

Understanding the Problem: WMD-T JIPOE Information Ingest, Structuring and Characterization 8

Coding with Unstructured Text and Operationalizing the JIPOE Nexus 10

U.S. Army Intelligence & Security Command 10

Session 1: Breaking it Down: Information Extraction and Coding 11

Understanding the Problem – Lessons from JIPOE Nuclear Smuggling Case 11

Rare Event White Paper 14

Art and Science of Data Coding 15

Extracting Context from Unstructured Text 17

Panel Discussion 1: Challenges of Multilingual Processing – Scaling and Tailoring for Global Missions 18

Bootstrapping Translation Capabilities 18

Translation for High Demand Languages 20

ARGUS 22

Session 2: "Yours, Mine and Ours": Operationalizing the JIPOE Nexus 23

Understanding the Problem – Lessons Learned 23

The Nexus Federated Collaboration Environment (NFCE) 23

Industry Perspectives 24

Art of the Possible: Functional DoD Federation 26

Biologically Inspired Innovation in Action 27

Technology Previews 28

Charlotte Visualization Center, University of North Carolina, Charlotte 28

Artificial Intelligence Lab, University of Arizona 29

BBN Technologies 29

Ontology Works 30

SAS / DETICA 30

Social Science Automation 31

NuTech Solutions Division of Netezza 31

SAIC 32

Open Source Center 32

SAIC 33

Sandia National Laboratories 33
Air force research laboratory 34
Strategic Analysis Enterprises (SAE) 34
Pacific Northwest National Laboratory (PNNL) 34
Indiana University 35
IBM 35
Sandia National Laboratories 36
Wrap Up 36
Appendix A: Workshop Agenda 37
Appendix B: Workshop Attendees 38

ABBREVIATIONS

ABA	Automated Behavior Analysis
ABEM	Agent-Base Evidence Marshaling
AFRL	Air Force Research Laboratory
AOR	Area of Responsibility
CENTCOM	United States Central Command
COCOM	Combatant Command
DARPA	Defense Advanced Research Projects Agency
DDR&E	Director of Defense Research and Engineering
DHS	Department of Homeland Security
DIA	Defense Intelligence Agency
DOD	Department of Defense
INSCOM	United States Army Intelligence and Security Command
JIPOE	Joint Intelligence Preparation of the Operational Environment
JOUST	Joint Operations on Urban Synthetic Terrain
NASIC	National Air and Space Intelligence Community
ODNI	Office of the Director of National Intelligence
OSD	Office of the Secretary of Defense
MAR	Minorities at Risk
NFCE	Nexus Federated Collaboration Environment
PNNL	Pacific Northwest National Laboratory
SINAPS	Small World Interagency Process System
SMA	Strategic Multilayer Analysis
SME	Subject Matter Expert
SNL	Sandia National Laboratories
START	National Consortium for the Study of Terrorism and Responses to Terrorism
STRATCOM	United States Strategic Command
TSWG	Technical Support Working Group
USDI	Under Secretary of Defense Intelligence
USDP	Under Secretary of Defense Policy
USG	United States Government
VNSA	Violent Non-State Actors
WMD	Weapons of Mass Destruction
WMD-T	Weapons of Mass Destruction - Terrorism

ADDRESSING MULTILINGUAL INFORMATION OVERLOAD PROBLEMS

INTRODUCTION TO THE FLOW OF THE WORKSHOP

Industry and government representatives spoke about challenges in information processing and multilingual data over the course of the two-day workshop. These speakers included:

- Facilitator: Laurie Fenstermacher, Air Force Research Laboratory (AFRL)
- Hriar Cabayan, Office of the Director for Defense Research and Engineering (DDR&E)
- Susan Numrich, Institute for Defense Analyses (IDA)
- Russell Richardson, U.S. Army Intelligence and Security Command (INSCOM)
- Robert Popp, National Security Innovation (NSI), Inc.
- Amy Pate, National Consortium for the Study of Terrorism and Responses to Terrorism (START)
- Rohini Srihari, Janya Inc.
- Ray Slyh, Air Force Research Laboratory (AFRL)
- Kathleen Egan, Technical Support Working Group (TSWG)
- Frank Connors, Defense Threat Reduction Agency (DTRA)
- Carl Hunt, Institute for Defense Analyses (IDA)
- Stanley Horky, General Motors (GM)
- Amy Henninger, Institute for Defense Analyses (IDA)

The workshop also scheduled time for technology demonstrations and displays. The technologies were selected to represent a sample of the work that is being done in the information field and were not meant to be comprehensive. The technology demonstrations included:

- University of North Carolina, Remco Chang
- University of Arizona, Hsinchun Chen
- BBN Technologies, Sean Colbath
- Ontology Works, Ryan Cole
- Detica / SAS, David Porter
- Social Science Automation, Robalyn Stone
- NuTech, Mario Inchiosa
- SAIC (Thememate), Gary Jackson
- Open Source Center, Nancy Chichor and Roberta Dobbins
- SAIC (LPA), Mark Clark, Anne Russell
- Sandia National Laboratories, Christy Warrender
- Air Force Research Laboratory, Mark Zappavigna
- Strategic Analysis Enterprises (SAE), Stephen Shellman
- Pacific Northwest National Laboratory, Antonio Sanfilippo
- Indiana University, Travis Ross
- IBM, Salim Roukos
- Sandia National Laboratories, Peter Chew

A list of workshop attendees is located in Appendix B.

ADDRESSING MULTILINGUAL INFORMATION CHALLENGE OVERLOAD PROBLEMS

Laurie Fenstermacher, AFRL

There is a pervasive need to deal with ingesting, characterizing, and extracting information from data, particularly unstructured and multilingual data. Many technologies that address this need are becoming mature enough to be useful to the intelligence and analysis communities. The availability of mature technologies to address the multilingual information overload problem has resulted in renewed interest and use of unclassified, open source information for missions across the globe.

Day one of the workshop focused on the technologies available to address the multilingual information challenge. The morning session of day one set the context for JIPOE. The afternoon was set aside for brief presentations and poster sessions for various technologies representative of mature, multilingual, data processing technologies.

ORIENTATION/ OVERVIEW TO WMD-T JIPOE

Dr. Hriar Cabayan, SMA Director, DDR&E

This workshop was designed to display the state of the art technology in multilingual data processing. Dealing with the overwhelming amount of information available from open sources has affected many agencies across the federal government, notably those that have sponsored this workshop. Priorities for the Strategic Multilayer Assessment (SMA) are set by the Joint Staff (J3) and the effort is executed through U.S. Strategic Command (STRATCOM), Global Innovation and Strategy Center (GISC), and the Director of Defense Research and Engineering (DDR&E). Its objective is to provide potential solutions for the combatant commands' operational requirements. The SMA process typically discovers and reports solutions within a six-to-eight month time frame. SMA provides the advice; the commands operationalize the solution.

The SMA program first realized the difficulties surrounding the information challenge problem during its Sudan study sponsored by United States Central Command (CENTCOM). Over 90 percent of the information about Sudan was available from open source material. The modelers soon came to realize that open source information was a Pandora's box that continuously generated problems for framing the problem and modeling the solution. These problems and others were addressed during the workshop.

UNDERSTANDING THE PROBLEM: WMD-T JIPOE INFORMATION INGEST, STRUCTURING AND CHARACTERIZATION

Dr. Susan Numrich, IDA

Sue Numrich from the Institute for Defense Analysis (IDA) discussed challenges faced by the JIPOE during previous efforts and case studies. Sue emphasized that all of the workshop participants were invited because each could contribute a piece of the solution in the information challenge problem space. Whether one employs low technology solutions, such as a "bunch of guys and gals sitting around a table" (Bogagsat) or whether one employs visualization or analysis tools to bring together the information, there is a critical need for better, more accurate information. The type of information to be primarily discussed during this workshop is found in text form.

To illustrate the problem, Numrich used the damage from the Katrina hurricane in Alabama as an example. A person may know where Alabama is, the language, and perhaps even the culture, but the problem remains of not knowing where to find the best sources of information, extracting relevant pieces, and the best way to store and archive the information. In a more complicated example, Numrich posed the question of whether rebels in the nation state of Georgia want to be part of Russia or just an independent state. In this example, the analyst does not know the language, culture, or geography. Additionally, the analyst does not know where to find the best sources of information or whether the sources potentially contain bias. But even still, this is a bounded problem. A more complex, open ended question might be, why has the drug trade shifted from the majority coming to the United States to a majority going to Europe and the Middle East. The problem space changes because the analyst must go back many years and investigate many countries and many languages. Additionally, the drug trade is a multi-stage process involving manufacturing, sales, finance, etc. The analyst must not only detect the credibility and bias of potential sources but determine the audience in order to know whether they are meant to understand or not. The problem space facing the JIPOE is how to better understand culture in order to understand what is going on in a nation or region. This information must then be compiled in a way that is easily understood using a variety of techniques including visualization. The information collected must then be operationalized to help establish causality. Finally, the information must be categorized so that analysts can compare apples to apples.

One of the problems of particular concern facing JIPOE today is understanding the propensity of any non-state actor to use WMD. JIPOE seeks to understand how to take all of the pieces of the multilingual, culturally disperse, and multidisciplinary problem and process them in a way to make them fit together in an analytic jigsaw puzzle. Many efforts are underway to use technology to extract, code, store, manage, and analyze data, but they largely work separately. A stronger effort needs to be made to integrate these technologies and tools to meet today's requirements.

The workshop challenged the participants to work together to better understand the pieces of the puzzle that will need to be assembled to address the information challenge.

Discussion

Kelcy Allwein of the Defense Intelligence Agency (DIA) stated that one of the things she has seen in immersive visualization is that sometimes it is not about the data you have, sometimes it is about the data you are missing. As you put data together, you must understand the data you have including gaps and inconsistencies. Maybe the fact that information is unavailable tells you about gaps or confirms theories. This is an area that must be further explored.

Sue Numrich responded that denial and deception campaigns complicate the information challenge problem. A government in crisis wants to deny information to reduce panic. How does an analyst understand that there is something going on *sub rosa*? The problems are hard and they are deeply layered. This workshop helped to determine where to look for the next layer.

CODING WITH UNSTRUCTURED TEXT AND OPERATIONALIZING THE JIPOE NEXUS

Laurie Fenstermacher reviewed the information overload problem facing the USG. She introduced the first brief of the day – an operationally focused brief on what United States Army Intelligence and Security Command (INSCOM) is doing in the information arena. Session one focused on lessons learned from previous JIPOE case studies. Panel discussion one focused on transitioning and scaling multilingual capability.

U.S. ARMY INTELLIGENCE & SECURITY COMMAND

Futures Directorate Initiatives / Projects

Russell Richardson, INSCOM

The INSCOM Futures Directorate mission is multifold. It provides management, oversight and synchronization of INSCOM related research and development efforts. It identifies emerging concepts, technologies and tools that fill existing INSCOM intelligence capability gaps or that may substantially improve Army Intelligence Enterprise capabilities. It integrates and operationalizes high payoff technologies and concepts through the Rapid Technology Prototyping process. Finally, it manages the transition of capabilities to an appropriate acquisition program.

The INSCOM Futures Directorate has been working on data extraction and management for about nine years. The directorate focuses on finding, testing, and fielding promising technologies early. The tools include but are not limited to data extraction, entity extraction, and XML coding.

The INSCOM Futures Directorate's charter is to bring intelligence products into a single repository so analysts can query and find data. Having a large data warehouse helps analysts conduct a federated search of data. The Directorate has fielded copies of the warehouse to Combatant Commands (COCOMs) and others in the operational theaters. The objective of the warehouse is to become a single source of query for analysts. Technology is needed to extract data and to make it receptive to queries across various data sources. The warehouse contains over 100 sources of both real time intelligence feeds and data repositories. Currently, the warehouse does not contain audio sources, but it stored and indexed audio for eight years until it became impractical. That mission was passed off to another organization.

INSCOM Futures Directorate is funding about 25 technology development projects right now. The goal is to do an evaluation of the technologies in the software integration lab using pseudo analysts to evaluate the technologies, determine maturity level, and approve for use in an operational environment. The target transition system in the near term is the Distributed Common Ground Station (DCGS). DCGS is a large Joint project to bring together sensor data in one central enterprise.

The objective of the directorate is to mature appropriate technology to hand off to the field. INSCOM does not have a production responsibility.

INSCOM has software processes to tag and clean data. Data feeds are updated constantly. The extraction process is split in two pieces. Upon entering the system, the tool finds names, extracts entities, tags, geopositions, cleans, and normalizes the data. From there, analysts can create their own corpus from 30 million records to 30,000 records.

INSCOM does not engage multilingual tools anymore. It can hold various language documents and they have dictionaries where you can look up words, but it is not truly multilingual.

Discussion:

Antonio Sanfilippo asked why INSCOM did not use semantic indexing. Richardson replied that it is an interesting example of how everything worked correctly, but was overcome by a different set of events. INSCOM looked into it very early and helped mature the technology and it had a lot of promise, but when SAIC sold the technology, licensing became no longer practical for the Army.

Antonio Sanfilippo asked if the tools cluster in English. Richardson responded that it did not. It had a corpus of documents that was translated into various languages through analysts, which was done as a normal part of workflow production. Developers trained on those corpora and created a single indexing corpus to do searches against. There were 5,000-10,000 docs that INSCOM supported in 20 languages.

SESSION 1: BREAKING IT DOWN: INFORMATION EXTRACTION AND CODING

UNDERSTANDING THE PROBLEM – LESSONS FROM JIPOE NUCLEAR SMUGGLING CASE STUDY

Bob Popp, National Security Information

One of the pilot JIPOE programs was nuclear smuggling, which provided valuable lessons learned. The nuclear smuggling project started in November 2007. It sought to understand what violent non-state actors (VNSAs) were doing in terms of nuclear smuggling. Initially, analysts from the Department of Energy and the intelligence community said this was a data rich problem set. They thought there were many instances of nuclear smuggling in the historic record. However, it turned out that data was scarce and the modelers were faced with a rare event problem. Plans for conducting fingerprinting (i.e., statistical analysis) were based on the expectation of a data rich problem. The resulting small n study made fingerprinting difficult. There were many lessons learned.

One of the first lessons learned was that defining the dependent variable was a critical issue. The problem was made difficult because both the unclassified and classified communities developed different definitions of the dependent variable. On the unclassified side, modelers were missing data that was available on the classified side. Having a consistent definition is critical to success.

Similarly, the data collection process was complicated by the fact that each provider of data (typically a US agency) collected data in various ways resulting in different attribute data. It was difficult to compare two databases and understand whether similar looking data complied with the definition of the dependent variable. Additionally, analysts were sometimes unable to tell whether two databases were even recording the same event. Ultimately, JIPOE could not resolve the database issue, but by better defining the dependent variable, it could better define what qualified as an event.

Modelers ended up spending between 50-90 percent of their time on consistently coding data. The coding process involves finding data, extracting it from a database, and determining how it fits with data extracted from other databases. Data could be integral, ordinal, or nominal - meaning that the modelers had to normalize and assign numerical values to the data. Some of this work can be automated, but it often required hand coding. The modelers put a code book together, but did not have time to train coders due to short timelines. Modelers had four months to do approximately 12 months of work. Due to lack of time and training, it turned out that data was coded inconsistently.

Another mistake modelers made was the failure to carry along confidences and uncertainties. This is especially important for rare events. There are statistical methods to deal with missing or scarce data, but it still requires confidences and uncertainties.

The value of classified data remained unexplored during this project because the modelers did not receive all available classified data. The effort also lacked law enforcement data. While no effort was made to determine what kind of information was available from the law enforcement community, it seemed a likely area of oversight.

Because of the rare event problem encountered by the modelers, statistical analysis techniques were stressed making the results of the analysis unexpected.

The program also highlighted the importance of analysts and modelers spending time together in a continuous way – meeting once a month at a workshop is not sufficient. It would be ideal for analysts and modelers to work together daily. If possible, analysts and modelers should be co-located for a set period of time. This would create consistency between what is desirable, what is possible, and what is value added.

Numrich added that at the beginning of the effort, there were multiple understandings of how to even define data. Data to some is a huge stack of paper text documents. People who said they had a ton of data were referring to stacks of paper. The quantitative data one could pull from those documents was not the same quality as pre-existing quantitative data set. The disconnect occurred because different communities speak different languages. To overcome this barrier, it would have been helpful if representatives from the various communities (historians, psychologists, linguists, computational analysts, etc.) could have sat down together to agree on a common vocabulary. Without this, it was a challenge to federate different capabilities in to a unified capability.

Discussion:

Carl Hunt asked Richardson whether INSCOM had any new tool to help bring modelers and analysts together. Richardson responded that cooperative efforts go through phases. There have been certain periods where the relationship between the two groups have been strong and it they worked well together. Then there were phases where everyone was too busy to cooperate.

Hunt also mentioned that having analysts and modelers meet once a month does provide some utility and should not be disregarded. He noted that meetings kept the group and the effort fresh. He suggested that one possible finding of the effort was that a way needs to be found for technologists, analysts, and modelers to create synergies together.

Chen stated that a study he did for DHS using the Dark Web data also found that nuclear smuggling is a rare event, but it is something that terrorists are talking a lot about. There is far less content in the Dark Web about bioagents and very infrequent talk about nuclear weapons. Smuggling is more difficult because you need both capability and intent. Some groups can build nuclear weapons, but they lack the intent to do so. And some groups have the intent but lack the capability. Smuggling requires a different skill set than making nuclear weapons, but linking people who can make nuclear weapons together with smugglers is very difficult.

Popp said that when you talk to analysts, they say there is a lot of smuggling going on because people can make money. It is not necessarily about having ill will; it is about making money.

Sanfilippo said that the rare event problem is a big one. He attempted to create a nonproliferation model. Instead of looking at events themselves, they looked at precursors and looked at conflicts in the area. One could look at trading routes or the extent to which a state actor has supported non-proliferation treaties. By looking at several variables, one can obtain a lot of data about precursor activities.

Popp responded that on the other hand, in an effort with over 200 variables – some way left of boom – the data and variables run the risk of becoming very difficult to quantify because they are ephemeral. It is hard to use this kind of data to make conclusions.

Cabayan stated that all of the comments are good, but that modelers inexperienced in the domain have to respect analysts who make understanding an issue or area their lifetime work. Modelers cannot match that experience or insight. Analysts do not want models or tools to solve the problem for them – they want tools to help them with their analysis. The harder a tool is to use, the less likely they will be to use it. Many analysts are concerned about imminent crises, they are not convinced that they can start looking left of boom yet. It is important that modelers and analysts work together to show them how anticipation works and how it can help them in their work. Operators are comfortable with risk. They are willing to live with some false positives, but they do not want false negatives with large consequences. They are risk averse. Modelers must take these needs into consideration. Modelers do not know whether they can succeed in anticipating attacks, but the investment is miniscule in comparison to the cost of a real attack.

Sanfilippo reiterated that modelers, analysts, and coders must be able to work together.

Cabayan warned that analysts already work long hours. Modelers must be careful how they make the request because analysts are not sure exactly what they will get at the end of collaboration. Once you get the analysts on board, they will support the tool.

Popp agreed and pointed out that the initial value of the collaboration was not about the tools; it was about gaining respect for one another.

Cabayan stated that a great example of successful collaboration and a successful tool is Argus. Eight months ago, analysts from Argus were brought on board to help with the nuclear smuggling effort. Modelers worked closely with nuclear smuggling analysts. Through an iterative effort, analysts and modelers improved the tool. At the end, the analysts were comfortable with the tool.

David Porter stated that the primary reason to involve analysts in the development process is to prevent tautologies. It is as much to stop the developers from doing wrong things as it is to do the right things. The problem is that there will never be enough data. He asked what the community is doing to prevent the next generation from becoming nuclear smugglers or proliferators.

Fenstermacher spoke about a program in Jordan where they teach K-12 students to dispel the relationship between violence and Islam. There are also programs for military and other societal organizations to ward off extremism.

Chen stated that he was pessimistic. Violent groups are able to move much faster than the US can, especially in the virtual world. This means that violent groups are having a much greater influence on the next generation. There is not enough positive messaging on the internet to counter the extremist voice.

Fenstermacher spoke about a program in Brighton, U.K. where they have programs to work with Salafi clerics and youth to counter the influence of extremists.

Chen argued that the US must go into cyber space to counteract the extremist influence.

Larry Kuznar said that getting back to the data problem, the most fundamental challenge is the competing definitions of data. This is the point from which all other problems are derived.

Richardson stated that the problem with data is faced every day. The problem will never be entirely solved, but the community is moving towards an ontology based representation. The goal is to have useful information, not just hundreds of sources. We do not do this well yet. By moving towards better, semantic representations of data, it is helping to solve the problem.

Kuznar asked David Porter whether they are using ontologies to overcome these problems. Porter responded that you cannot underestimate good housekeeping and good data management. Creating appropriate metadata is more important than reshaping existing data sources.

Richardson said that metadata standards are improving. Through efforts such as those supported by INSCOM, metadata standards are reaching a new plateau. New technologies will be developed and we have to incorporate ontologies into those. That is a significant step.

Ann Speed said that it is not that moving toward ontologies is bad, but that there needs to be a recognition that it is only one method of representation. One thing we need to do is define variables clearly. It involves communication of perspectives. The more data is processed, the more it is occluded. There is no one right way to define things. Speed also warned analysts and modelers not to throw away documents once they have been coded.

Ned Snead stated that most of the lessons learned have been captured. However, another important lesson in processing classified and unclassified data is to understand the culture gap. No information came down from the classified side to the unclassified side. The problem is universal in that many do not understand how to move the information down.

Popp stated that the real analytic expertise is at the unclassified level, but these experts were not able to benefit from knowledge gained at the classified level.

Kathleen Egan stated that ontologies designed with the western mind looking at non western data can be perilous. By doing so, you insert bias and loose value. There are lots of data that do not extract easily including intents, motivations, and sentiments. These are very difficult things to put into ontologies. The models will not work based on poorly constructed ontologies.

Kuznar stated that there are new developments in ontologies that overcome many of these issues. They are much more flexible and new concepts can be introduced. The ontologies come directly out of the data and are related to other concepts. People should not think of ontologies as rigid structures.

RARE EVENT WHITE PAPER

In early October, JIPOE/SMA will issue a three part white paper focused on rare events. The first part deals with how the various disciplines deal with rare events. Perspectives from anthropologist, economists and others will be sought to understand how practitioners deal with rare events. Part two will assess the limitations in anticipating rare events from a various perspectives including cognitive and social psychology. Part three will assess how various disciplines work around limitations including resources. Contributions to the white paper will be five-to-eight pages long and will be written for the operational community within the DOD and DHS.

ART AND SCIENCE OF DATA CODING

Amy Pate, Research Director, National Consortium for the Study of Terrorism and Responses to Terrorism (START)

Minorities at Risk (MAR) is a university-based research project that monitors and analyzes the status and conflicts of politically-active communal groups in all countries with a current population of at least 500,000. The project is designed to provide information in a standardized format that aids comparative research and contributes to understanding conflicts involving relevant groups.

Coding is processing narrative and other information and distilling it according to a pre-determined framework. It is also taking qualitative information and transforming it to quantitative information.

The key to coding is clearly defining what concepts and relationships are of interest. The data must be suitable and mapped to those concepts and relationships. More data does not always mean better data. However, more does always mean more expensive, more time, and more people.

MAR has a well defined code book. The problem they experienced though was that there were 400 variables to code. To overcome this challenge, the program reduced the number of variables coded to about 70. MAR had to clearly define what concepts and relationships were of most interest to them or their clients. The data collected must be suitable to answer questions and be mapped closely to concepts and relationships.

In the terrorism example, MAR created a concept map to narrow down what they really wanted to measure and to identify related concepts. The concept map affects at what level you measure an event.

Key lessons learned from years of coding involve the following. First, abstract concepts must be broken down in order to be measurable. Second, clear definitions are crucial. Third, levels of a variable must be clearly distinguishable. Fourth, more granular measurement requires better source information, more time, and more money.

Lessons learned from teaching graduate students how to code includes the basic fact that training *is* important. Intercoder reliability is also a key determinant of coding accuracy.

In terms of sources, not all sources of information are reliable. Some sources may be reliable for some types of information, in some geographic areas, during some time periods. For example, TASS was a very good source for information on the Middle East, but it was terrible for the Soviet invasion of Afghanistan because China had an active interest result in biased reporting.

It is important to develop ways to assess source reliability and train coders (human or machine) on those assessment methods. Similarly, not all regions of the world are covered to equal degrees in open-source information. Even in highly saturated regions, not all actors are covered equally. Modelers need to know how gaps in coverage impact relationships found.

Pate warned about other coding dangers. First, sources are replete with vague and ambiguous information. The more granular the measurement, the more difficult it is to code ambiguous, vague, or contradictory information. Modelers need to carefully choose their level of granularity to make best use of their resources. Additionally, modelers need to develop rules on how to deal with poor information

Discussion:

Chen asked whether in a social science coding context, whether Pate uses a computer program that learns as data is being coded. Pate responded that they are working with a group of computer scientists at the University of Maryland to develop a system to train algorithms for automatic data extraction and coding.

Cabayan asked how Pate managed the interaction between academic social and political scientists with the operational community. The two communities may have different perspectives that make collaboration challenging. Pate responded that it has been a learning process to work with people in government because MAR is a purely academic organization. The more interdisciplinary you get, the more you need to sit down and bang heads to reach agreements.

Popp stated that methodologically, the granularity level of coding is determined by operational problem specification and what the data will support. It is where the two meet. Where you want to go is not necessarily where you can get.

Pate stated that one of biggest challenges the MAR program faces is a high turnover rate. You have to calibrate people to ensure consistency across the entire dataset. This makes the training process extremely important. Modelers have to spend a lot of time upfront to design training protocols.

Porter asked how much of MAR's coding is based on the documents available and how much on background of the coder. Pate responded that it depends on the variable. Many variables require multiple source documents before it can be included in the database. There is very little done based on the coders' own expertise.

Porter asked whether MAR project has done any machine learning – taking post coded text and finding out what terms or phrases led to that coding. Pate responded that it is on the agenda, but has not been done yet.

One person from the audience asked why MAR was so careful to use multiple sources to validate the data. He asked why not allow the coders to note all data and let statistics deal with the uncertainties. Pate responded that they just started working on that problem with the computer scientists. However, MAR's market is political scientist academics. They would not know what to do with the data if it contained inaccuracies.

Chen asked how MAR deals with deception. Pate responded that coders should not make decisions about which sources is credible. She said deception is less of a problem in English language press, which MAR uses exclusively. However, MAR is working on this issue.

Another person from the audience suggested that gaps may cause biases in the data and in the results of statistical analysis. This goes back to optimizing what information is available and what question one wants to answer.

Remco Chang stated that it seems that Pate is saying that the data collected is not always reliable. It almost seems like you need a confidence level with every variable. Pate responded that that would be great to have.

Kuznar stated that there is another model done by Human Relation Files at Yale, which is a major set of information and structured database on 1000 cultures from all over world. They use a combination of structured, disciplined methods and do swarming. He is not sure how the database is coded, but it might be interesting to find out.

Another person from the audience asked about the extensibility and scalability of MAR's coding approach. She asked how reusable the indicators are in different areas (i.e., non terrorism related databases). Pate responded that it is interesting, because the word terrorism only appears once in the code book because it is such a highly loaded term. Acts of terrorism are defined by attacks against civilians.

Robalyn Stone stated that she had been in a similar position. You develop a methodology but then you have new questions and short timelines. She asked how you maintain some kind of methodology when you have a limited timeframe. Pate responded that the more you limit and clearly define the question, the faster you can get though information sources and the more you can discard. It is also helpful to look around and see what other people have done that is similar.

Popp asked about the state of art of automated coding. A member of the audience responded that coding events is in a pretty good state. The variables that are difficult are the ones where you have to pull in multiple pieces of information or the languages used are diverse. Popp stated that developers might have to work with computational linguists. Translations make coding very difficult, so modelers may have to work with multilingual people.

Sanfilippo asked Pate to say a little more about inter coder reliability. Pate responded that the classic way to assess inter coder reliability is to measure the output of two separate people. She has never done that because of budgetary issues. Sometimes it is easier to get reliability with rare events by guessing all zeros.

EXTRACTING CONTEXT FROM UNSTRUCTURED TEXT

Rohini Srihari, Janya Inc.

Multilingual text extraction and mining has received a lot of attention in recent years. One problem faced by multilingual extraction tools is that the low accuracy of machine translation (especially for names) results in poor search outcome. However, querying tools can help improve the outcome of various types of inquiries. The tool developed by Janya Inc. for AFRL and NASIC can query multilingual documents resulting in ranked documents or text snippets.

The objectives of the program are multifold. NASIC is primarily interested in searching Chinese scientific documents. The tool must be able to deal with very high volume of data from multiple journals and open sources.

A typical query is "Which authors/organizations have collaborated on research on new technologies for jet fuel?" In the current approach, the tool uses machine translation to convert documents to English. It then searches the English corpus. But the problem encountered is that translation is very poor and names get mangled. The solution is native language entity tagging, deeper content extraction, and topic classification used in conjunction with machine translation.

Janya found that by doing entity extraction in the native language first and then doing machine translation it results in better outcomes.

The content extraction engine generates entity profiles and events from English documents.

The cross document merging system consolidates profiles from multiple source documents into a database It performs cross-document entity disambiguation. It then enables search and visualization across an entire corpus and enables domain and/or language porting across the full document collection. The problem encountered is name matching and being sure that the names refer to the same person.

Each customer has different requirements. Therefore, rather than coming up with pre-packaged formats, Janya has spent time creating semi-automated domain porting tools. The hardest part of the process is defining client requirements.

The Semantex engine is deployed in high content volume areas. The engine has been retooled to work in Chinese; Russian and German are the next languages in the pipeline.

Discussion:

Chen asked if this project was similar to the nanotechnology project in that the objective is to retrieve Chinese documents, but queries are made in English. He asked what kind of noise was introduced. Srihari responded that accuracy is over 90 percent for tokenization but there are no hard numbers on entity tagging. The performance on Chinese entity tagging puts it in the top three systems based on ACE benchmark data.

PANEL DISCUSSION 1: CHALLENGES OF MULTILINGUAL PROCESSING – SCALING AND TAILORING FOR GLOBAL MISSIONS

BOOTSTRAPPING TRANSLATION CAPABILITIES

Laurie Fenstermacher introduced Ray Slyh and his presentation on bootstrapping translation capabilities. Indication and warnings is a global mission. There are some capabilities and some good translation tools available; however, they do not cover the entire spectrum of languages. High quality translation may not be a high priority for a low demand language, but ultimately some kind of capability must exist for all languages.

Ray Slyh works for AFRL on speech processing and translation issues including broadcast monitoring. Ray talked about solutions for low priority languages. He addressed how the community meets the needs for bootstrapping translation capabilities.

Ray Slyh, Air Force Research Laboratory

Ray Slyh works in the Human Effectiveness Directorate at the Air Force Research Laboratory (AFRL). There are approximately 6800 languages in world. While the Department of Defense (DoD) has a strategic language priority list, it needs at least some capabilities across a wide spectrum of languages. The DARPA GALE and predecessor programs have invested lots of money in translation and speech recognition for some priority languages such as Arabic and Mandarin, but the reality is that the DoD cannot afford the same level of effort for all languages.

The amount of training data required is a major driver of cost and development time. For speech recognition, one typically wants hundreds if not thousands of hours of transcribed data for training. For training large translation systems, one typically wants on the order of a hundred million words or more of parallel (i.e., source and translated) text. The costs of collecting, transcribing, translating, and performing various quality checks on training data can be substantial.

Complicating matters is the fact that languages change over time, so speech recognition and translation systems must be adapted to new words, people, and concepts. A system cannot be considered static or it will become stale; systems periodically need to be updated.

There are also many genres of sources to consider. For example, translation and/or speech recognition systems trained for broadcast news may not work well for other sources, such as internet blogs or audio found on various web sites. For each genre that one wants to process, training data need to be collected. Thus, there is a real need for tools, techniques, and algorithms to rapidly develop systems. The systems must not require a lot of training data and must be easy to update and maintain.

If one is interested in the finer nuances of meaning to support sophisticated downstream processing, one must be careful that the translations performed on the training data are done with this in mind. To date, much of the training data used by system developers has not been translated with such a goal in mind, so users need to be careful when using these systems. Of course, very careful translation of training data will add to the cost and slow down development time.

While the government cannot afford the stance of "there's no data like more data." Error rates are typically higher when systems are trained with less data. There are no silver bullets immediately on the horizon to dramatically reduce error rates with small amounts of training data, but a number of groups are working on this task.

AFRL is looking at ways of using additional front-end features to develop speech recognition systems with smaller amounts of training data. Some of the DARPA GALE participants have looked at training speech recognition systems on smaller amounts of training data and using these to bootstrap the development of improved recognizers using large amounts of untranscribed training data (i.e., a semisupervised approach). AFRL, the GALE participants, and others have been looking at ways to combine systems with complementary capabilities. The system fusion of different feature sets or classification algorithms can be especially useful when there is a high error rate.

For translation, small amounts of training data can be particularly problematic for morphologically complex languages such as Arabic. Such languages attach various prefixes and/or suffixes to base word forms, and large numbers of such combinations will not be seen in small training sets. Morphological analyzers are systems that segment "words" into their prefixes, stems, and suffixes, and these systems can help alleviate data sparsity issues in developing translation systems. Research is being conducted by many groups to improve morphological analyzers for various languages; however, the solution may not require a top notch morphological analyzer. Rather, a quickly developed system may provide more output for the resources invested, because of minimal gains in pursuing the best. AFRL has been pursuing the latter approach with some success. As with speech recognition, combining translation systems with complementary capabilities can provide considerable benefit.

In summary, research is ongoing to address the performance decrease with smaller amounts of training data, but it is important to manage end user expectations.

Discussion:

Fenstermacher stated that part of the problem occurs when a new area of the world suddenly becomes the focus of decision makers. When there are not enough linguists, the most promising arena is to use tools to sort data for scarce linguists. This is especially important in the speech-to-speech realm. Humanitarian crises, such as the tsunami in Indonesia, are a good example of this. There needs to be some kind of capability however imperfect.

Sanfilippo asked about efforts to improve narrowing the domain of relevant sources. If a tool can narrow the domain or use a classification system to isolate the most relevant documents, it saves the linguists time. Linguists can also go through and tag topics rather than translating all documents that come across their desk.

TRANSLATION FOR HIGH DEMAND LANGUAGES

Automated Contributions to Exploiting Multimedia Sources at the Technical Support Working Group (TSWG)

Kathleen Egan, TSWG

Egan stated that while the last presentation was presented from a research perspective, this one is presented from the operational perspective. The research community has to put on a user hat to bridge the gap with the operational community. The TSWG approach is to bridge the gap and respond to the emerging needs in the operational environment by partnering with research entities, developers, and operational users.

TSWG has effectively partnered with DARPA for transfer of research outcomes and components. TSWG fosters innovation and new approaches, but does not fund basic research or acquisition. Together with others, TSWG serves critical operational interagency information extraction and language translation shortfalls with near term technical solutions. TSWG's mission and resources do not extend to full life-cycle support of fielded systems. Ultimately, what is needed is a full government integration of systems in an operational environment to manage work flow.

There is no perfect human translation, so expecting a perfect machine translation may be ambitious.
Accuracy matters in translation. Language is at the heart of human communication. Words have infuriated populations, created misunderstanding, fueled hatred, and built barriers across cultures and civilizations. At the same time, words have opened doors, reconciled groups, gained peace, and created laws for justice, order, democracy and freedom. Culture and language can save lives.

However, translation is not sufficient to exploit data. With too much information, something might get missed. Therefore translation by itself is not sufficient and needs to be combined with other tools to support analytisis. Human analysts and Language analysts do not have enough time to go over all the translated data.

The goal at TSWG is to mitigate the current errors in the Human Language Technologies. By taking something that is not perfect and building it into something analyst can work with and trust, both operations is satisfied and the developers learn from working with real users. In the past few years, TSWG has been putting technologies into the hands of users to get their feedback. It took ten years to improve the Arabic model, six years to do Chinese, six months to do Farsi and planning on a few months to do Bahasa-Indonesian. If the methodology is good, developing a translation system becomes easier.

TSWG has a focus on combating terrorism. Sometimes translation may not be the best tool for the job. Sometimes text extraction could be more important. Trying to build the perfect solution takes too long and does not improve the state of the art. Inserting technologies at the right time and in the right environment is a challenge and objective.

The language program focus areas include the following:

- Foreign Media Exploitation from Open Source
- Conversational Speech Triage for Added Intelligence Value and Tactical Operational Activities
- New Languages
- Tools for Analysts and for Language Professionals
- Integration into Operational Workflow
- Cultural Awareness and Language Learning

End to end capability assumes certain developments from the research field. First, it assumes basic research and algorithm breakthroughs in capabilities such as speech recognition, machine translation, information extraction and other language processing technologies. Second, it assumes development of research engines that demonstrate capabilities (i.e., reductions in word error rate and/or translation edit measures) in a scientifically controlled experiment and results are published by the National Institute of Standards (NIST). Third, it assumes demonstration of utility through research prototype engines that can interface with casual users.

It is time to start working on languages other than Arabic and Chinese. Furthermore, analysts must be given tools in their own language to work with to enhance their own productivity. The challenge is inserting these tools into the operation workflow so that analysts can use them on a day-to-day basis.

An end to end capability also assumes the following from development and engineering (i.e., technology transfer). The first assumption is a proof of concept/demonstration system resulting from lessons learned from the research prototype. The second is an operational prototype with some engineering to withstand regular users but with hand holding from engineers. The third is an operational system that is hardened to transition in the user's environment. The last is integration of the system in the daily work of users and in an enterprise architecture.

Developers need to get analysts involved from the beginning. The government does not invest enough money in the beginning stages of a program. Developing useful technologies is an iterative process between the developers and the users. The government lacks a strategic view of moving research into the user environment.

The ability to query is just as important as machine translation. Most query tools available today require exact matches. The user needs to be able to query more broadly.

Operational users are the final judges of application's utility. "Don't take it back" is the highest praise a tool can receive from the users during the testing phase. Operational users are in great need of multilingual translation tools to help them accomplish their jobs. They have little time to rely on human translators to validate output. They need transparent tools to resolve their tasks. However, machine translation capability by itself is not yet reliable for full dissemination out of the box. But workflow systems can help speed the integration process and to meet operational needs. TSWG has success stories developing the BBN Broadcast and Webmonitoring, now deployed at CENTCOM, SOCOM and many other operational settings in the US and overseas.

The multilingual tools need to be developed to allow analysts to focus on what matters. Tools need to automate the ingest of data and cluster data and get rid of duplicates. Users need to be able to customize filters so they are only alerted about information that fits their needs. Systems need to include a flexible query system and the system needs to be able to learn from users. Ultimately, the tools need to reduce the analysts' workload and be a user-friendly system they can trust.

Discussion:

Cabayan stated that most people in the room will not interact with a real operator. In the early 1990s, the COCOMs were each doing their own analysis resulting in analytic chaos. They realized that they had to build a better integrated infrastructure. However, the DOD still lacks a true mechanism for integrating tools and analysis. What is needed is a human domain analysis reachback center. Currently, even when useful tools are deployed to the theater, they often get lost in the frequent troop rotations. The time to coordinate with USDI, USDP and DDR&E to create a long term solution is at hand. Pushing these tools to the commands is not a long term solution. There needs to be one place where tools and training can be delivered to analysts. There is a great research community out there who are unable to connect with the users.

Ben Riley stated that the dilemma is that the department has a requirement system. When something becomes a requirement, money and infrastructure for training and maintaining it are assured. However, this area and tools are still considered boutique products by many. There is not even an agreed set of terms of what these tools and problems are. That is why all of this work is funded through supplemental money. There is no sustainment. Many are content to keep it outside of this system because there is lots of supplemental budget money available, but it will come to an end sometime. Furthermore, the work is supported by contractors and there is no training money. The problem with supplemental money is that the user is not well represented – it revolves more around the interests of the researcher.

Riley stated that if you look at JIPOE, it is a very powerful idea. It focuses on an important area, but none of it is sustained within the budget. An airplane will win over information systems every time. One almost has to be subversive to initiate action. The system needs to be embarrassed into taking action. Rational arguments will not cut it.

Porter stated that there is a paradox at core of this. There are a lot of software developers who have no idea of actual problem because it cannot be discussed.

ARGUS

Frank Connors, Defense Threat Reduction Agency (DTRA)

Argus was supposed to be the third presentation in this panel, but Frank Connors was not able to attend. He submitted a word document instead, which is recorded here.

Project Argus is a global biological event detection and tracking capability that provides early warning alerts to the USG.

When speaking to the intelligence community, Argus shies away from using the word "analyst" as that word engenders a distinct definition within that community. On the other hand, in the appropriate settings, Connors refers to his valuable linguists as Argus Analysts.

Currently, Argus has a human linguistic capacity in 37 languages for Bio-Events work. For the Argus-N effort, there are six languages. Bayesian nets must be developed for all languages.

Argus analysts are native speakers (i.e., they are not only fluent but most of them were born in country and lived there long enough to have a good understanding of the cultural nuances of the language and how to properly translate and put the text into cultural context). One gap identified in an evaluation report is their fluency certification.

Chris Morrell (OSC) is working with Argus to ensure that our linguist meet USG Standards for certification. Argus analysts *are not* technical subject matter experts. There is an initial two week program to train them on the Argus hardware, software and rudimentary disease identification. Training for the nuclear Watchboard is done by Brad Clark.

Over time, the analysts get very good at detecting and reporting events. But the final word on what the event really is (i.e., what is posted to the Watchboard) is subject to the final validation of a physician on staff for the human bio events. As the agricultural/plant taxonomy comes on line, we have a plant pathologist on staff as well as experts from the Animal Plant Health Information Service (APHIS), United States Department of Agriculture (USDA), Fort Collins, Colorado. For the nuclear work, Brad Clark is the SME.

As previously briefed, the big fire hose of articles first goes through machine processing through the Bayes' net and only those that "make the cut" are read by the analyst. On average, each of the analysts reviews 300 articles per day.

As Argus begins works in areas other than Bio, it is seeing communities of interest (COI's) emerge. The Biological Indications and Warning Analysis Community is a federated group of USG biologists working in various agencies and all use the Argus Bioevents Watchboard. It is possible to consider that Argus and other tool sets may be instrumental in developing other COI's in the areas of Chemical, Nuclear, Radiological, Criminal, etc.

SESSION 2: "YOURS, MINE AND OURS": OPERATIONALIZING THE JIPOE NEXUS

UNDERSTANDING THE PROBLEM – LESSONS LEARNED

Fenstermacher introduced Susan Numrich and Carl Hunt of IDA to provide context for understanding the information challenge issue. Stanley Horky from General Motors then talked about transitioning technologies from a corporate and commercial aspect. This presentation focused on the JIPOE nexus.

THE NEXUS FEDERATED COLLABORATION ENVIRONMENT (NFCE)

Sue Numrich and Carl Hunt, IDA

Carl Hunt asked what it means to operationalize a federated environment. What does it mean when you have information marshaled? How do you distributed analysis and fuse the information? Unilateral/bilateral relationships among entities typically define how the Interagency and NGOs interact. Synergy and integration are difficult and innovation is not shared effectively. Leadership and alignment are unclear. This is the foundation of the operational dilemma.

The basic purpose of NFCE is to articulate and demonstrate the feasibility of sustained federation and collaboration between interagency partners where relationships are mutually supportive within a widely scalable and adaptive manner. The relationships and the outcomes the partners produce must be transparent and targeted towards maximum objectivity in the use of knowledge and generation of relevant directions for meaningful inquiry. These are key ingredients for the Nexus Federated Collaboration Environment.

The Nexus, however, is not the JIPOE. The Nexus is the supporting and underlying federation of diverse organizations that compose the WMD-T JIPOE effort, and the beginnings of the organizational design, processes and information technology infrastructure that would make it possible for people to work together to defeat WMD-T in the hands of terrorists, most notably non-state actors. The NFCE seeks to describe the application of some of those social science insights to our own behaviors as collaborators and partners in the JIPOE endeavor.

The JIPOE Nexus (NFCE) is a virtual place that transcends the center and the edges of its member organizations, facilitating the linkages of the members through concepts of social network science.

Federations do not exist long term within the USG except in extreme crises. Innovations occur in "cylinders of excellence." Even when entities do come together in a crisis and clear management is designated, individual organizations may dominate or lose influence. Some even drift off, taking their distinctive contributions with them. Innovation and collaboration may be stifled. Leader partnerships may even wane or disappear. Quite often, lessons learned and innovations are lost. Synergies dissolve and pathways that motivated and incentivized are obscured in the return to business as usual.

Organizations within the federated nexus environment sustain both their individual and corporate identities but through a common purpose based on transparency, bias mitigation and collaboration. Lead organizations may be designated, but they lead on behalf of all organizations within the nexus, aligning to purpose and enforcing maximum collaboration and sharing. Incentives emerge and are propagated. Hunt recommended leveraging the working being done in JIPOE and taking a more systemic approach.

The major concern is that even in 2008, the USG risk losing "the battle" due to inconsistencies of mission and individual organizational objectives. Inconsistencies block the path to protecting the greater interest. Incompatibilities of mission, force and resources self-deter success. The vision of the JIPOE "Nexus" is to create a virtual "place" where adaptation and innovation are encouraged and sustained. To do this, JIPOE attempts to create synergy of two important contemporary organizational and information management capabilities, aligning them to strategic intent.

Sue Numrich is writing a paper on the key attributes of the nexus, which will be released shortly. These attributes include:

- Dynamic alignment -- people, process, policy, and technology
- Transparency -- data and process
- Continuous improvement -- informed by performance metrics
- Adaptive
- Continuous learning -- evolutionary environment
- Flexible -- Framework defined by problems and results vice locked-in areas of responsibility, authorities
- Seamless -- All are "prosumers."

Prosumer refers to everyone connected within the Nexus. Each person is capable of and subject to the production and consumption process anytime and at the same time.

When thinking about systemic change, the key is to design the framework to empower and exploit convergence. Alignment is a very difficult objective. The USG needs to find ways of making individual organizational objectives align with federated objectives.

This effort needs to be a sharing effort. Diverse communities from both the classified and unclassified level must work together. There must be infrastructure built for sharing with allies. It is within this nexus that the USG can make a real difference to the warfighter.

INDUSTRY PERSPECTIVES

Stanley Horky, General Motors

Stanley Horky spoke to programmatic success factors and action from an industry perspective. His work at GM focuses on alternative fuels, alternative propulsion systems, energy, and advanced materials. He has been and is currently working with the DOD.

Horky brings in a perspective as an outsider. The discussion boiled down to assessments in pairwise comparison of programmatic success factors for high visibility programs to succeed. GM has familiarity with high visibility programs including fuel cell and electric vehicle development such as the Volt program. But it is important to note that the success factors described in this brief are independent of the type of program – defense or commercial sector. They are independent success factors. These are consistent with early to late stage programs. It is important that these factors are considered in how the USG evolves its programs.

The first category is strategic activities. These are listed in order of priority. First, the project must be unique and must satisfy a critical need. The work must be on the edge of research technology and the program always wants to be pushing towards that edge. Second, the decision maker must act as a "champion" of the program and actively support it. Third, an active senior executive advisory board needs to be established to represent the major stakeholders. Within the USG, it is important to have Congress involved from the beginning or you will lose. Fourth, the program must identify the final customer and need base. The "clarity of purpose" must be firmly established. The program should consider involving the customer in the management of the program. It is a difficult step to take, but it encourages customer support and buy in. Fifth, an "executive shepherd" must be identified for budgetary advocacy both internally and externally. Sixth, the key policy or decision maker staff must be kept informed and engaged. Finally, the program must highlight goals that "capture the imagination."

The second category is budgetary concerns. Budgets kill major programs. If you do not have continuity established, you will hit crisis points that you cannot control that will destroy program. If stakeholders are not committing budgets and resources, they are not a true stakeholder. The first success factor is ensuring long-term (multi-year) budget continuity. The second factor is establishing stakeholder budget and resource commitments.

The third category is creating a centralized program with centralized control. Leadership is essential to keep researchers from going off in their own directions. There are several success factors in this category. The first is that the organizational structure must be appropriate to the program stage. There are formal and informal chains of command. Encouraging redundancy in communication is extremely important. The program should establish informal mechanisms to manage the process from top to bottom. The second is to ensure key stakeholder management styles are compatible including complementary strengths and driving forces. Third, the program should recruit personnel based on enthusiasm and commitment to project goals. Put people on the program who want to be on the program. The program does not want passive actors on program. Fourth, allay potential stakeholder fears relative to programmatic, proprietary, and budgetary issues. If you do not get stakeholder buy in, you will not get support, funding, etc. The leadership team must be able to deal with these issues. Lastly, the program should maintain a conscious effort to network for the highest quality program execution. Nobody should think that you are not dealing with the best tool, techniques, etc. Any hint of mediocrity and you are in trouble. This is more important than cost.

The fourth category is creating a cooperative environment. First, program leadership must be coordinated and actively supporting and advocating. Everyone must know who the leadership is and must be engaged with them. In his early stage program, the researchers and management should be in the same geographic location. Second, positive stakeholder relationships must be actively managed. Third, "great chemistry" is apparent. People need to interact positively with each other and feel a sense of accomplishment. Fourth, utilize the best "state of the art" technology tools in program execution. The program should always push the envelope, especially in the early stages. The end results will be better.

Discussion:

One audience member asked about bias mitigation. Where does competitiveness come in? Horky responded that this issue goes back to whether you can resource competitive groups. Competitive groups can push each other in a positive way, but it can also become destructive. The management team needs to step in and manage competitiveness.

Hunt asked how the program ensures that leaders get honest feedback. How do you challenge leaders to look out for the whole federation? Horky responded that the senior leader cannot be viewed as championing a particular agency, but must represent the federation so that there is no question of bias. As far as individual areas of concern, they will have biases, but there will be counterbalance of leadership team, who should be able to get over biases as a group.

Horky stated that there are three critical decision points in any major program. Is it real? Can you do it? Should you do it? If an early stage succeeds, the program will have a valid proof of concept. The next steps are invention, engineering, and political questions. If you do not have a valid proof of concept after the early stage, do not proceed because phase two teams will not wait for innovation.

One member of the audience asked how automakers work together to share advances, which is similar to the USG interagency process. Horky responded that in the hybrid electric technology program, GM has relationships with other automotive manufacturers to deal with mature technology that everyone wants access to. At that point the companies are willing to negotiate an alliance agreement. The objective is not to try to change the final customer's interest, but to align 80-90 percent of the elements to meet needs. If you can align interests by putting people on the program, you are more likely to succeed because they have a vested interest.

ART OF THE POSSIBLE: FUNCTIONAL DOD FEDERATION

Sue Numrich introduced Dr. Amy Henninger from the Institute of Defense Analysis. Numrich stated that Henninger will speak about how interagency sharing might work. When you bring disparate groups together in the USG, there is always security problem. It is the first thing raised. The second thing raised is that the USG is resource constrained in people and funding. That translates into groups being unwilling to share their scarce resources to work with others on a new need. Henninger spoke about an interagency DARPA program that successfully worked in a federated environment.

At the invention phase in late 1970s and early 1980s, DARPA created and integrated various flight simulators together to train aviators. The only way to beat the machine in simulation is to put the human into the system at random points. What DARPA did was to make sure the simulator could work both independently for specific agencies and cooperatively in a federated environment. The services had to collaborate to make the simulators work. In this instance cooperation was encouraged by taking away funding from each service unless they cooperated on this development. These kinds of cooperative environments take time and effort. Henninger will talk in more detail about these issues.

Amy Henninger, IDA

Yours, Mine, Ours: Art of the Possible: Functional DOD Federations

While creating federations within the USG is a difficult undertaking, successful examples of cooperation exist. The problem is not in the technology development, it is in the process of creating a federation.

Federating modeling and simulation (M&S) systems are an important area of consideration. These systems support training and mission rehearsal; acquisition and testing; and analysis and experimentation. The simulations need to be able to exchange data and need to be able to understand the data that is exchanged.

The benefit of a federated M&S system is that it avoids monolithic "one size fits all" models and simulations. It establishes processes and mechanisms to integrate heterogeneous systems. And it promotes reusability of individual components. Challenges include cultural barriers, security barriers, and data barriers.

The Joint Operations on Urban Synthetic Terrain (JOUST) use case is a good example of a federated system. The objective of JOUST was to rapidly demonstrate, assess, and transition a capability to train forces in joint urban operations (JUO) using a mix of live, virtual and constructive (LVC) training environments.

In the end, the JOUST case showed that functional DOD Federations exist and are flourishing. In fact, the number of federated scenarios, DOD-wide, has been estimated at over 13,000 per year. Creating federations allows agencies to engage in cooperation without losing sight of an organization's core mission

Discussion:

Numrich stated that simulation is a boutique capability. But federation is the process which allows groups to work together. Federations are both possible and functional. However, it requires a high level person to initiate and support a federation, especially in regard to budget challenges.

BIOLOGICALLY INSPIRED INNOVATION IN ACTION

Carl Hunt, IDA

The solution to the information challenge may closely resemble biological synapses. Synapses are specialized junctions though which the cells of the nervous system communicate with non-nervous system cells. They are the foundation of perception and thought and allow the brain to communicate and coordinate the rest of the body.

SINAPS, in DOD terminology, is a small world interagency process system. It is the movement of information through billions of connections transmitted with great fidelity and speed.

In small worlds network theory, one starts off with an organization. Relationships build between nodes creating local communities. It does not make sense for all nodes to be interconnected and yet that is what the DOD often tries to do. Small worlds theory shows that a system can be highly effective without having to connect all the nodes to one another.

The system does not need to understand the message going through the system, it must just maintain fidelity. People already do this – tools are an enhancement of what people already do. The nexus is about people; it is not about the technology. SINAPS is essentially a virtual fusion cell in which every participants thinks of himself as a prosumer.

Within a SINAPS, there are hard and soft linkages. Hard linkages occur between organizations. Soft linkages occur between people who influence the organizations. SINAPS works best in a collaborative, transparent environment that helps prosumers remove biases and move forward.

Discussion:

Sanfilippo stated that within information fusion and the exchanging of protocols, one element that seemed relevant was the creation of schema that allows for the establishment of common language. How do you manage to preserve richness of the model when you share it with others? Hunt responded that on a philosophical level, you do not want to lose sophistication in order to share.

Pate stated that one of the things that M&S community has learned is how much of the object model to embedded in the architecture. As people started using distributed simulation more and more, they found the system to be less robust . Over the years, the object model became separated from architecture, which allows the users to tailor as needed. The systems are not as robust in using ontologies. That is the next generation.

Sanfilippo stated that it is more like intersections. Common information is shared but the richness of the materials stays with the agency.

Hunt stated that this goes back to the fractal nature of schema. There can be separate ontologies that interface at a known level as necessary.

Hunt stated that there is a cultural issue here as well. The workshop focused on the challenges inherent in large data sets, but people will judge progress on the performance of JIPOE. It is not a product, but it will be judged by some tolerance to false positive and no tolerance for false negatives. Some organizations call false positives all the time, but those organizations do not last very long. In a table top exercise several months ago, an effort was made to model a federated project. One group refused to state a hypothesis, one group came very close then veered away, the third group was all over the map. It is important to understand the cultural perspectives that each group brings to the table.

Hunt stated that the USG must incentivize change in order to achieve a federated environment.

Cabayan spoke about cognitive issues at the group level. He asked what kinds of things a JIPOE-like cell should incorporate at a cognitive level.

A member of the audience recommended using a simulated environment.

Numrich stated that a major driver is how people think about problems. People with work experience close to "boom" are most concerned about clear ties to everything. Operational users could accept more flexibility. The research community is ready to explore any option.

TECHNOLOGY PREVIEWS

CHARLOTTE VISUALIZATION CENTER, UNIVERSITY OF NORTH CAROLINA, CHARLOTTE

Roadmap for Visualization and Interactive Analysis
Remco Chang

During a previous workshop, Remco Chang spoke about the specific application of visualization tools. During this workshop, he talked about where visualization is going.

In the short term, the goal is to develop interactive tools for analysts to identify trends and patterns in massive amounts of data. Much progress has been made on this front. The common problems faced today are large data sets and the emphasis on human operators to discover patterns. Interactive tools help analysts sort through the large expanse of data to find patterns and trends not obvious to the analyst.

Farther out, the goal is to develop knowledge storage and analysis through understanding of user interactions. Once operators have begun using the tool and imputing knowledge into the system, the developer wants to know why and how analysts use the tools. The next step is then knowledge storage. Chang has just begun this work with some success monitoring operations with specific visualization tools. Some of the intent is derived as part of a reasoning process through post mortem analysis.

In the long term, the goal is to aid the user's decision making process through mixed initiative systems. The objective is to get the computer to do some of the work for and with the analyst. Computer-based tools can be used to forage for information or to guide or hint to the analyst when he or she is going down the wrong path and direct them to a new one.

During the demonstration period, Chang displayed two current projects. The first was a financial wire transaction detection program, supported by the Bank of America. It is currently beta-deployed at WireWatch. The second project was Visual Analysis of Interaction of Investigation. It visually depicts interactions and is being evaluated by financial analysts.

The Charlotte Visualization Center will team up with Dr. Stephen Shellman of the College of William & Mary (W&M) to help with the visualization aspects of the Civil Strife project.

ARTIFICIAL INTELLIGENCE LAB, UNIVERSITY OF ARIZONA

Dark Web Forum
Dr. Hsinchun Chen

Chen called the special section of the Web that is used by terrorists, extremist groups, and their supporters the "Dark Web." Information within Dark Web forums represents a significant source of knowledge for security and intelligence organizations. As is mentioned in Leaderless Jihad by Marc Sageman, extremist groups have moved from primarily face-to-face interactions to web interactions. In response to this threat, the Artificial Intelligence Lab at the University of Arizona developed systems supporting the large-scale collection, search, and analysis of Dark Web forums, specifically addressing the needs of security analysts.

During the collection stage, the Arizona Forum Spider creates an automated collection of forum communications and other content of interest. The spider only searches primary sources of information, which may include web sites, blogs, and video sites like You Tube. It collects information from 80-100 jihadist forums regularly. The spider is disguised to look like a human user.

During the search stage, information is deposited in the Arizona Dark Web Forum Portal. The goal of the portal is to allow the user to browse, search, and analyze the Dark Web forum collections. The current portal contains more than 2.5 million messages across seven forums. Within this stage, the Arizona Sentiment Analyzer, which is a web based portal for the analysis of opinion and emotions expressed in Dark Web forum communications, enhances the understanding of forum communities and participants.

During the analysis stage, the tool uses the Arizona CyberGate Text Analyzer, which is a comprehensive system for the analysis of forum communications. It uses automated text mining and visualization methods.

The effort focuses on terror threats, but also can be applied to criminal groups. It can process information in English, Arabic, and Spanish.

Web forms in the Dark Web offer extremists a rich medium for communication of their ideas. Likewise, information contained within Dark Web forums represents a significant source of knowledge for security and intelligence organizations. The tool developed by the University of Arizona is a comprehensive approach to Dark Web forum analysis and spans collection, search, and analysis.

BBN TECHNOLOGIES

Foreign Language Media Monitoring
Sean Colbath

A series of maxims capture why it is so difficult to monitor the media. These include:

- Communication is a two-way street
- There will never be enough linguists
- There is too much media to be monitored
- There are too many media types to monitor

While the problem space is large, various developments will certainly aid in the creation of a solution. One solution is for people and machines to work together. If machines are allowed to do what they do well (process mass quantities of information; filter, sort, quantify and prioritize; and never sleep), it will aid the analyst to do their job well (interpretation, judgment, assessment, cultural insights, and domain expertise).

CENTCOM Joint Staff Intelligence (J2) Open Source Intelligence (OSINT) Section has been using the BBN's Broadcast and Web Monitoring systems since 2004 funded by TSWG that allows non-linguists to access multilingual data from Broadcast and Web sites. The systems are deployed in multiple operational sites. From this, BBN has gleaned important insights. First, English-speaking analysts can use machine translation for alerting and selection of content. Human linguists can use transcription, translation, audio/video synchrony to produce high-quality translations. Analysts and linguists do not have to be physically co-resident. Finally, automation gives additional coverage when there are not enough analysts.

Languages are not one of America's strong suites, but technology is. Technology should be use as an assistive aid to cross the language barrier.

ONTOLOGY WORKS

Ryan Cole

Ontology Works is a product company offering a broad suite of semantic technologies including deductive information repositories (Ontology Works Knowledge Servers), semantic information fusion, and cost effective semantic federation of legacy databases, ontology-based domain modeling, and management of the distributed enterprise. Ontology Works seeks to create and employ ontology based databases.

These kinds of tools employ ontologies, such as those created by Larry Kuznar of NSI, against specific data sets, such as radiological nuclear data. The tool integrates these elements together in order to answer questions about potential indicators and connections between data.

SAS / DETICA

David Porter

Detica is focused 100 percent on high volume information intelligence in areas such as commerce, fraud, and organized crime. SAS is a leader in business intelligence, predictive analytics and data management. Together, SAS and Detica have created an extensive analytic toolkit to test insight-led strategies. Some of the tools and techniques used by Detica/SAS include data matching, data quality, and neural networks. The Detica/SAS Data Driven Environment delivers all components required to manage and act on multi-intelligence (INT) data rapidly

One featured tool, NetReveal, allows analysis of data from many sources. It takes raw data, creates social networks, and prioritizes high risk networks.

No matter what type of tool Detica/SAS builds, it is designed on a conceptual model that allows for changes in the environment, requirements, and client needs. The critical node is in the interchange between analytics and automation. For example, all types of intelligence (the "INTs") can be represented as an ontology. The output from analytics can be represented categorically. If you recognize that, you can mix them all up together. This results in several advantages. First, even if analysis has no relationship another INT, you can use the same set of tools to analyze them, creating a productivity bonus. Often there is a keystone document that allows one to link data across domains. Second, a global system helps the analyst understand where the gaps in information are located. Ultimately, this Detica/SAS tool is good at bringing other tools together.

SOCIAL SCIENCE AUTOMATION

Robalyn Stone

Social Science Automation (SSA), Inc. provides text analysis products and services to business, government, and academic clients. SSA also provides media analysis, campaign and election media evaluation, profiling, and forensic psycholinguistics all rooted in the company's core competency of automated text analysis.

SSA works with the Defense Advanced Research Projects Agency (DARPA) and BBN and has gained the attention of government clients through its media analysis. Today, the company is doing multilingual analysis, which can be process in near real time. An analyst can go online, go to a news website, download articles, and feed it into the SSA media analysis tool. Right now, the tool works in both Arabic and English and work is being done to expand to other languages.

SSA has completed efforts to attempt to forecast or anticipate international instability using international news feeds. The outcome of the effort resulted in a general tool that measures tension in news feeds – how "good" or how "bad" the news feed is. Media analysis can also be done on a similar scale measuring whether media is generally good versus generally bad. The tool can also compare and contrast English with Arabic language media in terms of identifying the most common descriptors of items or issues of interest, such as the US military.

NUTECH SOLUTIONS DIVISION OF NETEZZA

ABEM: Agent-Base Evidence Marshaling for Distributed Multi-Intelligence Analysis

Mario Inchiosa

The intelligence community is faced with large amounts of heterogeneous, distributed data. To deal with this massive amount of information, Nutech is developing self organizing information agents based on biologically inspired algorithms. They are simple agents that engage in local interactions in an explicit space. The process is based on agent-based evidence marshalling.

Agents containing tuples (e.g. Concept-Relation-Concept) can deduce new tuples, exchange tuples with their neighbors, and move within their environment.

In a National Institute of Standards and Technology (NIST) study, ABEM tripled the performance, doubled the accuracy, and halved the reported difficulty of an analyst's task.

The cluster amplification effect suggests an interesting spin-off for ABEM technology, namely its use as a preprocessor to enhance the performance of statistical document clustering algorithms. The effect illustrates how semantic understanding can be used to improve clustering.

ABEM's distributed, population-based approach is designed to scale well. ABEM has demonstrated excellent scalability when applied to geospatial problems.

In summary, ABEM is a self-organizing approach that synergizes deterministic and stochastic algorithms, and enhances analyst performance and text clustering. ABEM demonstrates interoperability and fertile cross-source inferencing. Finally, it exploits Small World Networks to speed reasoning and enable scalability via parallelization and distributed computation.

SAIC

Automated Behavior Analysis Extraction

Gary Jackson

Automated behavior analysis (ABA) is an automated assessment tool used by JIPOE to generate WMD-T COA hypotheses. ABA fuses the social and behavioral sciences with computer sciences for automated modeling. Automated Behavior Analysis is based on decades of validation and automation. Because applied behavior analysis is difficult and time consuming, the ABA model presents automated assistance for use on real world problems using information in electronic form.

Applied behavior analysis is a sub-field of psychology. Early models in psychology demonstrated that behaviors occur in response to environmental antecedents and is maintained by consequences of the behavior. Early psychology models were not capable of prediction. The proprietary ABC Predictive Model extended the early psychology model to include prediction, influence, pattern classification, and analysis within asymmetric warfare environments. ThemeMateTM and AutoAnalyzerTM applications present a behavior analysis model that provides prediction, methods of influence, and automation. Forecasting models have been validated across individuals and groups.

The objective of automated behavior analysis is to discover antecedents to a particular problem. It is very difficult to obtain actual observations, so the tool is based on text accounts. Using the five w's (who, what, where, when, and why) the tool can extract antecedents. The tool can also use multiple sensor data instead of text. The tool has been used to identify cyber attacks, anticipate hacking behavior, and identifying terrorism antecedents. However, it can be a complex process, especially if monitoring different groups or countries.

During the afternoon's demonstration, the tool was employed to look at 53 terrorist biographical sketches to look for commonalities and correlation.

OPEN SOURCE CENTER

Nancy Chichor and Roberta Dobbins

The Open Source Center (OSC) has an emerging media group that looks at all forms of new media including second life and web cameras. It also supports Argus, which is an early warning system for emerging diseases. Another big project is the Large Scale Internet Exploitation effort, but it is in an early prototype stage.

There other tools in the same problem space that deal with information challenges. Pacific Northwest National Laboratory (PNNL) created a tool called IN-SPIRE. It was developed by the lab to deal with large amounts of unstructured text data. It allows you to deal with large amounts of data collected in non-standard way. It also looks for themes and is used analytically. The tool is very rich and very complicated, but if you have a complicated problem, you need a complicated tool.

OSC is starting to split off more useful tools that allow analysts to quickly share results without burdensome interactions. The tools allow analysts to explore data more effectively. The OSC also has the Voice of America dataset in which they can do trend analysis, correlations, alternate competing hypotheses, sentiment analysis, and query by example.

SAIC

Linguistic Pattern Analyzer (LPA)

Mark Clark, Anne Russell

Link Pattern Analysis (LPA) helps analysts do their work better. It is a tool to help structure world views and use linguistic patterns to find indicators of world view. The LPA compiles "hits" and scores the content in the article, which is directly related to indicators of multiple computational social sciences models. The LPA seeks to capture the ways (patterns) in which the indicators are expressed in multilingual communications.

LPA operates by searching the documents or linguistic patterns related to specific indicators of the models it is populating. The LPA looks across the corpus to detect phrase variations for the same indicators, compares the rating and generates a histogram of the score for the indicator at this sample time. Documents with evidence for any of the indicators are further processed, looking for information to rate the significance/severity of the situation indicated in the detected text. When all of the documents in the corpus have been processed, a histogram of the results for each atomic element is generated. The ratings are then coded, compared to the ratings of similar "hits" in the corpus and a nonparametric, aggregate estimate of the rating is derived.

SANDIA NATIONAL LABORATORIES

Christy Warrender

Sandia National Laboratories (SNL) is working on tools for ingesting and analyzing raw text. They created a text analysis library (called STANLEY) and then built a number of applications on top of that library.

Two example applications were built to support the Yucca Mountain licensing effort. There are 20 years of data on whether it is safe to store nuclear materials there. One SNL tool helps analysts discover and document how data from experiments and simulations support specific conclusions. Another helps classify emails as to whether they are relevant to the licensing effort or contain privileged information.

In one demonstration, SNL built models of individual people based on the text documents they produce. The tool lets you query or explore those models in a way similar to a search engine. However, the tool allows you to identify whose text contains topics that are most similar to the query. It reveals how different people talk about the same subject in different ways.

The demo presented in the afternoon also included a preview of Peter Chew's work (see multilingual text characterization below). This approach uses the Bible as a Rosetta Stone to allow cross-language clustering. Chew can do this in 56 languages which cover about 99 percent of documents on web. He has used it to classify documents by whether they were ideological or not and by type of ideology (Marxism-Leninism vs. Nazism) with greater than 90 percent accuracy.

Dr. Chew's work is not based on STANLEY, but STANLEY can support analysis in various languages because it is based on a statistical technique.

AIR FORCE RESEARCH LABORATORY

Advanced Text Exploitation Assistant (ATEA)

Mark Zappavigna

Mark Zappavigna spoke about automated text extraction of command and control organizational hierarchies and supporting behavioral influence analysis. AFRL created an ATEA for the National Air and Space Intelligence Center (NASIC) to support command and control and behavioral influence analysts. ATEA extracts entities (people, organizations, GPEs, facilities, locations, weapons, vehicles), relations (personal-social, general-affiliation, physical, part-whole, agent-artifact, org-affiliation), and events (life, transaction, movement, business, personnel, justice, conflict, contact).

The prototype software is a combination of commercial off the shelf technology (COTS) and government off the self technology (GOTS) products. Bringing the various technologies together allows researchers to validate and assesses the ability of the semi-automated system to support downstream analysis.

ATEA provides visualization to support analyst in discovering information. It also creates a model area of responsibility (AOR) to allow the analyst to see how things may change in the region as well as supporting downstream analysis.

Ultimately, ATEA extracts explicit information and shows analysts where it is located in the text. It has a comprehensive user interface with document lists, document view screens, and network charts.

STRATEGIC ANALYSIS ENTERPRISES (SAE)

Stephen Shellman

SAE & College of William & Mary (W&M)

SAE turns unstructured text into quantitative data for analysis. Shellman's research using such data focuses primarily on conflict processes, mostly involving state and non-state actors. The work attempts to discover what processes lead to escalation/de-escalation, actions/reactions in conflicts. Shellman is also looking at different tactics groups use on a daily level - disaggregating people, targets, actions, and events. He is building tools to disaggregate data in the conflict world focusing on the characteristics of state and nonstate actors to produce a more nuanced level of data. The data can further be characterized under an array of ontologies such as the DIME framework.

During the demonstration, Shellman demonstrated how to turn information from text reports (media, NGO, blogs, etc) into quantitative data. The software is language agnostic and was first used and tested in English and is now being used to code Arabic sources. The tools ask questions such as: what is the action, where did it occur, and on what date. Shellman's research focuses mostly on political events including coups, riots, protests, armed clashes, terror attacks, criminal activities, etc. The resulting data is useful for forecasting political conflict and cooperation, analyzing intended and unintended consequences of actions, performing network analysis, and generating early warning models. SAE also developed a new parser, which acts as a baseline for future analysis such as the analysis of sentiment.

PACIFIC NORTHWEST NATIONAL LABORATORY (PNNL)

Automating Frame Analysis

Antonio Sanfilippo

Frame Analysis provides insights on two issues. The first is how communication sources construct issues to influence target audiences, e.g., framing suicide bombing as "martyrdom." The second is how target audiences respond to framing, e.g., by resonating or not with the message conveyed by the communication source. Recognizing framing intent leads to an understanding of the goals of the communication source.

Frame Analysis was pioneered by Erving Goffman in 1974 and has become an important analytical component in the study of group behavior and social movements. Despite significant recent theoretical advances, there still is no systematic method to identify and marshal frame evidence in a time and cost effective manner.

PNNL's goal in automating frame analysis is to address current limitations in the representation, acquisition and analysis of frame evidence. Frame analysis gives users the capability to understand and analyze what is communicated, why and to what extent. If you understand intent, you can understand communication goals.

The tool includes an ability to leverage complementary approaches to frame analysis from sociology and political science. It also includes the ability to combine theoretical insights from frame analysis and linguistics with information extraction capabilities and content analysis methods. The project is supported by the Human Factors division within the Department of Homeland Security (DHS).

The PNNL tool is designed to annotate frames and extract naturally occurring frames in text. The PNNL annotation scheme helps the user understand how a domain is organized. PNNL developed a semantically-driven and visually interactive search environment to query and quantify frame evidence.

INDIANA UNIVERSITY

Travis Ross

Ross spoke about using virtual worlds and video games as experimental environments to explore social science problems. Virtual worlds like Second Life and World of War Craft can be used as experimental environments to test ideas and courses of action. Virtual worlds can be copied and customized to change one variable at a time. The worlds act as control groups. The game system allows researchers to alter incentive structures to look at problems longitudinally.

The demonstration illustrated how virtual worlds could be used in a commercial environment as well as how they could be practically used by researchers and analysts alike. Typically, a commercial version of a virtual world could take up to five years and require millions of dollars of investment. Indiana University is working on creating these worlds with fewer resources through a grant provided by the Federal Reserve.

IBM

Translingual Automatic Language Exploitation System (TALES)

Salim Roukos

The Translingual Automatic Language Exploitation System (TALES) is a multilingual, multi-modal analytic system that lets English speaking analysts collect, index, and access information contained in foreign-language news broadcasts and websites. TALES technology is built on top of the IBM Unstructured Information Management Architecture (UIMA) platform and uses multiple IBM natural language technology components. TALES is a derivation from the Global Autonomous Language Exploitation (GALE) program for building translingual systems.

TALES incorporates technologies from various sectors. These include speech-to-text, machine translation, named entity information extraction, and presentation of translated transcription. TALES supports many languages and has a dialect detection system for Arabic. It uses one of the most accurate Arabic translators and can be customized to specific domains.

SANDIA NATIONAL LABORATORIES

Multilingual Text Characterization

Peter Chew

The multilingual text characterization tool conducts network analysis in a multilingual framework. In one experiment, Sandia mapped the books of the bible. The result showed that the tool clustered the books by concept. The books of Mathew, Mark, Luke, and John were clustered because they contained similar content.

The multilingual text characterization tool is designed as an exploration tool; it is not designed to give the user answers. It groups ideas by proximity. For example, the tool grouped the Hamas charter near Ahmadinejad and the Iranian constitution. It raises the question of whether those three entities use the same terminology deliberately.

The framework allows the tool to look at multilingual documents, strip them of their language specific noise and isolate concepts within those documents. Sandia uses a Rosetta Stone approach, bypassing the need for translation. The method uses a numerical representation instead.

Discussion

Kelcy Allwein stated that it was wonderful to see the previews of the demonstration, but the key issue of integration has not yet been addressed. In the intelligence community, not understanding how to use a tool and incorporate it into the analytic process is a major problem. Analysts also often do not trust new tools. Integration is a key area that needs to be addressed. Tradecraft and business processes must be built around the integration of a new tool.

It would be useful to have a testing center to figure out the integration problem before it is sent to the analysts. The Office of the Director of National Intelligence (ODNI) is trying to accomplish that at the unclassified level. The center would assure analysts that the tool is effective and provides value within the context of existing analytic processes. A workshop on how to transition tools effectively to the user community could be a useful follow-on workshop. Transition is a critical component that the intelligence community rarely looks at.

WRAP UP

Laurie Fenstermacher concluded the workshop by thanking the briefers and the participants. She reiterated the importance of a federated system to deal with information challenges. She asked the participants to send her any comments for incorporation into the workshop report.

APPENDIX A: WORKSHOP AGENDA

August 27: Addressing Multilingual Information Overload Problems

0745 Arrive, badge, continental breakfast fare

0815 Orientation/ Overview to WMD-T JIPOE – Hriar Cabayan and Nick Wager

0900 Understanding the Problem: WMD-T JIPOE Information Ingest, Structuring and Characterization – Sue Numrich

0930 **Introduction to the Flow of the Day**

Preview Briefs for the Technology Displays

Ontology Works, University of Arizona, Social Science Automation, Pacific Northwest National Laboratory, Open Source Center, SAIC (Thememate), SAIC (LPA), BBN , NSI, University of North Carolina, NuTech, Applied Tech Solutions, Indiana University, Detica, Strategic Analysis Enterprises, Sandia, AFRL, CTO Translation Technologies

1200 **Lunch**

1300 – 1630 **Demo Session: View Demonstrations, Interact with Technical Experts (See reverse side of page)**

1630 Day 1 Wrap-up

August 28: Coding with Unstructured Text and Operationalizing the JIPOE Nexus

0745 Arrive, continental breakfast fare, reclaim badges

0815 Special Briefing: INSCOM Futures Directorate Initiatives/Projects, Russ Richardson

0845 **Session 1: Breaking it Down: Information extraction and Coding**

- Understanding the Problem – Lessons from JIPOE NS Case Study – Robert Popp, NSI
- Art and Science of Data Coding – Amy Pate, START
- Extracting Concepts from Large Data Sets -- Ed Marquardt, NSA
- Extracting Context from Unstructured Text – Mark Zappavigna, AFRL

Panel Discussion 1: Challenges of Multilingual Processing – Scaling and Tailoring for Global Missions

- Bootstrapping Translation Capabilities – Ray Slyh, AFRL
- Translation for High Demand Languages – Kathleen Egan, TSWG

1200 – 1300 **Lunch (for those interested, Peter Chew (Sandia) will be demonstrating multilingual text characterization during lunch).**

1300 **Session 2: "Yours, Mine and Ours": Operationalizing the JIPOE Nexus**

- Understanding the Problem – Lessons Learned – Sue Numrich and Carl Hunt,, IDA
- Industry Perspectives – Stan Horky, GM
- Art of the Possible: Functional DOD Federation -- Dr. Amy Henninger, IDA
- Effective Team Composition: PLG
- Nexus General Forum Session and Wrap-up

1600 **End Workshop**

APPENDIX B: WORKSHOP ATTENDEES

Name	Organization
Adkins, Mark L.	GISC
Allwein, Kelcy	DIA
Atkinson, George	SAS
Bailey, Joe	JIOWC
Bemish, Nick	DIA
Benson, John	AFRL
Cabayan, Hriar	OSD
Canna, Sarah	NSI
Chadha, Bipin	NuTech
Chang, Remco	UNCC
Chen, Hsinchhun	U Arizona
Chesser, Nancy	DTI
Chew, Peter	Sandia
Chinchor, Nancy	OSC
Clark, Mark	SAIC
Colbath, Sean	BBN
Davis, Mike	Ontology Works
Dietrich, Paul	DARPA
Dobbins, Roberta	OSC
Egan, Kathleen	TSWG
Eick, Steve	SSS Research
Fenstermacher, Laurie	AFRL
Fuller, Doug	SSA
Hensley, Brian	JIOWC
Henze, Roger	GISC
Hirsch, John	INSCOM
Horky, Stan	GM
Hunt, Carl	IDA
Inchiosa, Mario	NuTech
Ingram, Greg	NSI
Jackson, Gary	SAIC
Jakubek, David	OSD
Kargar, Hastie	DARPA
Kuznar, Larry	NSI
Lasky, Josh	JS J39
Leonard, Hans	U. of GA
Marcus, Sherry	21st Century
Marquardt, Ed	NSA
Mather, David	ATSID
Meeker, David	INSCOM
Neely, Mike	SPADAC
Newland, Marcus	U. of GA
Numrich, Sue	IDA
O'Connor, Jennifer	DHS
Olive, Joe	DARPA
Pate, Amy	UMD/START
Patterson, John	Detica
Payne, Tom	STRATCOM
Popp, Bob	NSI
Porter, David	Detica
Rhem, Sam	SRC
Richardson, Russell	SETA
Riley, Ben	OSD
Ross, Travis	Indiana U
Roukos, Salim	IBM
Russell, Anne	SAIC
Sanfilippo, Antonio	PNNL
Shaneyfelt, Wendy	Sandia
Shannahan, Mike	GISC
Shellman, Steve	SAE
Silman, Mort	
Simon, David	ATS
Slepoy, Alexander	DOE
Slyh, Ray	AFRL
Snead, Ned	IDA
Speed, Ann	Sandia
Srihari, Rohini	Janya
Stone, Robalyn	SSA
Strasburg, Jana	DOE
Toman, Pamela	NSI
Urgo, Marisa Ann	ARGUS
Warrender, Christy	Sandia
Washington, Tyrone	SRC
Wilson, Richard	OSD
Zappavigna, Mark	AFRL
Zimbra, David	U of Arizona