

# Text analysis and entity extraction in asymmetric threat response and prediction

Erwin Chan<sup>1</sup>, Jason Ginsburg<sup>2</sup>, Brian Ten Eyck<sup>3</sup>, Jerzy Rozenblit<sup>3</sup>, Mike Dameron<sup>4</sup>

<sup>1</sup>Dept. of Linguistics, University of Arizona, USA, echan3@u.arizona.edu

<sup>2</sup>Center for Language Research, University of Aizu, Japan, jginsbur@gmail.com

<sup>3</sup>Dept. of Electrical and Computer Engineering, University of Arizona, USA,  
bteneyck@ece.arizona.edu, jr@ece.arizona.edu

<sup>4</sup>Ephibian, Tucson, AZ, USA, mdameron@ephibian.com

**Abstract**—The **Asymmetric Threat Response and Analysis Program (ATRAP)** is a software system for intelligence fusion, visualization, reasoning, and prediction. ATRAP consists of a set of tools for annotating and automatically extracting entities and relationships from documents, visualizing this information in relational, geographic, and temporal dimensions, and determining future courses of action of adversaries by creating situational threat templates and applying customized prediction algorithms. In this paper, we first describe the task of analyzing data in intelligence reports, and then provide an overview of major system components: the Text Highlighter tool, the ThoughtSpace™ visualization environment, and the Template Builder and prediction tool. Subsequently, we describe linguistic characteristics of intelligence reports, and describe ATRAP's named entity recognition system.

*Keywords*—intelligence analysis, entity extraction

## I. INTRODUCTION

An important topic of research for military intelligence production is in developing systems that assist in the analysis of reports on the activities of adversarial agents, who are engaged in asymmetric warfare and whose actions are constantly changing. The ability to collect, analyze, model, and share entity objects and relationships in the battlespace is a key to minimizing the need for combat operations and subsequent collateral damage. This capability has been identified as mission critical to overcoming insurgencies in areas of U.S. military operation such as Afghanistan, Pakistan, and Iraq, and even on the U.S. border with Mexico. Computer systems have the potential to help intelligence analysts in sharing information, identifying entities and events of interest, visualizing entity relationships and activity patterns, and predicting future courses of action. These types of systems also find application in other types of intelligence-gathering domains, such as criminal investigations, corporate fraud detection, financial markets, and other social networks.

The Asymmetric Threat Response and Analysis Program (ATRAP) is an advanced data fusion and visualization environment that is intended to augment commanders' situational awareness and decision-making capabilities. ATRAP is comprised of a suite of software tools that in the aggregate is able to ingest raw intelligence messages from multiple sources, employ advanced unstructured text

processing and correlation capabilities to generate link and pattern analysis, and display that information in a 3-dimensional, geotemporal visualization environment. It has also been imbued with unique prediction algorithms that take into account both activity-based threat models and specific organizational behavior models in order to predict enemy courses of action. All of this is delivered within a fully-transparent analytic path in a collaborative application.

ATRAP is not intended to replace the intelligence analyst, but rather to serve as a “cognitive amplifier” for the trained analyst. Mundane tasks such as information organization and entity extraction are automated or facilitated, and analysts' attention is focused on higher-level activities such as evaluating the validity and plausibility of codified tactics, techniques, and procedures. ATRAP includes pattern recognition components involving entity extraction and prediction of enemy courses of action. Any suggestions made by the system, however, must be verified by a human user. This consideration plays an important role in the design of the system.

In the rest of this paper, we describe ATRAP, and focus on the nature of the input data and initial efforts in named entity recognition. We first describe the intelligence gathering process in more detail and survey some relevant work in computer-aided systems for intelligence analysis. Next, we provide an overview of ATRAP and its components for data annotation, visualization, and prediction. Then, we focus on the language of intelligence reports, and initial research in automatic entity recognition. Finally, we discuss current and future work.

## II. INTELLIGENCE ANALYSIS

In this field, analysts receive myriad human-generated and sensor-based reports on the circumstances of a conflict or potentially dangerous situation. Reports and particular actions described within them may be time-stamped. Documents from newswire or broadcast news may also be examined.

With the information contained in these reports, there are several tasks for the intelligence analyst. One is to identify entities of interest (such as persons, organizations, events, and their relationships) and to study networks of associations between entities. Due to the dynamic characteristics of asymmetric adversaries, it is also important to take into

---

This research has been supported by the U.S. Army Intelligence Battle Lab. The technical agent for this work was Fort Huachuca under Task Order Numbers 9T7ZDAIS705, 9T8ZDAIS803, and 9Q9SDAIS903. The opinions expressed in this document are solely those of the authors, and should not be viewed as representing policies of the U.S. Army Intelligence Battle Lab or the United States government.

account the temporal and geographic dimensions of this type of intelligence analysis. Some of the common techniques for visualizing information include link diagrams that describe relationships between entities, matrices describing activities of entities, and flow charts summarizing how events occur over time. These types of information may be overlaid on maps to incorporate geographic, topographic, or even meteorologic information into the analysis.

Crucial (and often perishable) information regarding high value targets and their disposition, location, and timelines are typically analyzed from an historical perspective; there is rarely the time or appropriate tools for predictive analysis. In ATRAP, however, information that has been collected is used to help define and ultimately anticipate possible adversary courses of action. Specifically, in military intelligence terminology, an analyst is looking for a sequence of *Indicators*, or pieces of information, that fit a *situational template* for accomplishing some objective. *Specific Information Requests* (SIRs) indicate what particular entity details are needed, while *Specific Orders or Requests* (SORs) indicate when and where to look for specific entities. The intelligence that is collected from SIRs may satisfy Indicators. For example, Figure 1<sup>1</sup> shows a template for the sequence of actions of foreign fighters coming to Iraq. In the first step, SORs for data collectors fulfill an SIR for finding the group membership and location of individuals, which together fulfill the Indicator of being a foreign fighter, with respect to Iraq. Subsequent indicators are looking for evidence of fighters entering Iraq, and then either becoming combatants or suicide bombers. If the initial Indicators in a template are satisfied, subsequent events, as described by other Indicators in the template, may be predicted to occur in the future.

### III. PREVIOUS WORK

Software systems can provide assistance for intelligence analysts, who must deal with large quantities of raw data. In relation to our work, there have been two relevant areas of previous research. One has been in research programs for automated information extraction techniques, such as the Message Understanding Conferences (MUC) [1, 4]; the CoNLL-2003 shared task on entity recognition [13], and the National Institute of Standards program on Automatic Content Evaluation (ACE) [12].

Another area of relevant research is in visualizing relationships in documents. For example, the TexPlorer system [10] performs extraction of named entities and entity relationships across multiple documents. Information can be visualized through cluster diagrams, tables, and maps. Conceptual link diagrams are also available through an interface with the ConceptVista system [2]. Another project, more closely related to our own, is the ZENON system for extracting entities and actions from text, and visualization

through link diagrams [5, 6, 7]. This system was developed for application to human intelligence reports for the German Federal Armed Forces.

ATRAP makes advancements beyond previous work in Natural Language Processing and data visualization in several ways. First, it has rich visualization capabilities that combine relational, geographic, and temporal information in an interactive 3-dimensional environment. Second, it allows the analyst to indicate entities, relationships, and matching indicators within the different visualization modes. Third, the ultimate task of predicting adversary courses of actions sets up expanded problem definitions for pattern recognition, encompassing problems of automated information extraction and fusion, threat modeling, and behavior modeling. Finally, it stores all analytic information and artifacts, making them available to other analysts and for future use.

A precursor to ATRAP was Pathfinder [8], a system created in the mid-1990's by the National Ground Intelligence Center. Pathfinder was developed to allow intelligence analysts to automatically query databases and collections of unstructured text documents, and to indicate and visualize data relationships. Now owned by Science International Applications Corporation (SAIC), Pathfinder is a commercial software system consisting of a large suite of tools for information processing and analysis, that is applicable to many different domains.

### IV. ATRAP'S CORE COMPONENTS

ATRAP is optimized to run in a 3-tier, client-server architecture, although it is also capable of operating as a standalone application. Documents and metadata are ingested into a central database, and analysts use ATRAP's tools to detect and visualize entity relationships and activity patterns in the data. The system's core functional components are briefly described below.

1. The *Text Highlighter Tool* enables users to highlight entities in text and add relevant metadata. Entities are assigned to one of six base entity types: Person, Location, Organization, Equipment, Event, and Undefined. Beneath these base entity types, ATRAP's ontological structure is extensible to any depth. The tool also includes an option for automatic entity recognition (to be discussed in more detail in Section VI). Entities suggested by the recognition algorithm must be confirmed by the user before being added to the database. The tool utilizes intelligent word boundaries, so the user can highlight a portion of a word and the tool will automatically expand to include the entire word. In addition to adding entities, the tool enables the user to associate attributes in the report to an entity. Finally, the tool provides a "quick add" mode, which enables the user to pre-define the entity's type, location, and reliability score. In this mode, the user can quickly highlight a list of people in a report without

---

<sup>1</sup> Screenshots are based upon the Sinjar records, a set of documents describing the entry of foreign fighters into Iraq.

having to classify each one separately. A screenshot of this tool is shown in Figure 2.

2. The *entity dossier* facilitates entering, viewing, and editing information about a selected entity, including its name and type, known locations and aliases, associated media files (e.g., pictures, videos), additional user-defined attributes and values, and any analyst comments. The dossier also provides a link chart displaying relationships to other entities, a map displaying geospatial locations, and an analytical time wheel, as well as a link to all source documents referencing the entity.
3. *Information Search* enables searching through all intelligence reports that have been ingested into ATRAP based on either keywords or specific entities of interest. The “visual search” function enables users to draw an area of interest on a specific map; optionally designate a date, time, and/or specific entity types of interest; and generate a ThoughtSpace using these parameters as the search criteria. “Relationship search” enables users to select any number of entities, and lets ATRAP automatically find and display any or all relationships between those entities.
4. The *ThoughtSpace™ visualization environment* provides an interactive 3D workspace for visualizing links and patterns between entities (all rendered based on Military Symbolology Standard 2525C). After adding entities of interest (or all entities within a specified area of interest) into a ThoughtSpace, users can search for locations by name or latitude/longitude (or MGRS or UTM), and the map will automatically center on the specified location. From there, users can continuously pan, zoom, and rotate the entire dataset, which renders entities and relationships based on geotemporal metadata, as well as the manner in which they support specific predicted courses of action. Maps may be loaded either from a library or from online map providers. ThoughtSpace also offers the ability to view time slices (set by the user) within the geotemporal dataset, allowing the user to pan through time or “play” the dataset and view the dataset on a timeline histogram.
5. The *Template Builder Tool* enables analysts to model and graphically represent enemy tactics, techniques, and procedures through a sequence of indicators. Template objects can be rendered within ThoughtSpace in order to visualize satisfied indicators and potential future courses of action.
6. ATRAP’s *predictive model* processes non-numerical data to arrive at automated assessment and confidence scores for these Templates. Prediction algorithms, described in [14], match information in the database to Indicators and attendant SIRs and SORs in situational templates to identify which enemy courses of actions may be currently active. The predictive model is traceable, transparent, and utilizes Human-in-the-Loop data fusion.

## V. ATRAP DATA AND NAMED ENTITY ANNOTATION

### A. Data Sets

The data sets that we were supplied with by our sponsors to help develop ATRAP consisted of corpora of English language text documents, totalling approximately 250,000 words. These (unclassified) documents are representative of actual intelligence reports, and other information sources used in intelligence analysis. For example, one of these corpora describes a foreign conflict and is used in training intelligence analysts. Documents may be concerned with a range of different topics, such as an overview of a conflict, descriptions of enemy actions, summaries of configurations of military units and their operations, etc.

### B. Linguistic Characteristics

Intelligence reports often display a writing style that is atypical of written news, broadcast speech, and other more commonly encountered genres of language. Some of the interesting characteristics of this data are listed below:

- While some documents may follow standard conventions for written English, many others may be in all uppercase letters, or a mixture of lowercase and uppercase. Words may be written in uppercase for extra emphasis. The following two sentences exemplify this: “We will Attack to Defeat enemy forces at OBJ HECATE and OBJ MERCURY”, and “Their base of operations is Moscow and they have a stronghold on most of the Bars, Casinos, and Nightclubs in the area.”
- Sentences may describe actions of military units, and frequently contain abbreviations. For example, “A/3 ENG CO is OPCON to 1st AR BN,” and “22 SIGNAL BN Commander is V Corps G6 and the V Corps Systems Control Element (SYSCON).”
- Entities are sometimes clearly indicated in text through particular stylistic conventions, For example, “JENAN, Arman” indicates a person, and “MOSCOW, Donovia” indicates a location.
- Entities and other specific types of information may be indicated through field names prior to the main body of a document. Field names include “Name”, “Suspect”, “Location”, “Date”, “Time”, and others. An example is “NAME Abbasavo, Adil.”
- Sentences can consist of long, comma-separated lists of entities, as in “Intelligence Cell Personnel: Rupen Melidonian, Nareg Saatjian, Madteos Abassian, Yermi Hagpoian, Khacho Elmassakian, Gosdant Kemikisizian.”

### C. Entity Definitions and Annotation

In developing an entity extraction system for ATRAP, it was necessary to define the types of strings that would be useful for the intelligence analyst. For the current stage of the project, we were advised to identify strings that directly refer

to entities by name (instead of more general referential noun phrases, such as “the group” or “they”). The entity types under consideration were Person, Location, Organization, and Unknown type, with an emphasis on identifying Persons.

We first considered utilizing a corpus already annotated for named entities, namely the Automatic Content Extraction (ACE) corpus [9]. Several problems were apparent, due to the wide range of linguistic constructs that ACE defines as entities. For example, consider the sentence fragment “Can [you guys]<sub>PER</sub> call [me]<sub>PER</sub> [Monica]<sub>PER</sub>?” [she]<sub>PER</sub> suddenly asked [the grand jury]<sub>PER</sub> ...”. In this sentence, only the name “Monica” would be identified by the intelligence analyst, while pronouns and referring noun phrases would not be desired. If these constructs were indeed proposed as entities, it could lead to an overload in the amount of information presented to the analyst in a document. ACE also allows for very long noun phrases as Persons, such as the string “the nearly 35 million babies born in this country from 1983 to 1991.”

Instead, for Person entities, we have chosen to extract names of persons along with titles, such as “Jon Postel,” “Dr. Arnold Levine,” and “Commerce Secretary Ronald H. Brown.” A Linguistics graduate student annotated the corpora for Persons. These annotations have been double-checked, and initial annotations have also been created for Organization and Location. In total, 3,948 Persons, 5,298 Locations, and 5,742 Organizations were identified in the ATRAP corpora.

## VI. NAMED ENTITY RECOGNITION

Named Entity Recognition (NER) is the task of discovering string sequences in text that refer to Persons, Locations, Organizations, and other types of entities. Some examples of entities extracted from ATRAP data are the following: for Person, “Dr. Dungwat Seit”, “Mohammad al-Fahid”, “Vincente Rojas-Moreno”, “Jelena Shecherbakov”, “DANIEL J. CALLAGAN”; for Organizations, “Democratic Front for the Liberation”, “Donovia Gangs”, “New People’s Army”, “Taliban government”, “Irish Republican Army”; for Locations, “US Embassy”, “Aleppo, Syria”, “Sonoran Desert”, “810 13TH STREET”, “Fresno County Hospital”.

### A. Techniques Used

Due to the lack of a large, suitably annotated, and domain-appropriate corpus at the beginning of the project, our initial system for entity recognition largely consisted of pattern matching through regular expressions, rather than through machine learning. Our goal was to develop a high-precision system, selectively using regular expressions to conservatively make suggestions about what entities might be in a text. This was important for our goal of providing assistance to analysts, rather than having them spend large amounts of time post-editing system output. The Person extraction system is described in detail in [3].

For Person recognition, we searched for field names (such as “NAME:”), titles (*President, Mr.*, etc.), highly-frequent preceding word n-grams (*named, according to*, etc.), and combinations of first name, middle initial, and/or last name. Lists of names from the U.S. Census were also utilized.

For Organization recognition, we searched for field names (“Affiliation:”, “ORG:”, etc.) and phrases including words that denote organizations (*committee, organization, cartel*, etc.), perhaps beginning with locational modifiers, which are signalled by words ending in *-ian, -an, -ese, -i, or -ic*.

For Location recognition, we searched for location field names (“LOC:”), address patterns, combinations of city and state or city and country, proper names preceded by directional modifiers (*West, Southeastern*), phrases preceded by locational titles (*Fort, Downtown*), phrases following verbs of travelling or attacking a position, nouns with locational complements (*vicinity of, people of*), lists of building types and landmarks, and membership in predefined lists of locations, with common English words filtered out.

In addition to code specific for Person, Location, and Organization, other entities were found by formulating regular expressions to match the writing style: first word in all-caps, second word with the first letter capitalized (for cases such as *MOSCOW, Donovia*), and also long, comma-separated lists of strings. Potential foreign names were detected through a statistical system that looks for character trigrams that are uncommon in English.

### B. Evaluation of Named Entity Recognition

For a quantitative evaluation, the named entity recognizer was tested on the ATRAP datasets and the MUC-3<sup>2</sup> data. It was also compared with a statistical named entity recognizer<sup>3</sup>, whose implementation was based on the Sparse Network of Winnows (SNoW), a modern machine learning algorithm [11]. This system was trained on the CoNLL-2003 data [13], which consisted of news articles annotated for named entities.

Performance was quantified through the standard measures of precision, recall, and F-measure. For a particular entity type, *precision* is defined as the number of correctly identified entities, divided by the number of entities proposed by the system. *Recall* is the number of correctly identified entities, divided by the total number of entities in the gold standard. *F-measure* is the harmonic mean of precision and recall.

Table I shows the performance of the ATRAP and SNoW named entity recognizers on the ATRAP data, for Person entities. It can be seen that the ATRAP recognizer outperforms the SNoW-based system. Evaluations for other entity types are not shown, as since the correctness of these annotations is still being verified.

<sup>2</sup> [http://www-nlpir.nist.gov/related\\_projects/muc/](http://www-nlpir.nist.gov/related_projects/muc/)

<sup>3</sup> <http://l2r.cs.uiuc.edu/~cogcomp/software.php>

Table II shows the performance of the two systems on the MUC-3 data, a corpus of news articles concerning Latin American terrorism. To produce a gold standard, values were extracted from the following fields of the MUC answer keys: for Persons, PERP: INDIVIDUAL ID, HUM TGT: TYPE, and HUM TGT: DESCRIPTION; for Organizations, PERP: ORGANIZATION ID; and for Locations, INCIDENT: LOCATION, PHYS TGT: FOREIGN NATION, and HUM TGT: FOREIGN NATION. We have added scores for Person entities with non-proper names manually removed, since these are not intended to be extracted within ATRAP (see Section V-C). Precision scores are artificially low because the articles are not fully annotated; both systems find many named entities that are not listed in the MUC answer keys. The recall of ATRAP is somewhat degraded compared to SNoW, but this is to be expected due to differences in language genre: the ATRAP system was developed primarily for military intelligence reports, while the SNoW-based system was trained on news articles, the same type of text as the MUC data. The ATRAP recognizer nevertheless outperforms the SNoW-based system in F-measure in most cases. Strings found to be entities by ATRAP, but of unknown entity type, have not been included in these calculations.

TABLE I. PERFORMANCE OF ATRAP AND SNoW PERSON ENTITY RECOGNIZERS ON ATRAP DATA

	Precision	Recall	F-measure
ATRAP	0.843	0.663	0.743
SNoW	0.346	0.489	0.405

TABLE II. PERFORMANCE OF ATRAP AND SNoW NAMED ENTITY RECOGNIZERS ON MUC-3 DATA

		Prec.	Rec.	F-meas.
ATRAP	PER w/ filtering	0.893	0.057	0.107
	PER w/o filtering	0.739	0.127	0.217
	ORG	0.878	0.081	0.148
	LOC	0.911	0.035	0.067
SNoW	PER w/ filtering	0.988	0.050	0.095
	PER w/o filtering	0.945	0.130	0.229
	ORG	0.922	0.068	0.127
	LOC	1.000	0.030	0.058

### VII. CURRENT AND FUTURE WORK

There are several areas, both practical and theoretical, in which we are currently working to improve ATRAP:

- *Improved information extraction.* We are exploring more sophisticated techniques for entity, event, and relation detection and resolution from text, involving

machine learning, parsing, and lexical semantics. Additionally, we would like to apply algorithms in social network analysis to see if partial information manually provided by the analyst can be used to automatically detect additional entity relationships. Further efforts in corpus annotation are planned.

- *System testing.* To facilitate user adoption, we are currently conducting usability tests on ATRAP with Army intelligence analysts, examining how well the system assists them with data entry, information access, and prediction. Additionally, ATRAP will be applied to large extant intelligence databases, so that its performance may be profiled at scale, and its predictive abilities may be refined.
- *System architecture developments.* ATRAP’s hardware configuration was designed to be deployable within the United States military operational systems. Efforts are underway to adapt ATRAP to interface with other intelligence gathering and data processing systems.
- *Behavioral modeling.* Actions taken by adversaries partially depend upon behavioral characteristics, such as their culture, organization membership, technical capabilities, geographic reach, and other factors. Future versions of the prediction model will include a behavioral modeling component that will adjust scores assigned to predicted courses of actions based on knowledge of these relevant characteristics.

### VIII. CONCLUSION

ATRAP is a software-based “cognitive amplifier” that enables intelligence analysts to rapidly identify and predict enemy actions within a particular theatre or operational environment. It represents an integrated system of software tools for assisting the analysis of text-based intelligence reports through a unique combination of automated entity extraction, document highlighting and annotation, situational template creation and prediction, and three-dimensional geotemporal visual displays of entity relationships. This combination of system features also provides a unique, real-world testbed for the application of information extraction and pattern recognition techniques.

### ACKNOWLEDGMENTS

We are grateful to Ephibian, Inc. for their software engineering assistance, and to Neil Garra of the S2 Company for his subject matter expertise.

### REFERENCES

- [1] N. A. Chinchor, D. D. Lewis, and L. Hirschman, “Evaluating message understanding systems: an analysis of the Third Message Understanding Conference (MUC-3)”, *Computational Linguistics*, 19, 409-449, 1993.
- [2] M. Gahegan, R. Agrawal, A. R. Jaiswal, and K. Soon, “Measures of similarity for integrating conceptual geographical knowledge: some ideas and questions”, *Proceedings of the Conference on Spatial Information Theory (COSIT)*, 2007.
- [3] J. Ginsburg. “A person extractor”, internship report, Human Language Technology Program, Dept. of Linguistics, University of Arizona, 2009.

[4] R. Grishman and B. Sundheim, "Message Understanding Conference - 6: a brief history", Proceedings of the International Conference on Computational Linguistics, 1996.

[5] M. Hecking, "Navigation through the meaning space of HUMINT reports", Proceedings of the 11th International Command and Control Research and Technology Symposium, 2004.

[6] M. Hecking, "System ZENON - semantic analysis of intelligence reports", Proceedings of LangTech, 2008.

[7] C. Jenge, S. Kawaletz, and U. Schade, "Combining different NLP methods for HUMINT report analysis", Proceedings of Information Management and Exploitation, NATO Research and Technology Organisation, RTO-MP-IST-087-11, 2009.

[8] M. G. Knapp and T. B. Hendrickson, "Project Pathfinder: breaking the barriers to more effective intelligence analysis", Military Intelligence Professional Bulletin, January-March, 1996.

[9] A. Mitchell, S. Strassel, et. al., "ACE-2 Version 1.0", LDC2003T11, Philadelphia, PA: Linguistic Data Consortium, 2003.

[10] C. Pan, A. R. Jaiswal, J. Luo, and A. Robinson, "TexPlorer: an application supporting text analysis", Proc. of IEEE VAST Contest, at the Visual Analytics Science & Technology Symposium, 2007.

[11] D. Roth, "Learning in natural language", Proc. of the International Joint Conference on Artificial Intelligence, pp. 898-904, 1999.

[12] S. Strassel, M. Przybocki, K. Peterson, Z. Song, and K. Maeda, "Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction", Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.

[13] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition.", Proceedings of the Conference on Natural Language Learning, 2003.

[14] M. Valenzuela, "A non-numerical predictive model for asymmetric analysis", Master's Thesis, Department of Electrical and Computer Engineering, University of Arizona, 2010.

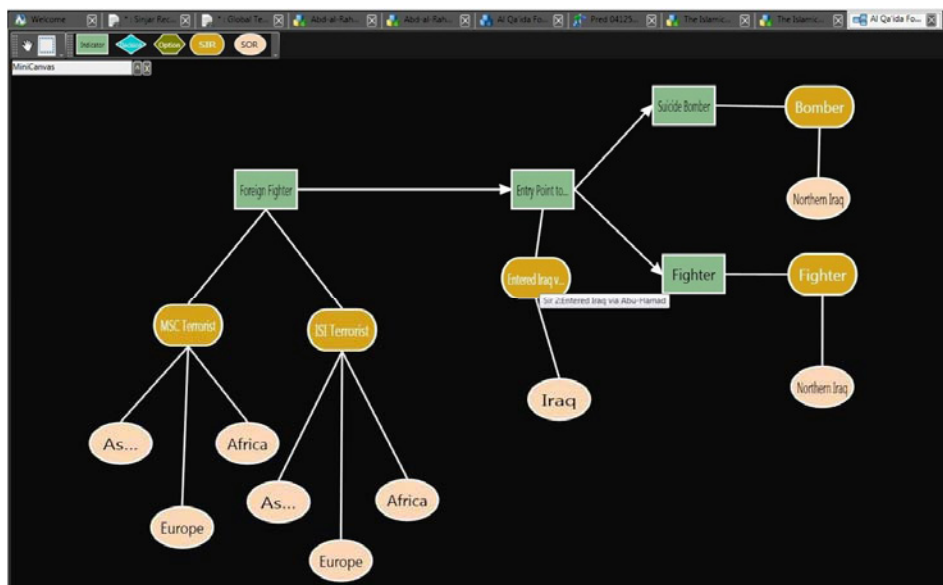


Figure 1. Template Builder.

Figure 2. Text Highlighter Tool.