Systemic Test and Evaluation of a Hard+Soft Information Fusion Framework

Challenges and Current Approaches

Geoff A. Gross[a], Ketan Date[c], Daniel R. Schlegel[ab], Jason J. Corso[ab], James Llinas[a], Rakesh Nagi[c], Stuart C. Shapiro[ab]
[a] Center for Multisource Information Fusion (CMIF), State University of New York at Buffalo, Buffalo, New York, U.S.A.
{gagross, jcorso, llinas} @buffalo.edu
[b] Department of Computer Science and Engineering, Center for Cognitive Science, State University of New York at Buffalo, Buffalo, New York, U.S.A.
{drschleg, shapiro} @buffalo.edu
[c] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, U.S.A.
{date2, nagi} @illinois.edu

***Abstract*— The area of hard+soft fusion is a relatively new topic within the information fusion community. One research effort which has confronted the subject of hard+soft fusion is the Multi-disciplinary University Research Initiative (MURI) titled "Unified Research on Network-based Hard+Soft Information Fusion". Developed on this program is a fully integrated research prototype hard+soft fusion system in which raw hard and soft data are processed through hard sensor processing algorithms, natural language understanding processes, common referencing, alignment, association and situation assessment fusion processes. The MURI program is currently in its 5th (and last) year. During years 1 through 4, the MURI team dealt with the research issues in developing a baseline hard+soft fusion system, while identifying a number of design alternatives for each of the framework processing elements. For example, within natural language understanding different stemmers or ontologies could be utilized. The mathematical nature of hard or physical sensor processing and data association involved design choices about numerous parameters which affect the solution quality and solution quality/runtime tradeoff. While traditional experimental or training approaches may be used in assessing these processes in isolation, the nature and dependencies of hard+soft fusion require a systemic approach in which the integrated performance of framework components are understood. In this paper we describe the design of a test and evaluation framework for systemic error trail analysis and parametric optimization of hard+soft fusion framework sub-processes. We will discuss the performance metrics utilized including notions of "system optimality," issues in defining the parametric space (design variants), cross-process error tracking methodologies and discuss some initial results. The presented system results are based on the Synthetic Counterinsurgency (SYNCOIN) dataset which is a dataset developed within the program and utilized for training and system optimization. Future work, including plans for the validation of experimental results will also be discussed.**

Keywords *— hard+soft information fusion, system test and evaluation, system under test, evaluation metrics, error audit trail*

I. INTRODUCTION

A topic which has recently received much attention within the information fusion domain is the topic of Hard+Soft information fusion. Hard+Soft information fusion considers both hard, physical sensor (e.g., radar, acoustic, etc.) and soft, linguistic (e.g., human reports, Twitter feeds, etc.) data sources. Many modern domains both in the military and private industry settings (e.g., counterinsurgency [1],[2], disaster relief [3],[4], consumer marketing [5],[6], etc.) have come to recognize the importance of the fusion of numerous data sources, broadly including both hard and soft data. One research effort which has confronted the subject of hard+soft fusion is the Multi-disciplinary University Research Initiative (MURI) on Network-based Hard+Soft Information Fusion [7].

The MURI program in Hard+Soft Information Fusion has developed a fully integrated hard+soft fusion research prototype system in which raw hard and soft data are processed through hard sensor processing algorithms (e.g., detection and tracking), natural language understanding processes, common referencing, alignment, association and situation assessment fusion processes. The MURI program is currently in its 5th year. During years 1 through 4, the MURI team dealt with research issues in developing a baseline hard+soft fusion system, while identifying a number of design alternatives for each of the framework processing elements. A recent focus (to continue through program completion) is in the systemic test and evaluation (T&E) of the developed hard+soft information fusion framework.

While traditional experimental or training approaches may be used in assessing processes of a hard+soft information fusion framework in isolation, the nature of dependencies across framework components requires a systemic approach in which the cross-component affects are understood. Past efforts in the T&E of hard, soft and hard+soft information fusion systems have largely focused on the evaluation of situational awareness of the human or machine consumer of system output (e.g., [8], [9], [10], [11]). While this assessment is an important measure of system effectiveness,[1] these past studies generally do not include assessments of sub-process performance and its effect on overall system performance (i.e., producing an error audit trail). In this paper we describe the design of a metric-centric test and evaluation framework for systemic error trail analysis and parametric optimization of hard+soft fusion framework sub-processes. We will discuss the performance metrics utilized including notions of "system optimality,"

[1] Although "situational awareness" provides a measure of the degree to which the system supports user understanding, many systems require further support, and an assessment of the degree to which the system facilitates action on this obtained understanding. Not much work toward this higher level objective exists within the literature and this topic is noted for a direction of future work.

issues in defining the parametric space (design variants), cross-process error tracking methodologies and discuss some initial results.

The remainder of this paper is structured as follows: Section II defines the exemplar system under test and Section III describes issues in defining metrics and the parametric space (or system variants) to be considered within T&E. Section's IV-VIII provide an overview of the framework processes within the exemplar system under test and provide both individual process and cross-process evaluation metrics. Specifically, Section IV introduces one physical sensor processing element within the MURI framework (as an exemplar of T&E approaches for these hard data processes), Section V presents an overview of the natural language understanding evaluation methodology, Section VI describes the system benefit of the common referencing process (uncertainty alignment), Section VII explains the evaluation of the data association process (readers are directed to [12] for a more detailed description) and Section VIII identifies a variety of graph analytic techniques which are applied on the cumulative associated data to enable situation assessments. Finally, Section IX discusses some initial T&E results across these framework processes and plans for future work and Section X provides conclusions.

II. System Under Test (SUT)

A necessity when performing system T&E is the definition of a System Under Test (SUT), which is the set of functional components and connections to be evaluated. The definition of the SUT must consider the larger project schedule beyond the T&E efforts. For example, continuing research and development (R&D) work during the T&E period may make the SUT a moving target. A decision may need to be made whether to freeze the SUT or allow for the continuing evolution of framework processes (see Section III for additional thoughts on tracking SUT performance through R&D iterations). Particularly if R&D efforts are to continue throughout the T&E period, version control and version logging must be carefully followed such that results and process settings of any test run may be replicated.

While the methods and metrics developed in this paper are fairly general, we will consider specific applications to the system architecture developed within the MURI project [1] (see Figure 1). Within the MURI framework, hard (or physical sensor) input data enters the hard sensor fusion and track creation processes which convert the raw sensor data (video, acoustic, etc.) into semantic tracks, containing the entity and attribute evolution over the duration of the data and some interaction events. Evaluation of the hard sensor fusion processes is described in Section IV.

Soft (or linguistic) input data within the MURI framework enters the Tractor Natural Language Understanding (NLU) process which performs processes including: dependency parsing, within-source co-reference resolution, named entity identification, morphological analysis to find token root form, context-based information retrieval and syntax-semantics mapping. The resulting propositional graph from Tractor is ideally fully semantic content (versus syntactic), containing all of the semantic propositions which would be identified by a human interpreter of the original message. Evaluation of this capability is described in Section V.

After a conversion from a propositional graph to attributed graph, the soft data stream is run through a common referencing and uncertainty alignment process. This process seeks to account for observational biases and variances in human observation, accounted for by contextually-based human error models, developed within this program. Evaluation of the benefit of this process to the fusion tasks of data association and situation assessment is described in Section VI.

Next, the hard and soft data streams enter the data association process. Data association algorithmically identifies common entities, events and relationships across data sources and data modalities, associating the entities, attributes, and relationships based on computed similarity criteria. The objective of data association is to form a single node for each unique entity or event or a single edge for each unique relationship within the cumulative data (see Section VII).

Upon the formation of a cumulative, fused body of evidence (the cumulative associated data graph), analyst-guided graph analytic processes reason over this data in an attempt to obtain and maintain situational estimates. Some graph analytic processes which were developed under the MURI effort (along with initial evaluation considerations) are described in Section VIII.

III. Defining Metrics and the Test Space

With the SUT defined, a determination of evaluation points within the SUT must be made. The evaluation points within the MURI SUT are separable along process lines including: physical sensor processing, natural language understanding, data association and graph analytic processes (situation assessment) as shown in Figure 1. For each of these processes we define evaluation metrics which are expected to be reflective of overarching system performance. Potential performance metrics are broadly classified as quality and runtime-based metrics, with the simultaneous optimization of both typically resulting in a conflicting objective. Depending on the operational environment, solution quality or runtime may be at a premium. Due to the basic research nature of our program and lack of a specific target data environment, our focus was on quality-based metrics.

While the physical sensor and natural language understanding processes operate on raw data which is expected to be factually correct,[2] downstream processes of data association and graph analytics may be subject to upstream errors. As a result, these downstream processes must consider the notion of both process and cumulative system optimality. The performance metrics for each process are described in detail within Section's IV-VIII.

[2] We assume the hard and soft data streams contain factual information not resulting from intentional attempts to deceive. While we understand these data (in particular soft data) may be subject to contradictions, inconsistencies or deception, the resolution of these elements was not a focus or expectation of the MURI program.

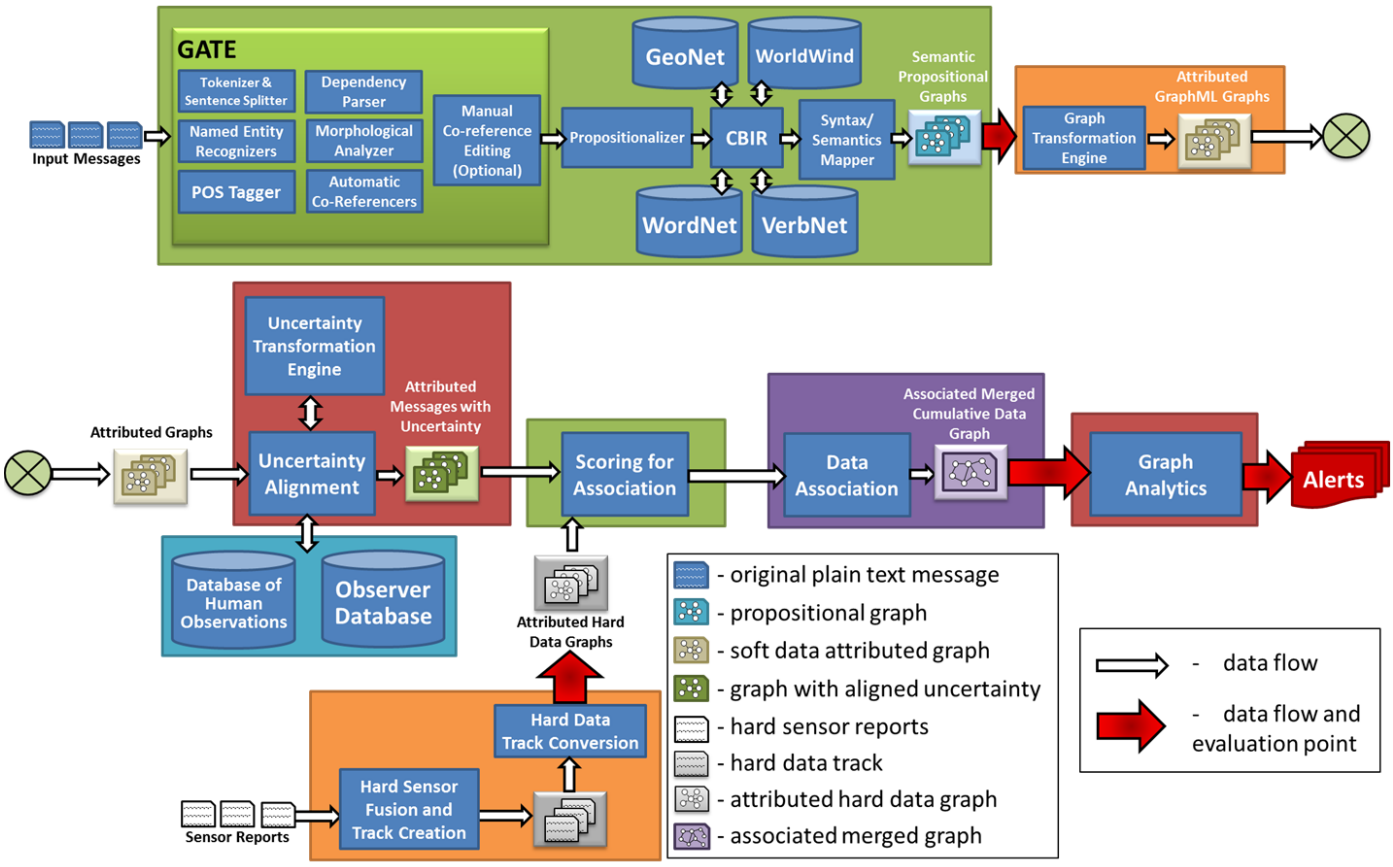

Figure 1 – MURI System Under Test.

In addition to identifying a system configuration resulting in "system performance optimality," another overarching goal of system T&E is to measure the main effects and interactions of design alternatives on both process and system-level performance metrics. While the number of design alternatives which could be considered is theoretically infinite (e.g., numerical parameters), some pilot study or process expert guidance may be used to prune the potential training and evaluation space. In addition to utilizing identified performance metrics as a basis for spiral (incremental) system development, a number of experimentation questions were developed, thus defining a *test space* for experimentation.

In addition to process parameters, elements of the test space include input data qualities. A natural interest in the nascent area of hard+soft information fusion is the quantification of the value of hard versus soft versus hard+soft information to some system level objective (e.g., to situational awareness performance measures). An additional input data interest within the test space is the robustness of processes to varied levels of input data quality, whether raw data or machine processed. The assessment of situational awareness metrics after the graph analytic processes in our SUT remains as future work (see Section VIII).

In addition to the optimization of each of the many process parameters, a sampling of process variation questions to be assessed via the T&E processes described subsequently are as follows:

1. How general are each of the processes to variations in input data? What are the input data qualities which affect system performance?
2. What is the effect of alternate stemmers within the NLU process?
3. How do different ontologies used within NLU processing (and downstream processes) affect performance?
4. How robust is the data association process to variations in input data quantity and quality?
5. What is the ideal recall/precision tradeoff in data association to best support situational awareness at the graph analytic processes?

The metrics identified in support of the evaluation of the above experimental questions are described subsequently.

IV. Physical Sensor Tracking and Attribution Evaluation

We use a Deformable Part Model [13], abbreviated DPM, to detect specific instances of object categories in the hard data video frames. The DPM method is the state-of-the-art object detection method in the computer vision literature [14]; it depends heavily on methods for discriminative training and combines a margin-sensitive approach for data mining hard negative examples within a formalism called latent SVM (Support Vector Machine). The DPM model represents an object as a set of parts that are permitted to locally displace

(translate; despite the name deformable, there is no actual deformation in the model) allowing it to adapt to variations in object structure, articulations, and weak visual evidence. The model uses histograms of oriented gradients [15] as local features extracted from the images. During inference, the parts are allowed to displace locally and the reported detection score is the one that yields a maximum score over all configurations of the local parts.

To facilitate fair experimentation on the relatively small SYNCOIN physical sensor dataset (see Figure 2), we directly used the car and the human (upright pedestrian) DPM models that are available in the software package from Felzenswalb's PASCAL VOC experiments (see [13]). In other words, we do not train a separate DPM model specifically in our experimental scenario because the available samples are too few. The Felzenswalb's PASCAL VOC models are trained on the respective PASCAL VOC data, which are images and not video. Performance improvements are expected if trained on domain-specific data.

For tracking after detection, we use a tracking-by-detection framework and dynamic programming to compute best-fit tracks over the videos [16]. The basic method computes a best-fit path through the full set of detected objects over time. The best-fit minimizes a deformation penalty (penalizes large frame-to-frame motion) and computes the globally optimal tracks for the given set of detected objects.

For evaluation of the hard data extraction, we rely on well-established techniques from the computer vision community PASCAL VOC benchmark [14]. Specifically, for each semantic category, such as vehicles and people, we conduct a separate evaluation. Since we are concerned with detection, we essentially evaluate "for a given image, where are the instances of category X (if any)?" As in the PASCAL VOC, we will use the average precision metric to evaluate the detections. The first part of the evaluation is determining a positive hit for which we use the intersection-over-union criterion. Following PASCAL VOC, let the predicted bounding box for a given task be denoted by B_p and the ground truth be denoted by B_g. We compute an overlap ratio: $\rho = \frac{area(B_p \cap B_g)}{area(B_p \cup B_g)}$. When the overlap threshold exceeds a predetermined value (PASCAL VOC suggest 0.5) then the detection is considered a positive hit.

Given these positive hits, for average precision of a given task and class, we compute the standard precision-recall curve. The average precision is used to compute a summary statistic of the shape of the precision-recall curve. It is computed as the mean precision for a uniformly spaced set of recall values. The PASCAL VOC uses eleven such recall values, and we will follow this specification.

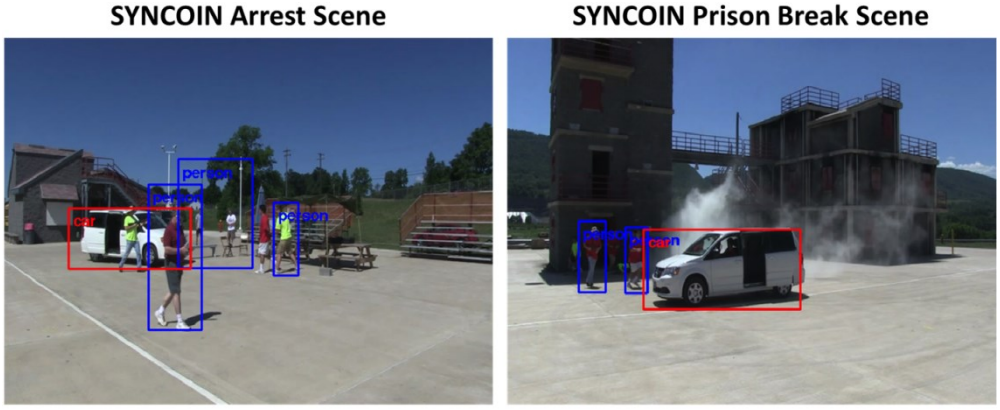

Figure 2 – Example detections on the SYNCOIN videos.

V. Natural Language Understanding Evaluation

Tractor [17],[18] is the subsystem of our hard+soft fusion system that is designed to understand soft information. In this context, understanding soft information means creating a knowledge base (KB), expressed in a formal knowledge representation (KR) language, that captures the information in an English message. Tractor operates on each message independently, and outputs a formal KB consisting of a series of assertions about the situation described in the message. The assertions include the categories (or types) each entity and event mentioned in the message is an instance of, the attributes of those entities and events, and the relations among the entities and events. The assertions are expressed in the SNePS 3 KR language [19],[20] and can be viewed as forming a propositional graph [21]. The assertions that are extracted from the message are enhanced with relevant ontological information from VerbNet [22] and WordNet [23] and geographical information from the NGA GeoNet Names Server database [24].

How is a system such as Tractor to be evaluated? Within Tractor evaluation the notion of "ground truth" does not apply, because regardless of the actual situation being described in the message, if the writer of the message described the situation poorly, no one would be able to reconstruct the situation from the poor description. Instead, the system should be judged by comparing it to a human's performance on the same task. We present a scheme for evaluating a message-understanding system by a human "grader" who produces an "answer key," then compares the system's performance to the key.

The answer key is created by the graders carefully reading the message and listing a series of simple phrases and sentences. The phrases should include all the entities and events mentioned in the message, with the entities categorized

into: people; groups of people; organizations; locations; other things, whether concrete or abstract; and groups of things. The simple sentences should express: each attribute of each entity, including the sex of each person for whom it can be determined from the message; each attribute of each event, including where and when it occurred; each relationship between entities; each relationship between events; and each relationship between an event and an entity, especially the role played by each entity in the event. If there are several mentions of some entity or event in the message, it should be listed only once, and each attribute and relationship involving that entity or event should also be listed only once.

If two different people create answer keys for the same message, the way they express the simple phrases and sentences might be different, but even though it might not be possible to write a computer program to compare them, it should still be possible for a person to compare the two answer keys. In this way, a person could grade another person's performance on the message-understanding task. Similarly, if a message-understanding program (e.g., Tractor) were to write a file of entries in which each entry has at least the information contained in the answer key, a person could use an answer key to grade the program.

Tractor writes a file of answers supplying the same kind of entries as the answer key, but with some additional information to help the grader decide when its answers agree with the answer key. For each entity or event other than groups, Tractor lists: a name or simple description; a category the entity or event is an instance of, chosen from the same list given above; a list of the least general categories the entity or event is an instance of; a list of the text ranges and actual text strings of each mention of the entity or event in the message. For each group, Tractor lists: a name or simple description; a category that all members of the group are instances of; a role that all members of the group fill; a list of mentions as above. For each attribute or relationship, Tractor lists an entry in the format (R a_1 a_2 ...), where R is the attribute or relation, a_1 is the entity, group, or event it is an attribute of, or the first argument of the relation, and a_i is the attribute value, or the i^{th} argument of the relation.

Given an answer key, a person can grade another person's answer key, Tractor's submitted answers, or the submission of another message-understanding program. Grading involves comparing the entries in the answer key to the submitted answers and judging when they agree. We call the entries in the answer key "expected" entries, and the entries in the submission "found" entries. An expected entry might or might not be found. A found entry might or might not be expected. However, a found entry might still be correct even if it wasn't expected. For example, some messages in our corpus explicitly give the MGRS coordinates of some event or location, and MGRS coordinates are also found in the NGA GeoNet Names Server database and added to the KB. If MGRS coordinates were not in the message, but were added, they would not have been expected, but may still have been correct. The grade depends on the following counts: a = the number of expected entries; b = the number of expected entries that were found; c = the number of found entries; d = the number of found entries that were expected or otherwise correct. These counts are combined into evaluation measures adapted from the field of Information Retrieval [25]: $R = b/a$, the fraction of expected answers that were found; P= d/c, the fraction of found entries that were expected or otherwise correct; $F = 2RP/(R + P)$, the harmonic mean of R and P. R, P, and F are all interesting, but F can be used as a summary grade. Average grades for 80 messages of the SYNCOIN dataset are, R=0.83, P=0.84, F=0.83.

VI. Common Referencing and Uncertainty Alignment

We consider the common referencing process of uncertainty alignment [26],[27]. Uncertainty alignment attempts to resolve a number of inconsistencies within the soft data stream including: qualitative language (e.g., "tall" person), human observational biases and variance and uncertainty transformations if required (e.g., enabling comparisons between fuzzy and probabilistic uncertainty representations). Due to the uncertain nature of inferences made by the uncertainty alignment process, it is difficult to quantify these results as "correct" or "incorrect." As a result, within our T&E of the uncertainty alignment process (see [26]) we have assessed the benefit of uncertainty alignment to the fusion processes of data association and situation assessment (through graph matching). This T&E process has shown a significant benefit of uncertainty alignment to both data association and graph matching.

VII. Data Association

A. Overview

If hard+soft data sources contain duplicate references to the same real world entity, event or relationship, the data association process needs to be performed, for merging common entities, events and relationships into fused evidence. This fused evidence is used in sense-making processes to make inferences on the state of the real world (obtain situational awareness) [1]. The data association problem can be modeled as a graph association problem. Different data association formulations (Graph Association or GA^N, Multidimensional Assignment problem with Decomposable Costs or MDADC and Clique Partitioning Problem or CPP) and their related algorithms for data association were studied on this program by Tauer et al. [28],[29] and Tauer and Nagi [30], each of which has its own strengths and weaknesses.

The first step of data association is to measure and quantify the similarity between pairs of nodes (or edges) in the input dataset. These similarity scores are calculated using a similarity function, which provides a positive score if two elements are similar; and a negative score if two elements are dissimilar. The absolute value of the similarity score is an indication of the strength of similarity or dissimilarity between a certain node/edge pair.

Given these similarity scores, data association tries to cluster (or associate) the nodes/edges which are highly similar, and produces a *cumulative data graph* (CDG), which is the cumulative fused evidence. The cumulative evidence should describe the real world as accurately as possible from the provided input data, so as to draw satisfactory conclusions on

the state of the real world. This calls for the development of an objective strategy for training and evaluating the performance of data association processes. This evaluation strategy also needs to be efficient with minimal human intervention. In this section, we will briefly describe the evaluation methodology that has been developed for assessing data association both with a "system perspective" and isolated "data association perspective."

B. Evaluation Methodology

The evaluation methodology for data association is divided into two tasks: ground truth development and an evaluation process, as discussed below.

1) Ground Truth Development

Development of the ground truth is a key step for evaluating the performance of any data association algorithm. The ground truth is typically prepared by one or more human analysts and it represents the answer key to the data association solution, against which the association algorithm is graded. The *soft ground truth* contains a list of unique entities, events and relationships with a unique identifier (UID) assigned to each of them; and another list containing observations of the unique entities, events and relationships (with respective UIDs) in various soft messages. The analyst also records the pedigree information related to each of entity, which represents the exact location and number of characters in the textual description of that entity in a particular text message. The *hard ground truth* contains similar lists of unique and observed entities and events, present in each of the hard data sources, with cross-modality UIDs carried forward from soft data ground truthing.

2) Evaluation Process

As mentioned before, the performance of data association is assessed at two levels. For assessing the cumulative system performance (the "system perspective") at the data association process, the CDG is compared with the ground truth and three types of entity pairs are counted: (a) *correctly associated*; (b) *incorrectly associated*; and (c) *incorrectly not associated*. These counts are obtained by programmatically comparing the pedigree records of the nodes in the CDG with those of the entity observations in the ground truth. After obtaining these counts, we quantify the performance of the data association, using Precision, Recall, and F-score, which are defined below.

- *Precision:* Ratio of correctly associated entity pairs to the total number of associated entity pairs (i.e. $\frac{a}{a+b}$).
- *Recall*: Ratio of correctly associated entity pairs to the total number of correctly associated and incorrectly not associated entity pairs (i.e. $\frac{a}{a+c}$).
- *F-score*: Harmonic mean of the Precision and Recall values i.e. $\frac{2\times precision\times recall}{precision+recall}$.

The higher values of these metrics typically indicate greater accuracy. Since maximizing Precision and Recall are competing objectives, our focus is on maximizing the F-score. For this purpose, we trained a logistic regression model on the feature scores for a separate training dataset. The training algorithm calculates the optimal values of the feature weights used in similarity score calculation, with an objective of maximizing the F-score. For a more in depth description of the scoring and evaluation processes, readers are directed to [12].

Assessing the "data association perspective" performance of data association is not so straightforward, because any imprecision in the upstream processes could influence the data association results. Two examples of such imprecisions are: incorrect or missing entity typing and incorrect or missing within message co-referencing. To assess the standalone performance of data association (the "data association perspective"), we need to identify and disregard imprecisions in the data association solution stemming from upstream processes. To this end, we will explain the *type restricted evaluation* method, which helps in isolating association performance on "correct" input data (see [12] for a detailed explanation). In this method, we identify the entity pairs which are incorrectly associated or incorrectly not associated due to NLU errors; and disregard them from the counts (b) and (c) mentioned above. To prevent an unfair inflation of the Precision and Recall, we also identify the correct associations which overcame the NLU errors, and disregard them from the count (a). Using these counts, we can calculate the "data association perspective" Precision, Recall and F-score for data association, which are likely higher than their "system perspective" counterparts.

Note that our current association perspective evaluation strategy does not support the nullification of the effects of within message co-referencing errors. However, modeling data association as clique partitioning problem (CPP) helps recover some of the missing within message co-references and improves the F-score (as seen in Table 2).

C. Testing

We tested our evaluation strategy on the three data association formulations and corresponding algorithms: sequential Lagrangian heuristic for GA^N, Map/Reduce Lagrangian heuristic for MDADC and streaming entity resolution algorithm for CPP (see [28]-[30]). The procedures were coded in Java and executed on Intel Core 2 Duo processor, with 3 GHz clock speed and 4GB RAM. We have used a sample vignette message set of SYNCOIN as the input data set, which contains 114 soft messages and 13 hard messages. The statistics related to the evaluation engine are presented in Table 1, and the computational results for the data association algorithms are presented in Table 2.

Overall 46,030 pairs of pedigree records were compared during the evaluation process, of which 1,302 are within-message and 44,728 are between-message. We see that the association perspective evaluation (the lower row performance metrics within Table 2) results in higher Precision, Recall, and F-score, as expected.

The sequential Lagrangian procedure for GA^N formulation takes the second longest time to solve because of the complexity of the model. The Map/Reduce Lagrangian procedure for MDADC is quite fast, as a result of parallelization. Thus, for large graphs, the sequential Lagrangian heuristic for GA^N will prove to be a bottleneck. On

the other hand, MDADC formulation solved using Map/Reduce can potentially provide a quick and accurate solution and it is easily scalable for larger graphs. The cumulative time required for Streaming Entity Resolution algorithm, is the largest; however it takes only 10 seconds per graph update. Streaming resolution also helps recover the missing within-message associations, improving the Recall of the system perspective evaluation.

Table 1. EVALUATION STATISTICS FOR SEQUENTAL GA^N.

Evaluation Mode	Correctly Associated	Incorrectly Associated	Incorrectly Not Associated
System Perspective	30,563	2,708	12,759
Association Perspective	29,349	2,382	8,836

Table 2. SYSTEM (UPPER ROW) AND ASSOCIATION PERSPECTIVE (LOWER ROW) ASSOCIATION PERFORMANCE BY ALGORITHM.

No.	Procedure	Precision	Recall	F-Score	Compute Time (s)
1	GA^N (Sequential)	0.918	0.705	0.798	794
		0.925	0.768	0.839	
2	MDADC (MR)	0.932	0.708	0.805	64
		0.938	0.772	0.847	
3	CPP (Streaming)	0.909	0.730	0.810	1,312 (10 s/graph update)
		0.915	0.796	0.851	

VIII. SENSEMAKING VIA GRAPH ANALYTIC PROCESSES

The situation assessment processes within the SUT utilize as input the cumulative associated data graph formed by the data association process. The graph analytic processes for situation assessment within our SUT are representative of just one analytic strategy for a hard+soft information fusion system, but they can be examined to illustrate some of the complexities of the broader evaluation issues for automated tools designed to aid sensemaking.

There are two major aspects for assessing a toolkit of automated methods to support a human-based sensemaking process: the performance of the algorithms in forming automated situational *assessments* (algorithmically-formed hypotheses), and the (possibly-separate) ability of these algorithms to aid in the formation of human-based situational *awareness*. While an automated algorithm (e.g., graph matching) may be *efficient* in *assessing* matches to specified situations of interest, this technique in itself may not be *effective* in supporting domain-wide *awareness.* This is in part because of the underlying discovery/learning-based approach to sensemaking and the limitations of deep knowledge in modern problem domains such as counterinsurgency (COIN). In complex and dynamic problem environments like these, even the best assessment-supporting technologies are of limited capability today and many produce what we will call "situational fragments," *partial* hypotheses representing situational substructures as patterns. Situational awareness at a more complete level is the result of a dynamic interaction with the assessment tools, possibly using other technology to connect these "fragments" (as the human is trying to do) and human judgment in a kind of mixed-initiative operation. The evaluation focus of the graph-analytic tools in our SUT is on measuring the situation *assessment* capabilities, with the evaluation of effectiveness in developing situational awareness left for future work (see Section IX).

Three graph analytic processes within our SUT have been previously evaluated: a link analysis tool, social networking tool and stochastic graph matching tool. The algorithmic computational efficiency, specifically with a focus on data size scalability, of the link analysis algorithm is described in [31]. The evaluation of the social network tool for social network extraction and high value individual (HVI) identification is described in [1]. Finally, the evaluation of the stochastic graph matching tool to efficiently identify situations of interest within the cumulative associated data is presented in [32].

IX. DISCUSSION AND FUTURE WORK

The example SUT and evaluation point process and system level performance metrics form the basis for error audit trail analysis. Through the utilization of this error audit trail numerous questions can be answered within the test space as described in Section III, for example: What is the value of hard+soft fusion (versus hard only or soft only) toward some system level objective? While the answer to this and other experimental questions is ultimately the goal of this approach in systemic testing, we are currently still completing the training phase of this effort. In addition to the assessment of the evaluation questions listed in Section III on an independent test data set, other issues in the evaluation of hard+soft information systems remain as future work. Additional questions which will be assessed as future work include: how does one assess generality of methods on independent training and test data [3]? What are the challenges of testing in a streaming environment and how are performance metrics in tune with the dynamic user requirements within these environments? What are the dimensions of scalability which must be considered both in input data and decision dissemination? What is the relationship between situational awareness and the resulting actions taken?

X. CONCLUSIONS

This paper presented a metric-based test and evaluation (T&E) framework for the assessment of a hard+soft fusion system. Issues in the definition of a System Under Test (SUT) and evaluation points in an active Research and Development program were discussed. An example SUT from the MURI Network-based Hard+Soft Information Fusion project is considered, with evaluation metrics at both the "process" and "system" level for each evaluation point provided. The future use of the evaluation framework in assessing design alternatives and incremental research and development efforts is also provided.

[3] We recognize there is some existing literature in the area of quantifying characteristics of a textual corpus via: statistical vocabulary analysis (lexicometry [33]), textural complexity (textometry [34]) and linguistic style (stylometry [35]) among other approaches. The investigation of these measures as an argument for framework generality remains as future work.

ACKNOWLEDGMENT

The authors gratefully acknowledge that this research activity is supported by a Multi-disciplinary University Research Initiative (MURI) grant (Number W911NF-09-1-0392) for Unified Research on Network-based Hard/Soft Information Fusion, issued by the US Army Research Office (ARO) under the program management of Dr. John Lavery.

REFERENCES

[1] K. Date, G. A. Gross, S. Khopkar, R. Nagi, K. Sambhoos. 2013. "Data association and graph analytical processing of hard and soft intelligence data," Proceedings of the 16th International Conference on Information Fusion (Fusion 2013), Istanbul, Turkey, 09-12 July 2013.

[2] C. L. Hornbaker II, "Tactical Fusion Centers: Restructuring Intelligence for Counterinsurgency," M.A. Thesis, (2012).

[3] J. Llinas "Information Fusion for Natural and Man-Made Disasters." Proc. of The 5th International Conference on Information Fusion, Annapolis, MD, 570-77, (2002).

[4] D. L. Hall and J. M. Jordan. "Information Fusion for Civilians: The Prospects of Mega-Collaboration." *Human-centered Information Fusion*. Boston: Artech House, 2010. 211-26.

[5] P. V. Puttan, J. N. Kok and A. Gupta, "Data Fusion through Statistical Matching," (2002).

[6] "Information fusion within the retail sector," letter, <http://archive.his.se/PageFiles/7158/ICA_scanned.pdf>, (2004).

[7] "Unified Research on Network-based Hard/Soft Information Fusion", Multidisciplinary University Research Initiative (MURI) grant (Number W911NF-09-1-0392) by the US Army Research Office (ARO) to University at Buffalo (SUNY) and partner institutions.

[8] D. L. Hall, J. Graham, L. D. More and J. C. Rimland. "Test and Evaluation of Soft/Hard Information Fusion Systems: A Test Environment, Methodology and Initial Data Sets." Proc. of The 13th International Conference on Information Fusion, Edinburgh, Scotland (2010).

[9] E. Blasch, P. Valin, and E. Bosse. "Measures of Effectiveness for High-Level Fusion." Proc. of The 13th International Conference on Information Fusion, Edinburgh, Scotland, (2010).

[10] M. R. Endsley, "Measurement of Situation Awareness in Dynamic Systems." Human Factors: The Journal of the Human Factors and Ergonomics Society 37.1 (1995): 65-84.

[11] J. J. Salerno, E. P. Blasch, M. Hinman, and D. M. Boulware. "Evaluating Algorithmic Techniques in Supporting Situation Awareness." Multisensor, Multisource Information Fusion: ARCHITECTURES, ALGORITHMS, AND APPLICATIONS. Ed. B. V. Dasarathy. S.l.: Int'L Soc For Optical, 2005. 96-104.

[12] K. Date, G. A. Gross, R. Nagi. "Test and Evaluation of Data Association Algorithms in Hard+Soft Data Fusion," Proc. of the 17th International Conference on Information Fusion, Salamanca, Spain, (2014).

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discrimi- natively trained part based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, pp. 1627–1645, 2010.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, 2010.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings of IEEE International Conference on Computer Vision, vol. 2, pp. 886–893, 2005.

[16] H. Pirsiavash, D. Ramanan, C. Fowlkes. "Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects" Computer Vision and Pattern Recognition(CVPR) Colorado Springs, Colorado, June 2011.

[17] M. Prentice, M. Kandefer, and S. C. Shapiro, "Tractor: A framework for soft information fusion," in Proceedings of the 13th International Conference on Information Fusion (Fusion2010), 2010, p. Th3.2.2.

[18] S. C. Shapiro and D. R. Schlegel, ``Natural language understanding for soft information fusion," in Proceedings of the 16th International Conference on Information Fusion (Fusion 2013), 2013, 9 pages, unpaginated.

[19] S. C. Shapiro and W. J. Rapaport, "The SNePS family," Computers & Mathematics with Applications, vol. 23, no. 2-5, pp. 243-275, January-March 1992, reprinted in F. Lehmann, Ed., Semantic Networks in Artificial Intelligence, Oxford: Pergamon Press, 1992, pp. 243-275.

[20] S. C. Shapiro, "An introduction to SNePS 3," in Conceptual Structures: Logical, Linguistic, and Computational Issues, Lecture Notes in Artificial Intelligence, B. Ganter and G. W. Mineau, Eds. Berlin: Springer-Verlag, 2000, vol. 1867, pp. 510-524.

[21] D. R. Schlegel and S. C. Shapiro, "Visually interacting with a knowledge base using frames, logic, and propositional graphs," in Graph Structures for Knowledge Representation and Reasoning, Lecture Notes in Artificial Intelligence, M. Croitoru, S. Rudolph, N. Wilson, J. Howse, and O. Corby, Eds. Berlin: Springer-Verlag, 2012, vol. 7205, pp. 188-207.

[22] M. Palmer, VerbNet: A Class-Based Verb Lexicon, University of Colorado Boulder. http://verbs.colorado.edu/~mpalmer/projects/verbnet.html. Last accessed 27 February 2014.

[23] Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu. Last accessed 27 February 2014.

[24] National Geospatial-Intelligence Agency, NGA GEOnet Names Server, http://earth-info.nga.mil/gns/html/. Last accessed 27 February 2014.

[25] C. van Rijsbergen, Information Retrieval, Second Edition, London: Butterworths, 1979.

[26] M. Jenkins, G. Gross, A. Bisantz, R. Nagi, "Towards Context Aware Data Fusion: Modeling and Integration of Situationally Qualified Human Observations into a Fusion Process for Intelligence Analysis," Journal of Information Fusion, (Accepted June 2013).

[27] M.P. Jenkins, G. A. Gross., A. M. Bisantz, and R. Nagi, "Towards context-aware hard/soft information fusion: Incorporation of situationally qualified human observations into a fusion process for intelligence analysis." In Proc. of the 2011 IEEE First International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), Feb. 22-24, Miami Beach, FL, pp. 74-81.

[28] G. Tauer, R. Nagi, and M. Sudit, "The graph association problem: mathematical models and a Lagrangian heuristic". Naval Research Logistics, vol. 60, pp. 251-268, April 2013.

[29] G. Tauer, K. Date, R. Nagi, and M. Sudit, "An incremental graph-partitioning algorithm for entity resolution," Under Revision. To be submitted to Transactions on Knowledge Discovery from Data.

[30] G. Tauer, and R. Nagi, "A Map-Reduce Lagrangian heuristic for multidimensional assignment problems with decomposable costs," Parallel Computing, 39(11), pp. 653-658, November 2013.

[31] G.A. Gross, Graph Analytic Techniques in Uncertain Environments: Graph Matching and Link Analysis, Dissertation, University at Buffalo, 2013.

[32] G. A. Gross, R. Nagi, and K. Sambhoos. "A Fuzzy Graph Matching Approach in Intelligence Analysis and Maintenance of Continuous Situational Awareness." Information Fusion 18 (2014): 43-61.

[33] Z. Harris, "Mathematical Structure of Language." John Wiley, New York, 1968.

[34] I. Aydin, and E. Seker, "Textometry: A Method for Numerical Representation of a Text," International Journal of Humanities and Social Science Vol. 2 No. 23; December 2012.

[35] M. Brennan, et al., "Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity," ACM Transactions on Information and System Security, Vol. 15, No. 3, Article 12, November 2012.