Tractor: A Framework for Soft Information Fusion

Michael Prentice, Michael Kandefer, and Stuart C. Shapiro
Department of Computer Science and Engineering,
Center for Cognitive Science,
Center for Multisource Information Fusion
University at Buffalo, Buffalo, NY 14260
{mjp44,mwk3,shapiro}@cse.buffalo.edu

Abstract – *This paper presents a soft information fusion framework for creating a propositional graph from natural language messages with an emphasis on producing these graphs for fusion with other messages. The framework utilizes artificial intelligence techniques from natural language understanding, knowledge representation, and information retrieval.*

Keywords: hard/soft data fusion, ontologies, propositional graphs, context.

1 Introduction

Information fusion is defined as the process of associating, correlating, and combining data and information from single and multiple sources to achieve refined estimates of characteristics, events, and behaviours for observed entities in an observed field of view [15]. Recently there has been an increased interest in applying information fusion to soft information sources such as natural language [16, 29]. The goal of such research is to fuse information contained within natural language data with more refined data in order to establish a better understanding of the domain in question. A system that is able to properly accept and utilize natural language information must solve two problems: understanding the contents of the message, a problem called *natural language understanding* [40], and evaluating the message in the context of its background knowledge, state estimate, and other—possibly conflicting—messages.

This paper presents the Tractor architecture for performing soft information fusion on natural language messages. The architecture utilizes artificial intelligence (AI) techniques from natural language understanding, knowledge representation, and context-based information retrieval to process messages and produce a data graph, which is a representation of a set of propositions about the world. The data graph is also called the *state estimate* [32].

Development of the Tractor architecture is part of ongoing research into using information fusion for counterinsurgency (COIN). Previous work has focused on COIN utilizing geographic entity tracking and using ontologies for soft data fusion [22, 23]. Tractor is part of a system focusing on Army operations involving counterinsurgency [10, 14, 31]. The goal of this research is to understand messages and other sensor data in the Army counterinsurgency domain and use the information to create estimates on the state of the current counterinsurgency operation. Further goals include using AI techniques to reason from this information to additional useful information that will aid in the operation.

2 Soft Information Fusion

Work on the Soft Target Exploitation and Fusion (STEF) research project [32] has resulted in a message repository that contains soft information from a manufactured counterinsurgency scenario. A message in this domain is a timestamped, small English paragraph containing information about activity that might be of interest to a counterinsurgency operative. An example message is:

> 02/10/07 American contractor in Yarmuk said ambulance service is deteriorating; he told of a friend who suffered a heart attack in Qahtan Square, a short distance south of the hospital. His friend had to wait nearly an hour before the ambulance arrived.

Such messages pose several challenges to information fusion [32]. A framework that can fuse the information in these messages with already processed information requires the following capabilities:

- **Natural language understanding** to create a suitable representation of the data contained in the natural language text

- **Background knowledge** in the form of ontological information and a priori knowledge of the domain.

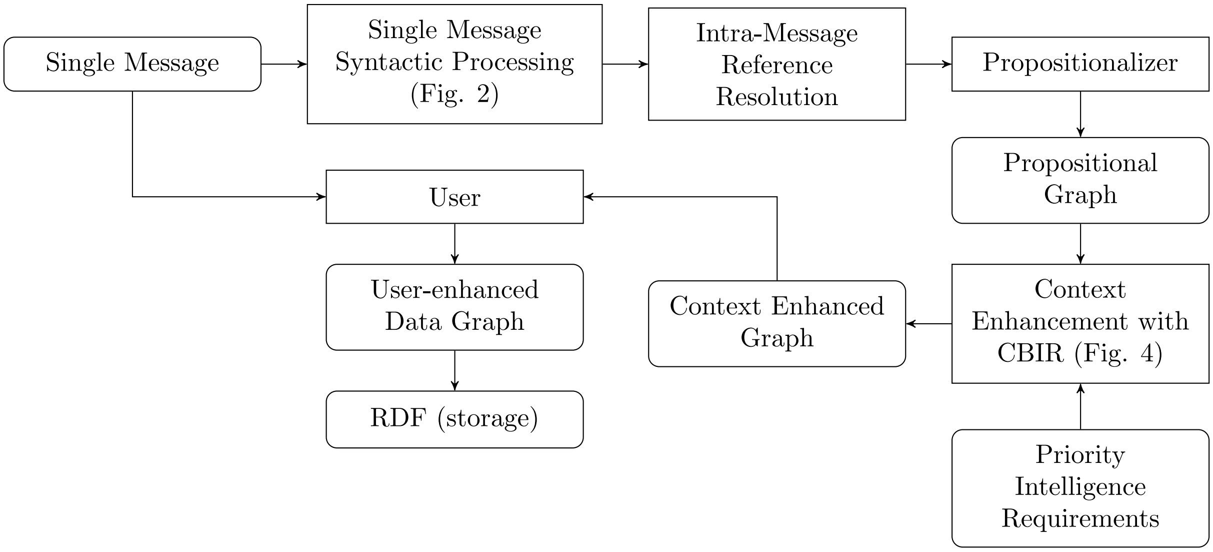
Single Message
Single Message Syntactic Processing (Fig. 2)
Intra-Message Reference Resolution
Propositionalizer
Propositional Graph
User
User-enhanced Data Graph
RDF (storage)
Context Enhanced Graph
Context Enhancement with CBIR (Fig. 4)
Priority Intelligence Requirements

Figure 1: The Tractor Architecture

- **Data fusion** to incorporate the new information in the state estimate. Since all counterinsurgency domains contain background knowledge, or ontologies, which are useful for reasoning about sensor data [4, 20, 25], a soft information fusion process must be capable of fusing the information in a newly processed message with the background information about that message. This allows future processes to make use of all the relevant information associated with the message, and places the contents of the message in context.

3 Propositional Graphs

Propositional graphs, sometimes called *propositional semantic networks*, are a knowledge representation (KR) technique that has been chosen to represent background information and the information contained in the natural language data sources. Since propositional graphs are types of semantic networks [24, 34, 35], they have all of the advantages of semantic networks, like RDF, for soft information fusion, such as [20]:

- The ability to reason about semantic relationships between domain entities and make inferences about them, and

- The ability to merge information together between concepts represented in the graphs, which is useful when it is discovered that the concepts represent the same entity.[1]

In addition to these advantages, propositional graphs have great expressivity in their ability to represent complex assertions. Propositional graphs have the following capabilities:

- Propositions are represented as terms in the representational language, thus propositions can be the arguments of other propositions (e.g., relationships). This is useful in situations such as representing what another individual said. In the message an American contractor says that ambulance service is deteriorating. There are two propositions contained in this message that should be represented. The first is the assertion that ambulance service is deteriorating, and the second is that the American contractor said this.

- Relationships are *n-ary* (i.e., they can have any number of arguments). In semantic networks relationships are binary, but propositional networks lack this restriction [34]. This is useful in any domain that requires representing complex relationships, such as the location of Qahtan Square in the message: “Qahtan Square, a short distance south of the hospital.” Qahtan Square stands in a ternary relation containing the location it’s relative to, the hospital; its direction, south; and its distance, short. Techniques for including n-ary relations in RDF are discussed in [38]. Other uses of n-ary relations in graphs for fusion applications are discussed in [36].

4 Tractor Architecture

The *Tractor Architecture* describes processes for implementing a program that can transform a natural lan-

[1] There are many ways of merging information about two concepts that are later discovered to be the same entity, advantages and disadvantages of some approaches are discussed by Maida & Shapiro [24].

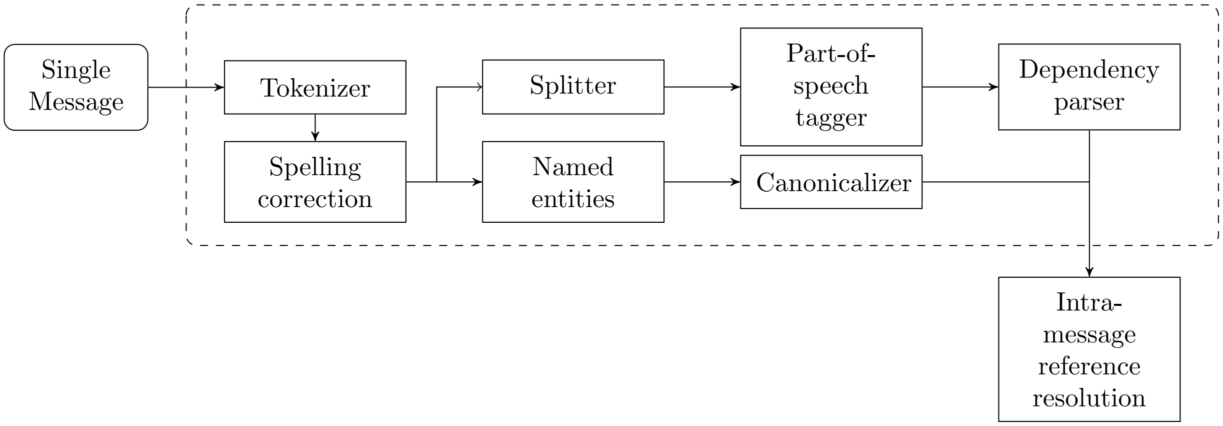

Figure 2: Single Message Syntactic Processing - Detail

guage input into a set of propositions about the world in the context of other messages and background information. This additional information is provided through a process called *contextual enhancement*. Though the architecture is described generally, some choices for implementing a program that instantiates the framework have been made, and others seem practical.

The architecture depicted in Fig. 1 implements a soft information fusion process. In the architecture messages are processed one at a time and recovered information is added to a state estimate. A message is processed through a modular, multi-stage pipeline:

1. **Single Message Syntactic Processing:** The syntax processing module performs lexical and syntactic analysis on the message, discussed in more detail below.

2. **Intra-message Reference Resolution:** A reference resolution process that establishes when two or more terms in the same message co-refer to the same entity. This process must be capable of recognizing different names that refer to the same entity, as well as identifying whether indexical terms (e.g.,"it", "there") in the message refer to the named entities identified in the previous process.

   An important assumption is that each message is written by a single author, whereas there is no guarantee that a different message has the same author. A consequence of this assumption is that anaphoric references across multiple sentences within a single message can be resolved. For example, the "his" in "his friend" in the example message refers to the American contractor.

3. **Propositionalizer:** This process takes the results of the reference resolution and single message syntax processes and constructs a propositional graph representation of the message [28, 34, 35, 39]. The propositional graph contains all assertions about the named entities. If an entity could still have multiple references (i.e., it is unresolved, but has potential resolution), this process will construct assertions for each possible referent.

4. **Contextual Enhancement with Context-Based Information Retrieval** Additional information is retrieved about the message entities, reasoning is performed, and the results are fused with the propositional graph. Details of this sub-process are discussed below.

5. **Display and Human Analysis:** This process displays the original message alongside the *enhanced propositional graph* and allows an analyst to select a referent for any unresolved references given the contextual information provided by the contextual enhancement process, as well as make tactical decisions based on the contents of the graph. Though not specified in the architecture, this process could be used to provide modifying feedback to any implementation decisions that rely on such feedback. The results of this process are output as a *user-enhanced graph*, which is placed into storage using Resource Description Framework (RDF) [30] for future processing by other applications (e.g., automated reasoners). RDF is a standard format for representing propositional information, which aids data portability.

4.1 Single Message Processing

Tractor uses the GATE framework to process natural language text [37]. The Tractor application performs two main functions: reference resolution for named entities and intra-sentence co-referential information; and syntactic parsing. These functions are divided into submodules.

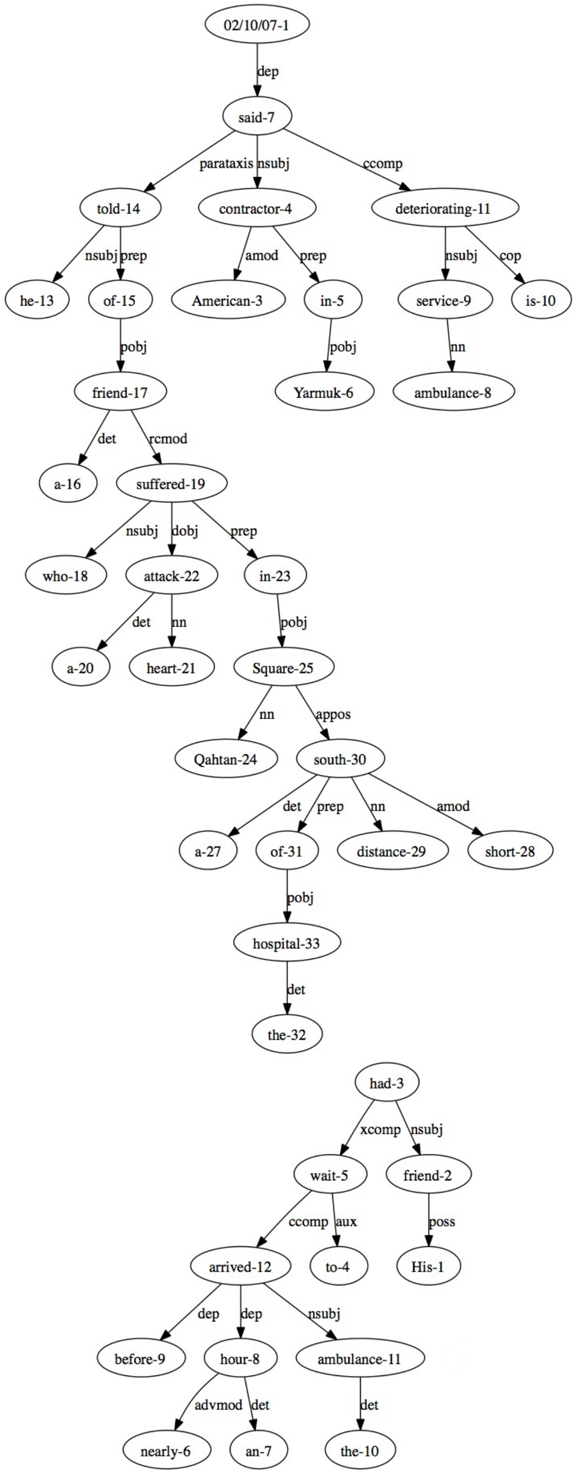

Figure 3: Dependency Parse of "02/10/07 American contractor in Yarmuk said ambulance service is deteriorating; he told of a friend who suffered a heart attack in Qahtan Square, a short distance south of the hospital. His friend had to wait nearly an hour before the ambulance arrived."

- **Tokenizer:** Recognizes word boundaries and punctuation. From this point on, lexical items are treated as atomic *tokens*. Tractor uses GATE's default tokenizer plugin.

- **Spelling correction:** Tokens are looked up in an internal dictionary. If a word is unrecognised, spell checking is performed. The token is tagged with possible spellings from the dictionary to aid in further processing, but crucially no replacement is performed (the message text is left intact). Tractor uses a separate spelling correction plugin to GATE, currently in development.

- **Named entities:** Named entity reference resolution is the process of resolving a name (which can consist of a word or phrase, and optional clarifying text) to the entity to which the name refers. In order to bring natural language text into a searchable database, these named entity references need to be resolved into either: (a) the unique entity to which it refers or (b) a set of possible referents.

  In case (b), the reference can be disambiguated by a later processing step, or, in systems with human interaction, by a human user (see [6]).

  Problems include automated disambiguation of the named entity from clarifying text within the message itself, or from background knowledge of the message domain. In order to select the referent(s) of the named entity references in the text, it is often helpful to perform a dictionary lookup in an external data source. The lookup can fail, in which case the reference is not resolved, or the lookup can succeed but still not uniquely identify the referent.

  Tractor depends on external data sources to identify references to named entities, and must also overcome the problem of *semantic heterogeneity*: different data sources may refer to the same entity using different names. This problem is especially relevant when using geographic information sources [13].

  When support needed for disambiguation is not provided within the message text, Tractor should continue its attempt to disambiguate the reference. Support can come from the ontology [17] and from contextual enhancement.

  Once a named entity is resolved using named entity resolution, it can then be fused with an external data source such as a database and information about the entity can be used in further message processing tasks. Supplementing the text with additional information informs military intelligence information fusion of natural language text [3].

  In Tractor, each word and contiguous sequence of words, and suggested spellings, is matched against

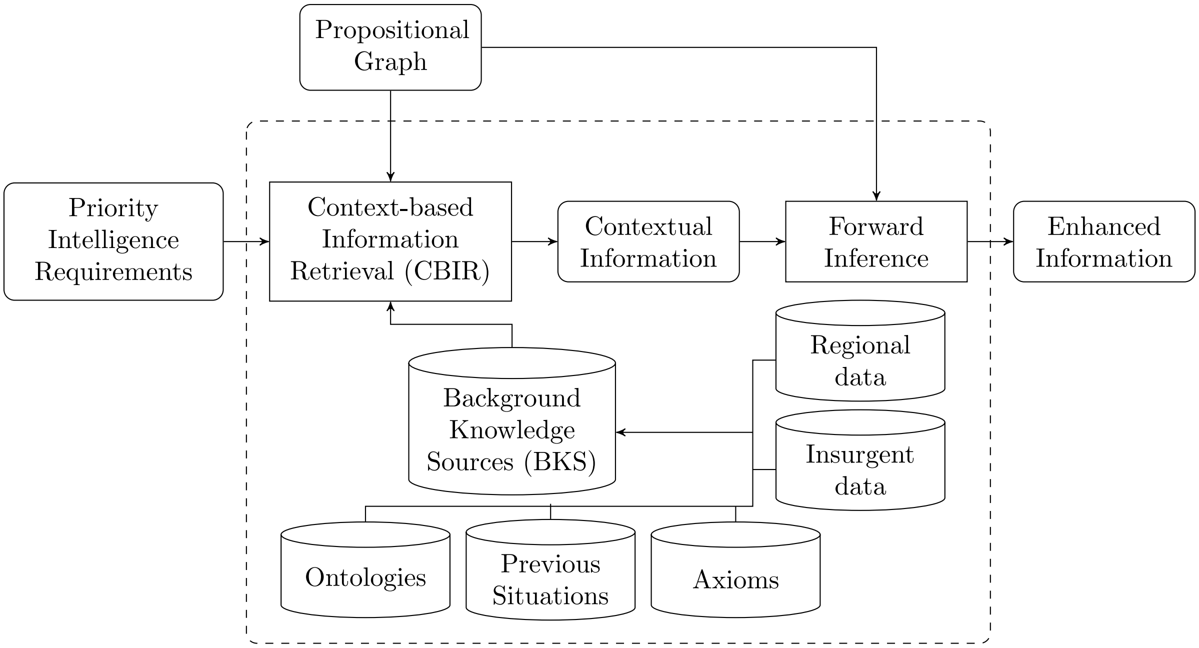

Figure 4: Context Enhancement - Detail

a set of *gazetteers*, which are used to identify possible entities to which the word or phrase refers.

Matching items are tagged with a type and supporting information from a GATE-specific ontology of tag types. This ontology needs to be enhanced with a domain-specific ontology to provide more geographic information to the system. The basic ontology provides rough categories of information. For example, a mention of "U.S." in a text has an associated *Location* tag with a *locType* (location type) of country.

We are currently using a geographic lookup plugin using the National Geospatial-Intelligence Agency GEOnet Names Server [27], which contains information about geographic entities, including aliases. This database is also being used to construct the ontological knowledge base used in the *contextual enhancement* process.

A particular name can be associated with multiple entities. The text processing pipeline of Tractor tags all possible matches for later disambiguation.

- **Canonicalization:** A referring token can be an abbreviation or alternate spelling for an entity (e.g. "U.S." for the geopolitical entity the United States). A geographic location can have multiple names associated with it, variations in spelling, and restrictions based on the originating document's character set and language. In addition, English documents frequently contain Anglicizations of foreign words, especially words from languages with a non-Latin alphabet.

  Canonicalization allows us to treat all references to the entity as the same entity by choosing one preferred representation for an entity and using it consistently. For example, Tractor canonicalizes geographic named entities to their Unique Feature Identifier in the National Geospatial-Intelligence Agency's GEOnet Names Server (GNS) [27].

- **Sentence splitting:** Prior to sentence-by-sentence syntax parsing, the message is split into separate sentences. Tractor uses GATE's default sentence splitting plugin.

- **Part-of-speech tagging:** Tokens are tagged with a part of speech using a statistical model of the language. This aids in parsing. Tractor uses GATE's default plugin.

- **Dependency parser:** The sentence is parsed into a set of *dependency relations* between tokens. These are purely syntactic dependencies based on the grammar of the language. Typed dependencies "provide a simple description of the grammatical relationships in a sentence" [9]. Fig. 3 shows the Stanford typed dependency parse of the message shown in section 2.

  The Stanford dependency parser is robust to "ungrammatical" sentences, i.e. sentences that don't fit into a formal grammar but that nonetheless contain meaning. This is important for processing real-world, uncleaned natural language data.

4.2 Contextual Enhancement with Context-Based Information Retrieval

The general approach of contextual enhancement has many applications. In human computer interaction, mobile devices provide context-aware capabilities such as context-aware information retrieval, procedure recommendation, and user preference learning [5, 12, 21, 11]. In common sense reasoning the information retrieval component can aid in knowledge retrieval [18, 19, 26] and situation dependent acting [2]. Context-based information retrieval can aid information fusion processes that depend on efficient retrieval of information from ontologies [4, 20, 25].

Contextual enchancement within the Tractor architecture is a process that uses the propositional graph as context in order to retrieve additional, relevant information about the entities discussed in the current message. The output of this process is an *enhanced propositional graph* containing information from the original message conjoined with the retrieved information. The process has two components, a Context-based Information Retrieval (CBIR) process, and a Forward-Inferencing process. These sub-processes, and their respective data sources and intermediaries are depicted in Fig. 4.

The CBIR procedure receives message input in the form of a propositional graph, which contains the resolved entities, and potentially some with multiple resolutions, and the assertions about those entities as contained in the message data. The CBIR procedure also receives input from a set of Priority Intelligence Requirements (PIRs), which provide a focus for the counterinsurgency operation (i.e., what is the goal of the operation). The third source of input is the background knowledge sources (*BKS*), which contain knowledge that will be retrieved by the CBIR procedure. Some useful sources for soft information fusion include:

- **Ontologies**: Ontological information about message entities, categorical relationships among them, and restrictions governing those relationships.
- **Previous Messages**: Information stored from previously processed messages, including the propositional graph representations of those messages, the authors, and timestamps.
- **Axioms**: Additional axioms for reasoning about the domain, including relationships not reasoned about in ontologies (e.g., part-hood)
- **Insurgent Data**: Information about known insurgents and activity (e.g., affiliations, contacts, pseudonyms), and rules for reasoning about counterinsurgency.
- **Regional Data**: Information about the regions of the world (e.g., geospatial location, population, alternative names) and how to reason about regions (e.g., connected regions, proximity).

The sources composing the BKS should be, but are not required to be, in propositional graph form like that used for the propositionalized message input. However, a standard makes implementation of a CBIR procedure and combining the results of the CBIR procedure with the original message (i.e., the creation of the enhanced message) easier. With these inputs the CBIR procedure retrieves a subset of the background knowledge, called the *contextual graph*, which contains relevant, additional information about the message contents (i.e, contextual information). The *contextual graph* is then passed to the forward-inferencing process.

The forward-inferencing process performs two tasks. The first utilizes a reasoner that performs forward inferencing on the *contextual graph* with the original *propositional graph* in order to derive new information about the message entities. This new information is derived through a combination of assertional information contained in the original message, along with rules governing relationships, as contained in the ontologies, additional axioms, regional data, and insurgency data. The second task is to merge the *propositional graph* with the results of the forward-inferencing process, as the results of reasoning process may not include all of the information contained in the original message. This combined data is called the *enhanced graph*, the output of the *contextual enhancement* process.

For the *contextual enhancement* process the choice for implementing the *context-based information retrieval* (CBIR) component is crucial, and one made on practicality. Apart from expressivity there is a need to reason to new information utilizing the *Ontologies*, *Axioms*,[2] and *Previous Messages* comprising the *background knowledge sources* (BKS). Classic ontological reasoners only perform reasoning among the class relationships of ontological BKS. As such, we have chosen to include a reasoner that utilizes both an expressive propositional graph representation, and includes more sophisticated rules of reasoning [33]. Due to this choice in representation, the CBIR component utilizes *spreading activation*, an information retrieval procedure that was developed for propositional graphs, and based on models of cognition [1, 7, 8]. Though this choice is practical, there are many methods for implementing *spreading activation*, and many constraints that affect their operation, as such a means of evaluating them in use is required [19].

[2] The background axioms include rules of reasoning that are constructed by an internal team that are used to address specific problems in the domain (e.g., identifying buildings that might be used as safe houses).

5 Conclusions and Future Work

Tractor is a framework for solving soft information fusion problems for natural language messages. The framework utilizes artificial intelligence approaches in concert to parse, represent, and reason about messages in order to produce propositional graph representations of the messages. These propositional graphs are useful for soft information fusion because they represent *n-ary* relationships, easily merge information together, and are used in reasoning. An approach called *context-based information retrieval* aids in retrieving additional, relevant information about the messages for subsequent processes in the soft information fusion stream.

The architecture is a specification for how to implement a natural language processing fusion application. Much of the architecture needs to be implemented and tested. For the CBIR process we are currently investigating the use of spreading activation and testing using a method described in [19].

Acknowledgements

This work has been supported by a Multidisciplinary 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] J. R. Anderson. Human associative memory. In *How Can the Human Mind Occur in the Physical Universe?*, pages 91–134. Oxford University Press, NY, New York, 2007.

[2] R. Arritt and R. Turner. Situation assessment for autonomous underwater vehicles using a priori contextual knowledge. In *Proceedings of the Thirteenth International Symposium on Unmanned Untethered Submersible Technology (UUST)*, 2003.

[3] J. Biermann, L. de Chantal, R. Korsnes, J. Rohmer, and Çagatay Ündeger. From unstructured to structured information in military intelligence—some steps to improve information fusion. In *RTO SCI Symposium on Systems, Concepts and Integration (SCI) Methods and Technologies for Defence Against Terrorism.* RTO-MP-SCI-158, 2004.

[4] A.-C. Boury-Brisset. Ontology-based approach for information fusion. In *Information Fusion, 2003. Proceedings of the Sixth International Conference of*, volume 1. 2003.

[5] N. A. Bradley and M. D. Dunlop. Toward a multidisciplinary model of context to support context-aware computing. *Human-Computer Interaction*, 20:403–446, 2005.

[6] R. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6, 2006.

[7] A. M. Collins and E. F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

[8] F. Crestani. Application of spreading activation techniques in information retrieval. *Artif. Intell. Rev.*, 11(6):453–482, 1997.

[9] M.-C. de Marneffe and C. D. Manning. *Stanford typed dependencies manual.* Stanford Natural Language Processing Group, `http://nlp.stanford.edu/software/dependencies\_manual.pdf`, September 2008.

[10] Department of the Army. *Counterinsurgency (FM 3-24)*, December 2006.

[11] A. K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1):4–7, February 2001.

[12] P. Dourish. What we talk about when we talk about context. *Personal Ubiquitous Computing*, 8(1):19–30, 2004.

[13] M. Duckham and M. Worboys. Automated geographical information fusion and ontology alignment. *Spatial Data on the Web: modeling and management*, pages 109–132, 2007.

[14] G. Gross, R. Nagi, and K. Sambhoos. Soft information, dirty graphs and uncertainty representation/processing for situation understanding. In *Proceedings of the 13th International Conference on Information Fusion*, 2010.

[15] D. L. Hall and J. Llinas. An introduction to multisensor data fusion. In *Proceedings of the IEEE*, volume 85(1), pages 6–23, January 1997.

[16] D. L. Hall, J. Llinas, M. McNeese, and T. Mullen. A framework for dynamic hard/soft fusion. In *Proc. 11th Int. Conf. on Information Fusion, Cologne, Germany*, 2008.

[17] J. Hassell, B. Aleman-Meza, and I. B. Arpinar. Ontology-driven automatic entity disambiguation in unstructured text. *Lecture Notes in Computer Science*, 4273:44, 2006.

[18] M. Kandefer and S. C. Shapiro. A categorization of contextual constraints. In A. Samsonovich, editor, *Biologically Inspired Cognitive Architectures, Technical Report FS–08–04*, pages 88–93. AAAI Press, Menlo Park, CA, 2008.

[19] M. Kandefer and S. C. Shapiro. An f-measure for context-based information retrieval. In G. Lakemeyer, L. Morgenstern, and M.-A. Williams, editors, *Commonsense 2009: Proceedings of the Ninth International Symposium on Logical Formalizations of Commonsense Reasoning*, pages 79–84, Toronto, CA, 2009. The Fields Institute, Toronto, CA.

[20] M. M. Kokar, C. J. Matheus, K. Baclawski, J. A. Letkowski, M. Hinman, and J. Salerno. Use cases for ontologies in information fusion. In *Proceedings of the Seventh International Conference on Information Fusion*, pages 415–421. 2004.

[21] B. Kurz, I. Popescu, and S. Gallacher. FACADE - a framework for context-aware content adaptation and delivery. In *Proceedings of Communication Networks and Services Research*, pages 46–55, 2004.

[22] E. Little, K. Sambhoos, and J. Llinas. Enhancing graph matching techniques with ontologies. In *Information Fusion, 2008 11th International Conference on*, pages 1–8, 2008.

[23] J. Llinas. New challenges for defining information fusion requirements. In V. V. Popovich, M. Schrenk, and K. V. Korolenko, editors, *Information Fusion and Geographic Information Systems Proceedings of the Third International Workshop*. Springer Berlin Heidelberg, 2007.

[24] A. S. Maida and S. C. Shapiro. Intensional concepts in propositional semantic networks. *Cognitive Science*, 6(4):291–330, 1982.

[25] C. J. Matheus, M. M. Kokar, and K. Baclawski. A core ontology for situation awareness. In *Proceedings of the Sixth International Conference on Information Fusion*, pages 545–552. 2003.

[26] P. Mylonas, D. Vallet, P. Castells, M. Fernandez, and Y. Avrithis. Personalized information retrieval based on context and ontological knowledge. *The Knowledge Engineering Review*, 23(Special Issue 01):73–100, 2008.

[27] National Geospatial-Intelligence Agency. NGA: GNS Home. `http://earth-info.nga.mil/gns/html/index.html`, 2010.

[28] A. Peñas and E. Hovy. Semantic enrichment of text with background knowledge. In *Proceedings of FAM-LbR*. NAACL, 2010.

[29] M. A. Pravia, O. Babko-Malaya, M. K. Schneider, J. V. White, C.-Y. Chong, and A. S. Willsky. Lessons learned in the creation of a data set for hard/soft information fusion. In *12th International Conference on Information Fusion*, 2009.

[30] RDF Working Group. Resource Description Framework (RDF). `http://www.w3.org/RDF/`, January 2010.

[31] J. G. Romero, J. Garcia, M. A. Patricio, J. M. Molina, J. Llinas, S. C. Shapiro, M. Kandefer, and M. Prentice. Strategies and techniques for use and exploitation of contextual information in high-level fusion architectures. In *Proceedings of the 13th International Conference on Information Fusion*, 2010.

[32] K. Sambhoos, J. Llinas, and E. Little. Graphical methods for real-time fusion and estimation with soft message data. In *Proceedings of the 11th International Conference on Information Fusion*, pages 1–8, 2008.

[33] S. Shapiro. SNePS: A logic for natural language understanding and commonsense reasoning. In M. Iwanska and S. C. Shapiro, editors, *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. AAAI Press, Menlo Park, 2000.

[34] S. C. Shapiro. Cables, paths and "subconscious" reasoning in propositional semantic networks. In J. F. Sowa, editor, *Principles of Semantic Networks*, pages 137–156. Morgan Kaufmann, San Mateo, CA, 1991.

[35] S. C. Shapiro and W. J. Rapaport. SNePS considered as a fully intensional propositional semantic network. In N. Cercone and G. McCalla, editors, *The Knowledge Frontier*, pages 263–315. Springer-Verlag, New York, 1987.

[36] A. N. Steinberg and G. Rogova. Situation and context in data fusion and natural language understanding. In *Proceedings of the 11th International Conference on Information Fusion*, pages 1–8, 2008.

[37] The University of Sheffield. *Developing Language Processing Components with GATE*, 5.2-snapshot (development) edition, February 2010.

[38] W3C. Defining N-ary Relations on the Semantic Web. In Natasha Noy and Alan Rector, editors, *W3C Working Group Note 12*. `http://www.w3.org/TR/swbp-n-aryRelations/`, April 2006.

[39] C. Welty, J. Fan, D. Gondek, and A. Schlaikjer. Large scale relation detection. In *Proceedings of FAM-LbR*. NAACL, 2010.

[40] A. Zaenen. *Survey of Human Language Technology*, chapter Language Analysis and Understanding, pages 95–123. Press Syndicate of the University of Cambridge, 1997.