Extracting Semantic Information Structures from Free Text Law Enforcement Data

Johnson, James R. (Bob); Miller, Anita
ADB Consulting
Carson City, Nevada, USA
james_r_johnson@earthlink.net

Khan, Latifur; Thuraisingham, Bhavani
University of Texas at Dallas
Richardson, Texas, USA
lkhan@utdallas.edu

Abstract — A detective distributes information on a current case to his law enforcement peers. He quickly receives a computer generated response with leads identified within hundreds of thousands of previously distributed free text documents from thousands of other detectives. The challenges lie in the nature of free text – unstructured formats, confusing word usage, cut-and-paste additions, abbreviations, inserted html/xml tags, multimedia content, and domain-specific terminology. This research proposes a new data structure, the semantic information structure, which encapsulates the extracted content information on classes of information such as people, vehicles, events, organizations, objects, and locations as well as the contextual information about the connections and measures to enable prioritization of files containing related pieces of content. The structure is organized to be a result of automated natural language processing methods that extract entities, expanded entity phrases and their links which are driven by ontologies, DLSafe rules, abductive hypotheses and semantic composition. Importance and significance measures aid in prioritization.

Keywords – semantic information structure; semantic content; semantic context; law enforcement, free text, ontology, abductive reasoning; expanded entity phrase, related information of interes,.

I. INTRODUCTION

This paper proposes a powerful new data structure, the semantic information structure, to capture semantic content and related semantic context created from extracted entities, expanded entity phrases, ontological rules, abductive hypotheses, and extracted syntactic links [1], [2], and [3]. This research is part of a three-year research project lead by the University of Texas at Dallas. The research arose in response to a need to identify information of interest in an email sent by a law enforcement detective followed by a search of a massive number of historical law enforcement emails to find information that could help the detective resolve the case, as described in [1]. Key word searches and ontology-enhanced searches were shown to be helpful but highly inadequate. The semantic information structure is a key component of the overall project in that it builds on the expanded entity phrases described in [3] and extended by an ontology in [1]. Reference [1] introduced, but did not fully define semantic content items and semantic context items defined in this paper. It also includes measures of importance and significance to support prioritization of results. Reference [4] introduced three concepts: (1) syntactic context as a group of linguistic text; (2) semantic context as a physical entity or situation; and (3) pragmatic related context as links to a reason or purpose for distinguishing a section of text. The semantic content defined in the next section is closely aligned with the semantic context definition in [4] and the semantic context is defined in this paper as a linkage of entities based on the ontological relationships. This semantic context definition appears to provide a more global structure than the definition in [4].

This work was sponsored by the Air Force Office of Scientific Research under Contract FA-9550-09-1-0468.

II. SEMANTIC CONTENT

A semantic content item is defined in [3] as the union of an entity phrase and associated attributes.

The attributes can be thought of as refining the semantic content. Semantic content can be explicitly stated or inferred. For example, an inferred attribute can result if the model of a vehicle is mentioned but not the make. A database of makes and models of vehicles is maintained for this purpose.

This research adds a quantitative measure to the semantic content to capture the importance of each attribute. Importance is specific to the domain and is an important component of the semantic information structure. The importance factor is used to facilitate quantitative comparison and prioritization between semantic information structures [5].

It was recognized that some attributes were far more important than others. For example, discovering that a person had brown eyes is not very valuable information because brown is the dominant eye color in humans. Finding that the person had a tattoo on her left arm would be a more important extracted attribute because of the uniqueness of the attribute.

The information for a given attribute is weighted by its importance to the domain expert. The importance can be determined in two ways: (1) by using the inverse of the number of potential values as a factor and (2) by domain expert specification. The inverse of the number of potential values approach was used for this research but experienced law enforcement officers validated the approach by reviewing representative prioritizations.

TABLE I. IMPORTANCE VALUES FOR EXAMPLE ATTRIBUTES

Vehicle Attribute Type	# Potential Values	Importance (Inverse of # Potential Values)
Make	68	.015
Model	48	.021
Color	50	.020
LicensePlateNumber	1	1.000
License Plate State	50	.020

For example, if a semantic content is defined by the ontology-defined Vehicle class, the extracted attribute categories are Make, Model, Color and LicensePlateNumber,. Importance values can be assigned to attributes associated with a motorized vehicle as shown in the Table I.

III. SEMANTIC CONTEXT

The semantic context item serves two purposes:

- adding contextual information to the semantic information structure and greatly enhance the identification of related information of interest in free text, and
- prioritizing results.

A. Definition of Semantic Context

Reference [6] proposed a functional definition for semantic context item as a function of task, ontology and data. This definition assumes a task is driving the context as would be the case in a task-driven user interface environment. A somewhat similar situation to that in [6] has been observed, however. The most valuable semantic context is frequently centered on an event or crime such as the burglary of a motor vehicle event where a person and a location can be associated with the crime. Non-event focused context can be important as well. An example would be a person about 6' in height wearing a grey hoodie associated with a white Chevrolet at a number of crime scenes. In this case, there is no event, only an association.

Selection of the central entity around which the semantic context is focused is determined by identifying the entity with the largest number of outgoing links to other entities.

Borrowing some terms from graph theory, the number of links outgoing from an entity, i, can be called its out-degree, $R^{out}(i)$. Likewise the total number of incoming links from an entity is called its in-degree, $R^{in}(i)$. Define the residual-degree, r(i) as

$$r(i) = R^{out}(i) - R^{in}(i). \quad (1)$$

If the residual-degrees of entities are sorted from maximum to minimum, then the entity with the maximum residual-degree is selected as the initial focus entity.

A semantic context item is defined by the following:

A semantic context item is the set of all links and secondary links from an entity with the maximum residual-degree .(4)

An example graph of the semantic context items from a document is shown in Fig. 1. The Event node was selected as the initial focus entity to its large residual degree of 5.

IV. SEMANTIC INFORMATION STRUCTURES

Previous research has been published on the representation of semantic information. Examples include [7], [8], and [9]. Reference [7] establishes a semantic representation model by using a template knowledge base. The system can only understand the sentences that match to a semantic node in the semantic representation model, so the semantic node needs to include a description of knowledge feature. Conceptual graphs have been developed in [8] as an intermediate language for mapping natural language questions and assertions to a relational database. They are a visual representation of text extracted from its predicate calculus formulation. Reference [9] extends conceptual graphs to semantic graphs consisting of nodes, links and direction of links where nodes (entities) are represented by a rectangle, links are relations between nodes. The nodes can be labeled with one or more semantic features, and a link is labeled with a semantic case. The semantic graphs support representation of words, sentences and larger texts. The current research extends these earlier works to include (1) ontologies incorporating domain knowledge and processes, (2) inferred relationships expanding the semantic context, and (3) measures of semantic content to support relatedness assessments.

The definition of a semantic information structure is:

A semantic information structure is a data structure consisting of semantic content and their associated importance measures, and semantic context with their associated significance measures. (2)

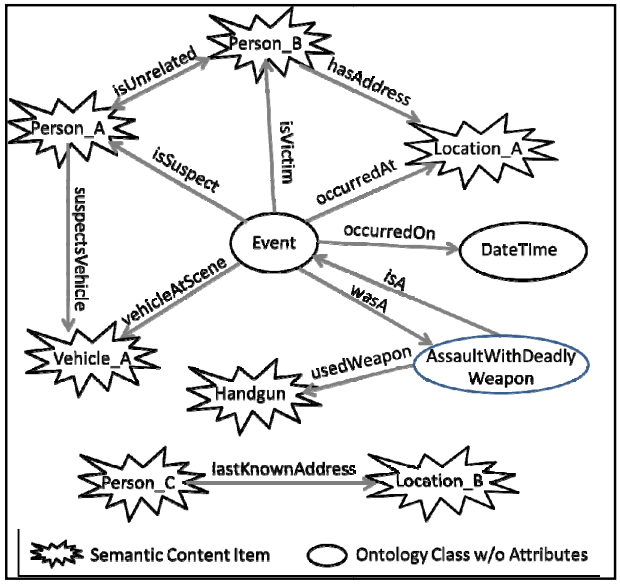

Figure 1. Semantic context links connect the event with a vehicle, person, another person, a location, a crime, and a weapon. There is a large semantic context item centered around the Event class node.

The data structure of the semantic information structure is shown in Table II. In the implementation, the location of the extracted data is preserved so as to allow semantic tagging for search as described in [10], [11], and [12]. Note that this approach is more powerful than key word tagging because it contains contextual relationships and inferred information to both expand the context and broaden the search space.

V. EVALUATING THE EXTRACTION OF SEMANTIC INFORMATION STRUCTURES

Over 1000 links, 600 semantic content items and 300 semantic context items were evaluated against ground truth. The entities and links were extracted from sanitized emails from a large law enforcement confidential LISTSERV network connecting nearly 1000 local, state, and federal law enforcement agencies over 4 years. Each data file contained the email "To", "From", "Subject", "Date" and the "Body". In addition, many emails have attachments which are integrated into the text body message.

This dataset was purposefully selected because of the extreme messiness of the data which lead to a thorough test of the algorithms. The results were captured in a spreadsheet for subsequent analysis of the statistics for recall and precision. The ground truth evaluation objectives were to quantify the recall and precision of the semantic information structures.

The results in Table III show good performance for semantic content items upon which much of the relatedness between documents will be based. While the performance is poorer for semantic context items, there is a basis for aiding in prioritization of results. Room for improvement was observed during the collection of performance measures, especially in extracting certain entities and links from syntactical patterns, and these will be implemented during the next phase. Refinements are also planned for the extraction techniques. Examples include improving inaccuracies due to mixing between neighboring semantic content and addressing the failure to extract attribute links that cross prepositional boundaries and lists using compositionality techniues such as described in [13] and [14].

TABLE II. DATA STRUCTURE FOR THE SEMANTIC INFORMATION STRUCTURE

Content	Description
Semantic Content Item (may be several in a document)	[0]: Entity phrase tied to an ontology class [1]: Anchor terms [2]:Attribute type and value … [n]: Attribute type and value
Semantic Content Importance (one per attribute)	$x_{1d}\|x_{2d}\|\ldots\|x_{nd}$ a vector representation of the semantic content measures for each attribute of each semantic content item
Semantic Context Item (may be several in a document)	[0]: From entity [1]: Link function [2]: Linked entity [3]: Sentence containing from entity [4] Position in sentence [5]: From entity type [6]: Linked entity type
Semantic Context Significance (one per semantic context item)	[0]: Importance associated with from entity [1]: Importance associated with linked entity [2]: Importance associated with link function [3]: Link function type

TABLE III: GROUND TRUTH ASSESSMENT RESULTS

	Recall	Precision
Semantic Content	94%	71%
Semantic Content Importance Measure	100%	68%
Semantic Context	50%	46%
Semantic Context Significance Measure	100%	40%

REFERENCES

[1] J. Johnson, A. Miller, L. Khan, "Law enforcement ontology for identification of related information of interest across free text documents," European Intelligence and Security Informatics, Athens, September 2011

[2] J. Johnson, A. Miller, L. Khan, B. Thuraisingham, and M. Kantarcioglu, "Identification of related information of interest across free text documents," IEEE Intelligence and Security Informatics, Beijing, July 2011.

[3] J. Johnson, A. Miller, L. Khan, B. Thuraisingham, and M. Kantarcioglu, "Extraction of expanded entity phrases," IEEE Intelligence and Security Informatics, Beijing, July 2011.

[4] J. Sowa, "Syntactic, Semantics and Pragmatics of Contexts," from AAAI Technical Report FS-95-02, 1995.

[5] J. Johnson, A. Miller, L. Khan, "Relatedness measures across law enforcement free text documents," European Intelligence and Security Informatics Conference (submitted), Odense, August 2012.

[6] M. Janik, A. Nanda, and R. Rabbani, "Semantic context specification," Report University of Georgia, http://lsdis.cs.uga.edu/~mjanik /presentation/20040506-SemanticContextSpecification.pdf , May 2004.

[7] Y. Chen and X. Fan, "A semantic representation model based on multi-graph," Journal of Computational Information Systems, Vol 7, Number 6, pp, 1830-1837, 2011.

[8] J. F. Sowa, "Conceptual Graphs," Handbook of Knowledge Representation, ed. by F. van Harmelen, V. Lifschitz, and B. Porter, Elsevier, pp. 213-237, 2008.

[9] L.Hébert, "Dispositifs pour l'analyse des textes et des images," Limoges, Pulim, 282 pages, Publié en ligne le 5 décembre 2007.

[10] H. Hedden, "How semantic tagging increases findability," EContent Magazine, October 2008.

[11] J. C. Mills, "Tagging and the semantic web," www.designmills.com /2008/05/20/tagging-in-the-semantic-web, May 20, 2008.

[12] V. Miličić, "Case study: semantic tags," http://www.w3.org/2001 /sw/sweo/public/UseCases/Faviki/, December 2008.

[13] S. A. McDonald and C. Brew, "A distributional model of semantic context effects in lexical processing," Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004.

[14] J. Johnson, A. Miller, L. Khan, "Graphical Representation and Graph Matching for Semantic Information Structures," To be submitted.