Extraction of Expanded Entity Phrases

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

Khan, Latifur; Thuraisingham, Bhavani; Kantarcioglu, Murat
University of Texas at Dallas
Richardson Texas, USA

Abstract— **This research is part of a larger integrated approach for extraction of information of interest from free text and the visualization of semantic relatedness between phrases of interest. This paper defines a new structure which is a key component, the expanded entity phrase (EPx). This paper also presents an approach for extracting EPx's from free text. The structure of the EPx's facilitates quantitative comparison with other EPx's. A combination of part of speech-based template matching and ontology-driven NLP provides an effective technique for extracting complex entity structures that cross clause boundaries. This approach also uses ontology-based inferences to lay the ground work for linking EPx's for semantic relatedness assessments involving different named entities not explicitly stated in the text. The real world data used in this research were derived from a collection of law enforcement email messages submitted by hundreds of investigators seeking information or posting information about crimes, incidents, requests, and announcements. Performance data on the approaches used for extracting EPx's and links from this data are presented.**

Keywords- information of interest; natural language processing; expanded entity phrase; semantic relatedness

I. INTRODUCTION

This research is part of an integrated approach for extraction of information of interest and visualization of relatedness [1]. In the integrated approach, related information of interest between a selected document and archival holdings is extracted. Visualization of these related items allows a topical focus and drill-down of specific information even when a document predominantly contains other topics. Text mining of information across documents enables sharing between investigators working, perhaps unknowingly, on related cases.

This paper defines expanded entity phrases and presents an approach for extracting expanded entity phases from free text. Expanded entity phrases (EPx) are entity phrases with associated attributes identified from ontology classes and their properties. This vector-like structure provides a means for quantitative measures against other EPx's. A combination of part of speech-based template matching and ontology-driven NLP provides an effective technique for extracting complex entity structures that cross clause boundaries. This approach also uses ontology-based inferences to lay the ground work for linking EPx's for relatedness assessments involving different named entities.

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

The test data used in this research are a collection of sanitized law enforcement email messages submitted by hundreds of investigators and crime analysts for a large metropolitan region. The primary goal is not to determine the similarity of emails but to find related information of interest within the emails. The data are complicated by the inclusion of a high percentage of law enforcement terminology and abbreviations in an informal email communication style. Most of the communications lack the punctuation typically encountered in English text, forcing the detection of entities that are adjacent or in lists by detecting contextual changes in topics and leading indicators.

The next section describes the technical challenges encountered in extraction of expanded entity phrases. Section III provides foundational definitions and provides a description of the approach steps studied for extraction of expanded entity phrases as well as end-to-end performance information. The conclusion is provided in Section IV.

II. TECHNICAL CHALLENGES

A. Challenges in Extracting Entity Phrases

The key technical issues concerning the extraction of entity phrases include:

- Handling domain-specific jargon and abbreviations – Most law enforcement jargon and product items are not included in a basic Thesaurus. Also specialized domain-specific jargon, acronyms and abbreviations are typically not included in Thesauri. Law enforcement emails typically lack target words as used in FrameNet applications [2]. Reference [3] authors have advocated enhancements to Wordnet, however, this presents issues for sharing and long-term maintenance because of the extensive customization.
- Inadequacy of anchor terms – Anchor terms are useful for identifying a foothold into a candidate entity phrase. With lack of punctuation and liberal use of capitalization, acronyms, and abbreviations, anchor terms may not be detectable and extraneous terms may be included into entity phrases. AutoSlog, [4], utilized conceptual anchor point heuristics to identify sentence components (subject, verb-types, and preposition-types) to extract a linguistic pattern of interest. This approach requires a domain-specific heuristic database to handle a large range of expected patterns. Further, new patterns that are not in the database will likely be missed. The linguistic patterns in the Riloff approach

are quite short, typically no more than three words. Hence lists, items in parentheses, and multi-clause patterns are excluded. Other approaches extend this technique of slot-based heuristics and operate on grammatical free text. Some of these approaches extract an exact match while others use syntactic matching [5]. Syntactic matching has shown poor performance with the data considered in this research due to its frequent omission of punctuation, use of all capital letters, and use of domain-specific acronyms.

- Handling lack of punctuation – Lack of punctuation causes problems in recognizing sentence boundaries and words within lists. This causes uncertainty in the location of multiword lists such as a person's name followed by a city name. Reference [6] applied tree kernels to model a syntactic parse tree. However, the lack of punctuation and overuse of capitalization in the law enforcement data result in errors and uncertainty in part of speech tagging, thus rendering the tree kernels ineffective.

Other complexities may include handling text (unstructured and semi-structured) that has been pasted into a free text document; handling an indirect reference to an event or requiring deduction to extract entities from partial information about an object, organization or person; identifying attributes associated with an entity phrase occurring in subordinate phrases; handling poorly stated phrases leading to a possible misinterpretation of the author's intent; handling misspellings or corruptions in the text; and minimizing loading resulting from complex algorithm elements.

B. Challenges in Forming Expanded Entity Phrases

This section discusses the technical challenges associated with the extraction of connected items (both attributes and associated entity phrases) for a specific expanded entity phrase. An entity may be explicitly stated or implied by associated attribute. The larger constructs (larger than an entity phrase) require identifying syntactical and semantic relationships embodied in free text. Technical challenges include:

- Understanding entities in context – Entities generally occur in a logical order within the information flow of the text. An entity can be semantically related to a preceding or following entity. Understanding the context of these entities allows incorporation of semantic content. Reference [7] developed a framework for modeling coherence of entities in discourse. The authors evaluated the relation of coherence to the inference demands of different types of referring expressions. Simple entities were utilized.

- Identifying description qualifiers – Description qualifiers can occur before or after an entity phrase, providing additional information about that entity phrase. Placing description qualifiers with the entity phrase enhances the semantic context. Reference [8] utilized linguistic patterns to extract entity types. They employed extended regular expressions utilizing Subject-Verb-Object macros to match entity types such as an event. They also utilized WordNet relations. Identifying Subject-Verb-Object contexts has increased complexity, because the data typically lacks punctuation, use of capital letters, and incorporates domain-specific acronyms that are not included in Thesauri or WordNet.

- Associated entity linking – Identifying associated entities provides context between pairs of entities. A technical challenge is the determination of whether or not two entities should be linked. Current link analysis software tools are capable of visualizing criminal networks but offer little structural analysis resources [9]. Structural patterns include the identification of central groups, subgroups and interaction patterns between groups. Reference [10] presented an approach to associated entity linking through the application of abductive reasoning over ontologies as a means of extracting links between entities through a measure of salience. This incorporates the strengths of ontologies to capture semantic structures between entities.

III. APPROACH FOR EXTRACTING EXPANDED ENTITY PHRASES

In this section, definitions and examples are provided as well as processes for extracting entity phrases and forming expanded entity phrases. Performance results are presented.

A. Definitions and Examples

An entity phrase is taken to be an ordered (in occurrence) near-contiguous (within the same clause) sequence of nouns (primarily) describing a person, a place, an organization, an event, a date, or an object. An expanded entity phrase is quite different from a keyphrase, term, named entity, or collocation in that it is tied to an ontology class which has attributes identified by properties of that class. The extracted text items are linked to the ontology elements through a Thesaurus using entries corresponding to each ontology class. A term is equivalent to a word. A named entity is a term which is categorized into one of certain groups. An entity phrase can consist of one or more terms. Noun collocations correspond to unnamed entity phrases [11].

An expanded entity phrase (EPx) is the union of an entity phrase and associated attributes with the following format:

$$EPx = (EPT{:}t,\ W_a,\ P_1{:}A_1,\ P_2{:}A_2,\ \ldots,\ P_n{:}A_n) \qquad (1)$$

where EPT = entity phrase tied to an ontology class, W_a = anchor word(s) if they exist, and A_i = i^{th} attribute (value) associated with the i^{th} property P_i.

Examples of expanded entity phrases are shown in Table I. Note the absence of punctuation and the broad use of capital letters and acronyms which are common in the test data. Note that the terms in the EPx are converted from abbreviations using the Thesaurus.

TABLE I. EXAMPLE SENTENCE FRAGMENTS AND THEIR ASSOCIATED EXPANDED ENTITY PHRASES

Sentence Fragment from Law Enforcement Data	**Extracted EPx's**	**Entity Extraction Type**
Smith, Everest Lee WM 03/12/1992 5'6", 115lbs …	{EP:Date, dateValue:03/12/1992}	Date
Montebello Police Department was out with two subjects in Busby''s parking lot …	{EP:Organization, organizationValue:Montebello Police Department}	Organization
… mail dfreeman@elapdcom BULLETIN #09 - 103 Inv Assistant K Robertson #609 March 13, 2009 SOFIA REYES DOB: 7/18/1991 CA DL# 27151579 …	{EP:Person, personValue:K Robertson, title:Investigator Assistant}	Titled person
… LKV: Dark Blue '01 Toyt 4- runner "MWN5721 …	{EPT:Possible vehicle, AchorTerm:LKV, Color:Dark Blue, Make:Toyota, Model:4-Runner, LicensePlateNumber: MWN5721}	Vehicle (Anchor Term)

A semantic link, L, between two expanded entity phrases is a triple where the first and third elements are the two expanded entity phrases and the second element is the functional relation between the ontology classes mapped to those expanded entity phrases. For example, let C_1 and C_2 be two ontology classes which map to corresponding expanded entity phrases EPx_1 and EPx_2. Further, let $R(C_1,C_2)$ be a relation between C_1 and C_2. Then the link between EPx_1 and EPx_2 is given by

$$L = (EPx_1, R(C_1,C_2), EPx_2). \quad (2)$$

As an example, suppose C_1 = Passenger, and C_2 = Automobile with a relation passengerOf between the two classes. Further suppose the extracted expanded entity phrases are

EPx_1 = {EP:Passenger, firstName:John, lastName:Doe} and
EPx_2 = {EP:Automobile, TriggerTerm:Suspect Vehicle, Make:Dodge, Color:Red}
=> $L = (EPx_1$, passengerOf, $EPx_2)$ (3)

where detailed attributes contained in each expanded entity phrase provide linked metadata. Inheritance provided by the hierarchical class structure and associated properties can supply other links extending the semantic content associated with these expanded entity phrases.

B. *PROCESSES FOR EXTRACTING ENTITY PHRASES*

The processes of extracting entity phrases are grouped by the type of approach used to extract them. The first three approaches have been studied previously in [12], [8], and [5]. The fourth approach is a new approach. It utilizes templates on a part of speech (POS) string derived from the POS tags combined with properties identified from an ontology.

Pattern-based extraction utilizes regular expressions to match string patterns. For example, telephone numbers have specific patterns such as ddd-ddd-dddd or ddd-dddd, where d is a digit from 0 to 9. The dash can be replaced by no delimiters or by other delimiters. Dates can exhibit a variety of formats, including the standard structure mm/dd/yyyy or yyyy/mm/dd. Other formats such as Jan 23, 2011 or just January can be encountered. Regular expressions are very successful at extracting patterns as exhibited in phone numbers, dates, times, and specific numerical types such as social security numbers.

Trigger terms in combination with patterns enable both extraction of the entity and determination of its type. For example, in law enforcement, CALP can be an abbreviation for California License Plate. It is typically followed by a string containing both alpha characters and numbers, such as 123ABC. There could also be other characters such as spaces and dashes separating groups of characters. Regular expression matching can allow for these additional characters.

Trigger term based growth is employed for more complex entity phrases, typically, organizations and people with titles. The trigger words and abbreviations are defined in the Thesaurus corresponding to the class and properties being searched. Titles of people can include Dr., President, Chief, Deputy, Officer, and Investigator. These terms are typically followed by the actual person's name tagged by a noun and beginning with a capital letter. Candidate names are checked against a U. S. Census database of names to augment the confidence level.

The anchor terms to slot technique first searches for anchor terms that can occur at the beginning (leading indicator), middle (embedded indicator), or end (second embedded indicator) of the expanded entity phrase. This approach is suitable for complex objects such as vehicles, events and addresses where many terms comprise an expanded entity phrase. The anchor terms are specified by ontology classes of interest and their associated Thesaurus property terms (synonyms, acronyms, and abbreviations). The positional indicators provide boundaries from which to search for specific property values. In this research the ontology is a law enforcement ontology described in [13].

Words can be mapped into one or more entity phrases depending on the context. For example, "General Electric" might be interpreted either as a high ranking military officer or as a large corporation. The order in which entities are searched affects the extraction and naming performance. Entity phrases involving trigger terms or common patterns are searched first because they are simpler and can be pulled out of consideration. Then the more complex entity phrases involving anchor terms are searched and pulled out of consideration. The person names and other terms tagged as nouns are searched in the remaining text. The sequence in which the entity phrases are searched is the following:

a. Pattern of trigger term (date of birth, any date, license plate, drivers license, email, social security number)

b. Pattern-based growth (organization, person with title)

c. Anchor terms to slot (vehicle, address, and event)

d. Person name

e. Other noun

The remainder of this paper focuses on entity phrases extracted via anchor terms and their links to the employed ontologies.

C. PROCESSES FOR EXPANDING ENTITY PHRASES USING ANCHOR TERM BASED APPROACH

The anchor terms to slot technique assumes the potential for an EPx with a number of attributes, although they do not necessarily have to exist. These anchor terms are identified from a Thesaurus based upon the particular ontology class mapped to the entity phrase name. The ontology provides a list of classes which serve as entries to the Thesaurus. Cases are established if none, one, or more of the anchor terms are present. With each case a baseline process has been developed to search for entity phrase attributes.

The domain-specific Thesaurus is established with 1) ontology class name, 2) definitions, 3) parts of speech, 4) synonyms, 5) acronyms, and 6) abbreviations. The primary term is equal to class name in an associated ontology. In this research, a law enforcement-focused Thesaurus has been created along with a law enforcement ontology. The reason for a domain-specific Thesaurus is that regular Thesauri do not contain most domain-specific law enforcement terms, nor do they contain terms pertaining to product names such as Toyota or Camry. Code violations are often acronymed or abbreviated. Examples are assault with a deadly weapon (ADW), health and safety code (HS), felony (fel) and misdemeanor (misd). Many common expressions are shortened as well. Examples include Automated Fingerprint Identification System (AFIS), be on the lookout (BOLO), gone on arrival (GOA), last known address (LKA), green (grn), suspect vehicle (S/V), victim (vic) and Victoria (Vic).

At a high level, the process for expanding entity phrases using the anchor term to slot technique consists of the following steps:

a. Detect sentences in the free text

b. Parse into major clauses and minor fragments

c. Process POS tags for each term including punctuation

d. Create POS string from POS tagged text

e. Search for anchor terms and process to slots for selected entity phrase types

f. Construct candidate EPx's

g. Search POS string and pattern for entities and links

h. Update expanded entity phrases using discovered links

These steps are illustrated in an example below. A brief discussion of parsing the sentences into major and minor fragments and performing the POS pattern search are provided in the next two paragraphs.

Parse into major and minor fragments – Sentences (fragment, simple, or complex) can have zero, one or more than one entity phrases. If the number is zero, then the sentence can be discarded from further consideration. If a sentence has exactly one entity phrase, then the attributes can be used to refine the naming of the entity phrase and construction of the EPx. For sentences with more than one entity phrase, construction of links between these entity phrases can refine their naming. The complex sentences are parsed into major clauses using the major cut terms such as "where", "which", "that", "before" and a semicolon. Minor fragments are delineated using minor cut terms such as "for", "and", "to", and a comma.

Search POS string with a pattern – A POS string is constructed from POS tags of the words in a sentence, removing less important tags by assigning them as an "other" category. A database of search patterns based on POS substrings is used to search for linkage patterns and remaining entities within the simple sentences, fragments and major clauses. These patterns can cross minor but not major clause boundaries. The search starts with the longest substring pattern and works down to the shortest substring pattern. The longest matches are preferred over the shorter matches if a pattern subsumes a shorter pattern.

The rest of this section illustrates the process for the following text: "Has anyone worked a scam involving two black males with heavy African accents with the pretext of one being robbed by a cab driver that then goes to a family member was kidnapped and he had to dispose of the inheritance before he returns to Africa?".

The tagged POS elements for this text are <vbz>Has</vbz> <nn>anyone</nn> <vbd>worked</vbd> <det>a</det> <nn>scam</nn> <vbg>involving</vbg> <cd>two</cd> <jj>black</jj> <nns>males</nns> <in>with</in> <jj>heavy</jj> <nnp>"African"</nnp> <nns>accents</nns> <in>with</in> <det>the</det> <nn>pretext</nn> <in>of</in> <cd>one</cd> <nn>being</nn> <vbn>robbed</vbn> <in>by</in> <det>a</det> <nn>cab</nn> <nn>driver</nn> <in>that</in> <rb>then</rb> <vbz>goes</vbz> <to>to</to> <det>a</det> <nn>family</nn> <nn>member</nn> <vbd>was</vbd> <vbn>kidnapped</vbn> <cc>and</cc> <prp>he</prp> <vbd>had</vbd> <to>to</to> <vb>dispose</vb> <in>of</in> <det>the</det> <nn>inheritance</nn> <in>before</in> <prp>he</prp> <vbz>returns</vbz> <to>to</to> <nnp>Africa</nnp> <pp>?</pp>

The constructed POS string in this example is VNVONVOJNIJNNIOIONVIONNIOVOONNVVONVOVIONINVONO. Note that all noun types are mapped to N and all verb types are mapped to V. This allows searching for consecutive noun substrings representing candidate entity phrases as well as template-based searches for links between candidate entity phrases. These template-based searches (linkage patterns) are constructed from rules identifying links between subjects and predicates in text. Once an entity phrase has been identified, a block of the sentence can be removed from consideration in the search for another entity phrase. Thus a second pass can be made to detect links that do not actually connect two entity phrases, but link an entity phrase and a verb or adjective.

Each section of Fig. 1 shows a text segment followed by the mapping of parts of speech to the corresponding part of speech string. The algorithm uses keywords to identify breaks in major and minor clauses listed in the three boxes. Several part of speech patterns were found (NVON, NIJNN, NVIONN, NNVV, NVOVION, and NVON) by substring matching. The links resulting from these substring matches are shown in the bottom half ((anyone, worked, scam), (males, with, African accents), (being, robbed by, cab driver), (family member, was kidnapped, null), (he, had to dispose of, inheritance), and (he returns to, Africa). Note that some words are missing and this complicates the process. For example, "family member was kidnapped" should be "family member who was kidnapped". Further it is difficult to assign an identity to "he" in the last two links. It could be interpreted that the family member had to dispose of inheritance as part of the scam, or one of the males had to dispose of the inheritance.

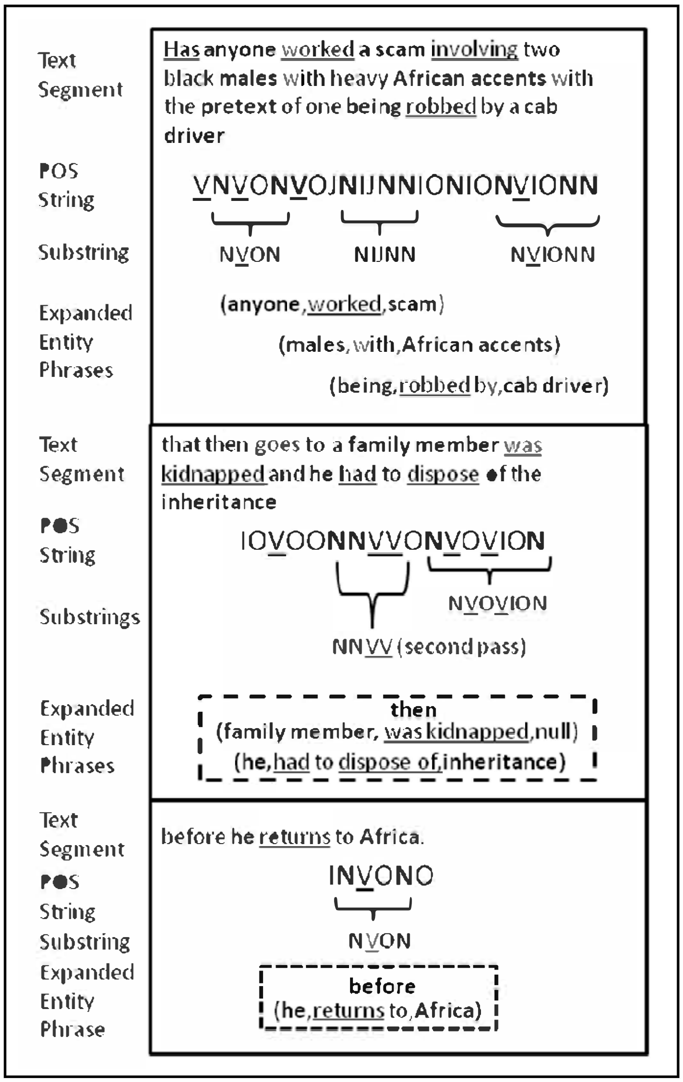

Figure 1. Example extracting entity phrases and links using POS templates

Expanded entity phrases are formed from the individual entity phrases, surrounding adjectives linked to ontology properties, and the links between the entity phrases. EPx's extracted from the example on the previous page are the following:

EPx1 = {EP:male, AnchorWord:males, startPos:6, endPos:12, number:2, race:black, accent:heavy African}

EPx2 = {EP:object, AnchorWord:scam, startPos:4, endPos:4}

EPx3 = {EP:Person, AnchorWord:cab driver, startPos:22, endPos:23}

EPx4 = {EP:object, AchorWord:inheritance, startPos:40, endPos:40}

EPx5 = { EP:country, AchorWord:Africa, startPos:45, endPos:45}

startPos and endPos denote the starting and ending positions within the sentence. Note for the complex EPx1 the position spans 7 characters and is located just after the entity "scam". Further analysis using the major clause breaks "then" and "before" shown in Fig. 1 can infuse semantic context relating these entity phrases.

D. PRELIMINARY RESULTS

An experiment was set up to process over 800 law enforcement documents and to extract both entity phrases and expanded entity phrases as an initial test of the NLP methods without the application of the full supporting ontology. This experiment utilized a minimal ontology with basic classes of person, vehicle, license plate, drivers license, but without rules or a Thesaurus. These preliminary results were then compared to manually extracted items. A second experiment using a full law enforcement ontology driven methodology and over 2000 emails is being presented in another paper [13]. The performance for entity extraction are measured by recall and precision. For this application, recall is defined as the *number of relevant entities extracted divided by the total number of existing entities*, while precision is defined as the *number of relevant entities extracted divided by the total number of entities extracted.*

The test data consisted of sanitized law enforcement emails notifying or requesting information on criminal incidents. The emails pertained to a range of topics including notification of thefts and major crimes, requests for information such as mug shots, and notifications of meetings, and courses offered. This provided a base of data that represents a real-world situation, one where emails are written quickly and conversationally with many incomplete sentences, pasted-in information, misspelled words, broad use of capitalization, law enforcement specific terminology, as well as heavy use of law enforcement acronyms, slang terms, and abbreviations.

Ground truth data were manually collected for each entity phrase in the documents. Statistics were compiled for each entity phrase type as well as for the total of all entity phrase types (Fig. 2 and Fig. 3). Not all rules have been implemented for identifying relevant entities, affecting the precision results.

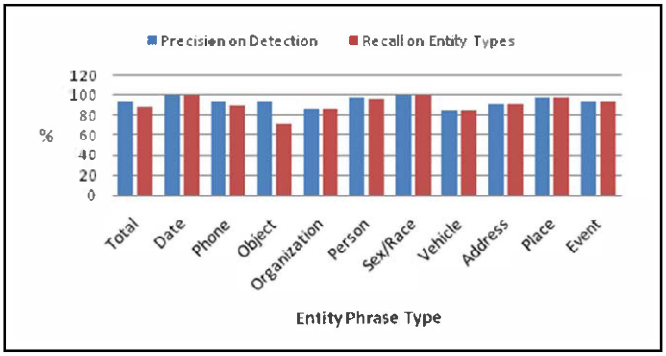

Figure 2. Preliminary results for entity phrase extraction from free text law enforcement data

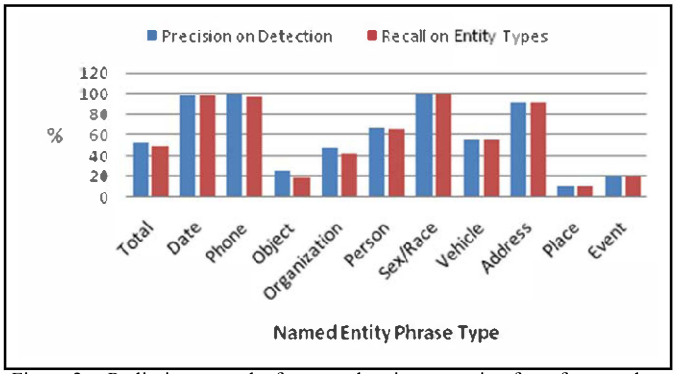

Figure 3. Preliminary results for named entity extraction from free text law enforcement data

Note that the performance is dependent on the type of entity phrase such that pattern-based approaches, for example, score very high. More complex entity phrases are more difficult to identify in the complex environment of this particular data.

Note that in extracting the named entity phrases, the less complex types yield higher performance as expected. Entity phrases that exhibit more consistence patterns, such as addresses, perform better than those with highly variable structures, such as vehicles and places. It is expected, and will be tested in phase 2, that the methodology should yield much improved results if the text includes punctuation and full sentences.

The implications of this experiment is that NLP approaches are promising for the poor quality of data supplied. The named entities results were quite poor suggesting that the inclusion of the ontology and Thesaurus structures will improve that aspect. Comparison to performance by previous NLP methods is currently not possible due to the nature of the data tested and the fact that data similar to those experiments have not been tested.

Preliminary results for extraction of expanded entity phrases was carried out for vehicle and address classes only. The precision and recall for addresses was 66 and 63 percent, respectively. The precision and recall for vehicles was 61 percent in both cases. The remaining issues to work are that not all of the properties belonging to an EPx are successfully attached.

IV. CONCLUSION

This research has defined a new construct encapsulating entity phrases and their attributes. This expanded entity phrase is identified by integrating ontology structures with a domain-specific Thesaurus which handles the acronyms, synonyms, and abbreviations found in real-world free text.

A new POS string-based pattern search method was also presented. This method is important for pulling out complex information that spans clauses and complex information that is embedded in lists. It is also used to extract links between entities for the purpose of refining their entity types and for providing semantic information in link structures.

The next steps in this research are to 1) extend the inference rules of an ontology enabling the discovery of links between expanded entity phrases and 2) address event types, which are very challenging, through abductive reasoning over the ontology.

REFERENCES

[1] 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, in press.

[2] A. Moschitti, P. Morăreascu, and Harabagíu, "Open domain information extraction via automatic semantic labeling," Proceedings of the 2003 Special Track on Recent Advances in Natural Language at the 16th International FLAIRS Conference, 2003.

[3] R. Basili, M. Cammisa, and A. Moschitti, "A semantic kernal to classify texts with very few training examples," Informatica, vol 30, pp 163-17, 2006.

[4] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," Proceedings of the Eleventh National Conference on Artificial Intelligence, AAAI Press / MIT Press, pp. 811-816, 1993.

[5] I. Muslea, "Extraction patterns for information extraction tasks: A survey," Proceedings of the AAAI '99: Workshop on Machine Learning for Information Extraction, 1999.

[6] F. Zanzotto, M. Pennacchiotti, and A. Moschitti, "A machine learning approach to textual entailment recognition," Natural Language Engineering, vol 15 (4), pp. 551-582, 2009.

[7] B. Grosz, A. Joshi, and S. Weinstein, "Centering: a framework for modeling the local coherence of discourse," Computational Linguistics 2(21), pp. 203-225, 1995.

[8] S. Harabagiu, M. Surdeanu, and P. Morărescu, "Automatic discovery of linguistic patterns for information extraction," FLAIRS-01 Proceedings, American Association for Artificial Intelligence, 2001.

[9] J. Xu and H. Chen, "CrimeNet Explorer: A framework for criminal network knowledge discovery," ACM Transactions on Information Systems, Vol. 23, No. 2, April 2005.

[10] C. Davenport and R. Hill, "Fast abductive reasoning over ontologies," Proceedings of the American Association for Artificial Intelligence 2006 Fall Symposium, 2006.

[11] C. Manning and H. Schutze, "Foundations of statistical natural language processing," MIT Press, May 1999.

[12] C. Chang, and S. Lui, "IEPAD: information extraction based on pattern discovery," ACM 1-58113-348-0/01/0005, 2001.

[13] J. Johnson, A. Miller, L. Khan, B. Thuraisingham, and M. Kantarcioglu, "Law enforcement ontology for identification of related information of interest across free text documents", unpublished.