View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by NASA Technical Reports Server

Source of Acquisition
NASA Ames Research Center

Semantic Analysis of Email Using Domain Ontologies and WordNet

Daniel C. Berrios[1], Richard M. Keller[2]

[1]University of California, Santa Cruz, MS 269-2,
NASA Ames Research Center, Moffett Field, CA USA 94035
[2]Intelligent Systems Division, MS 269-2,
NASA Ames Research Center, Moffett Field, CA USA 94035
{berrios,keller}@email.arc.nasa.gov

Abstract. The problem of capturing and accessing knowledge in paper form has been supplanted by a problem of providing structure to vast amounts of electronic information. Systems that can construct semantic links for natural language documents like email messages automatically will be a crucial element of semantic email tools. We have designed an information extraction process that can leverage the knowledge already contained in an existing semantic web, recognizing references in email to existing nodes in a network of ontology instances by using linguistic knowledge and knowledge of the structure of the semantic web. We developed a heuristic score that uses several forms of evidence to detect references in email to existing nodes in the SemanticOrganizer repository's network. While these scores cannot directly support automated probabilistic inference, they can be used to rank nodes by relevance and link those deemed most relevant to email messages.

1 Introduction

In the last decade, a revolution has occurred in the way most computer users communicate. Electronic mail, once the purview of only sophisticated users, has become the *de facto* way many users exchange textual information. The problem of capturing and accessing knowledge in paper form has been supplanted by a problem of providing structure to vast amounts of electronic information. Most emailed communications cannot be retrieved in a timely manner, because even if they are accessible, there are no prevalent automated methods that characterize their content. Without such methods, finding the right piece of information in email corpora will become more and more difficult as the volume of emailed information continues to increase.

The development of collaborative knowledge management tools is an important step towards faster, more precise information browsing and retrieval; such tools create a central point of knowledge capture and access [1-3]. Users of such tools can build semantic webs of richly structured knowledge. For example, a user can explicitly specify links between a measurement and an experiment, a set of experiments and a scientific project, and the project and its participants. These semantic links provide important contextual information when browsing a semantic web, and, we contend, could be used to search and retrieve electronic documents such as email messages more rapidly than through current methods.

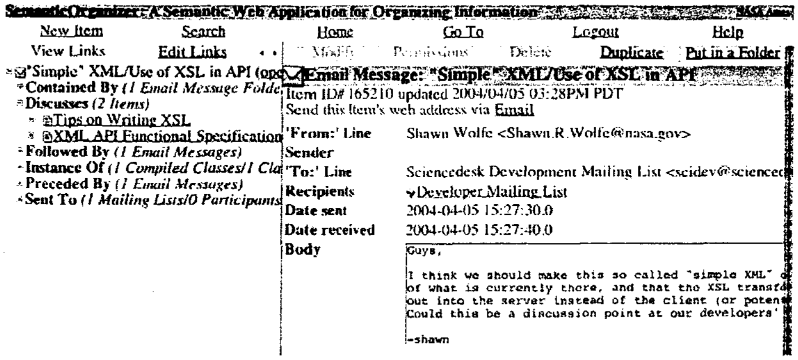
However, relying on humans to create a rich semantic web that includes electronic documents like email without substantial support is unrealistic [4]. There are several aspects of workflow that frequently limit the ability and willingness of authors to add structured semantic information to their documents:

- Domain information overload: As authors include more and more information in a semantic network, manually linking every new email message to all other relevant nodes becomes too time-consuming.
- Domain model complexity: If the number of potential types of links between email messages grows, it will becomes progressively more difficult for authors to specify the correct relationship between an email communication and other information.
- Insufficient domain knowledge: On some collaborating teams only certain users have sufficient domain knowledge to make appropriate semantic links between email messages and other information.
- Lack of technical sophistication: Some users will always lack the technical sophistication required to generate appropriate semantic links for email.

Systems that can construct semantic links for natural language documents like email messages automatically (e.g., [5]) will be a crucial element of semantic email tools for them to achieve any significant level of penetration outside research environments. We have developed a system that automatically extracts information from electronic text documents like email messages, and can be used to link email into an existing semantic web of information. The difficult problems of natural-language information extraction and understanding have been studied extensively, yet the precision and recall of general-purpose information-extraction systems have rarely exceeded 70% [6]. However, the vast majority of these systems were evaluated by extracting examples of knowledge (for example, the location and date of a terrorist attack) which a priori were completely unknown to the system. We have designed an information extraction process that can leverage the knowledge already contained in an existing semantic web, recognizing references in email to existing nodes in a network of instances from an ontology by using linguistic knowledge and knowledge of the structure of the semantic web. In addition, we suggest ways in which the system could present the knowledge it infers from email or other text documents to users.

2 Problem

In our experience certain types of email messages, such as those generated in the workplace, often have substantial content that can be "matched" semantically with instances in an ontology, if one exists. For example, an email message that contains the phrase "Baja Field Trip 2005" could be linked to an instance of a "field trip" in a domain ontology with the label "Baja California, Spring 2005." Such links could



"Simple" XML/Use of XSL in API (op
Contained By *(1 Email Message Folde*
Discusses *(2 Items)*
Tips on Writing XSL
XML API Functional Specification
Followed By *(1 Email Messages)*
Instance Of *(1 Compiled Classes/1 Cla*
Preceded By *(1 Email Messages)*
Sent To *(1 Mailing Lists/0 Participants*

Email Message: "Simple" XML/Use of XSL in API
Item ID# 165210 updated 2004/04/05 03:28PM PDT
Send this Item's web address via Email

'From:' Line	Shawn Wolfe <Shawn.R.Wolfe@nasa.gov>
Sender	
'To:' Line	Sciencedesk Development Mailing List <scidev@scienced
Recipients	Developer Mailing List
Date sent	2004-04-05 15:27:30.0
Date received	2004-04-05 15:27:40.0
Body	

```
Guys,

I think we should make this so called "simple XML" o
of what is currently there, and that the XSL transf
out into the server instead of the client (or poten
Could this be a discussion point at our developers'

-shawn
```

Figure 1. Use of SemanticOrganizer to view, archive, and semantically link email messages. An email message (shown in the right pane), is linked to nodes in the SemanticOrganizer ontology, including two documents relating to its content ("Tips on Writing XSL" and "XML API Functional Specifications", show on the left pane), its "Sender", its recipients, the mailing list to which is was sent, and preceding and following messages.

provide users reading the email with valuable insight regarding the meaning of the phrase, and conversely, could provide browsers of the instance with important details regarding the field trip and links to related information. The process of linking emails in this manner would be relatively straightforward for domain experts to perform manually, but tedious for users who exchange large volumes of email. Thus, the problem is how to perform such an analysis efficiently and automatically.

We explored several methods for analyzing email messages to uncover evidence of references to nodes (instances) in a semantic web, with the ultimate goal of linking the email to those nodes. We chose to develop and test our system using email messages sent to users of SemanticOrganizer, a web-based application that allows users to create networks of knowledge and data linked together by binary relations [3]. SemanticOrganizer users can set up project-related (or other type) mailing lists, and the system not only re-distributes these messages to list members via electronic mail, it creates nodes representing each message, and links these nodes to the mailing list node, as well as to individual sender and recipient nodes (if they exist, matching on email address property). Email message nodes already account for nearly half of the nodes in the semantic network, yet most of these nodes have no semantic links to the various equipment, experiments, mission activities, etc. that they discuss (Figure 1). Through the methods we present below, we hope to provide to a user who is browsing an email message node in SemanticOrganizer with meaningful links to other nodes in the network that have been deemed "relevant" to the message. Conversely, we hope to be able to support question-answer functions, such that a user could retrieve important supporting documentation when posing a query in terms of ontology concepts and relational contexts.

3 Approach

Our approach first involved pre-processing email messages, stripping out signature lines and embedded "quoted" email messages using some simple heuristics (similar to those developed in [7]). Next, we analyzed message bodies using a Hidden-Markov model syntactic parser [8] that identifies verb and noun phrases in the text. We then calculated a heuristic score that uses several forms of evidence of references to existing nodes in SemanticOrganizer for each node in the network. While heuristic scores cannot directly support automated probabilistic inference, these scores can be used to rank nodes by relevance and link those deemed most relevant to the email message (i.e., those above a specified threshold).

To calculate the heuristic scores, we considered direct and indirect evidence; we considered evidence as direct if it was based on comparing text in the email messages with a node's metadata (including its name – its label for display -- and type). Direct evidence relied only on noun phrase terms identified in the email. Indirect evidence included all other types of evidence, such as references to related nodes and how they may be related. We based indirect evidence largely on probabilities of contexts for nodes. These contexts were extracted using both noun and verb phrase terms in the email. Both indirect and direct evidence used linguistic knowledge contained in the English language taxonomy-thesaurus WordNet (v. 1.7, [9]).

3.1 Direct Evidence

To capture direct evidence using node property values, we developed and compared two methods: simple lexical matching of node attributes and weighted cosine similarity. In the former method, we merely scanned the email for strings matching node names and metadata (i.e., property values). We ignored meta-characters and nodes with common noun names (by look up in WordNet). We considered partial matches of node properties, splitting property values into tokens by simple heuristics, and weighting matches by the fraction of matched tokens. We also transformed names of persons ("Smith, John" to "John Smith") in both node names and email to improve matching. We then calculated an attribute score, σ_A, for a node A :

$$\sigma_A = L + \sum_{j=1}^{m} \frac{v_{Aj}}{m} \qquad (1)$$

where L is the number of exact matches to node A's name in the email, and v_{Aj} is the number of occurrences of each of A's m literal property values. This ad-hoc measure has the attractive feature of being able to recognize references to multi-word phrases in email that are part of node metadata.

Table 1 shows the top 20 attribute matching scores (and L component contribution) after analysis of an email with subject line "March Baja Filed Trip"

Table 1. Measuring direct evidence of nodes in email messages. Attribute-scores (s) are based on lexical matches to node names (L) and metadata.

	Attribute Scoring	
Node	σ	L
Brad's Objectives for May-June Trip	6.4	6.4
Nitrogen Headspace Measurements	6.1	6.0
Oxygen Headspace Measurements	5.8	5.7
Pond 2 B	5.3	5.3
BAJA-Gross Photosynthesis Measurements	5.3	5.3
Brad's Objectives for Fall 2	5.0	5.0
Tori's Objectives for Fall 2	5.0	5.0
Pond Survey	5.0	5.0
The Trip Down - Baja Pass.JPG	4.8	4.8
Scott's Field Camera	4.3	4.3
Pond 4 rep B	4.0	4.0
Pond Survey Data	3.7	3.7
Des Marais, David	3.7	3.7
Ecogenomics Focus Group	3.3	3.3
Pond 4 Near 5	3.3	3.3
Pond 5 Near 6	3.3	3.3
Motel Temperature Diel Methane Fluxes	2.9	2.8
Methane Fluxes 10-22-02 Diel	2.6	2.5
del13C methane – greenhouse	0.6	0.6
Bo's Sulfate Reduction	0.25	0.0

(sic). In this domain (or for this particular email message), most instances were referred to exactly, although a few matched only partially. For only 1 node, "Bo's Sulfate Reduction," were there no matches to any of the node's name tokens (only to its metadata).

The attribute scores lack at least one desirable feature: a heavier weighting of

Table 2. A portion of the mapping of node types to WordNet.

Type	Sense No.	Synset No.	Synonyms
institution	1	6689622	establishment
equipment	1	2869748	instrumentation, instrumentality
figure	1	5852382	fig
document	1	5421657	written document, papers
document	4	5452954	text file
light	1	9433880	visible light, visible radiation
oxygen	1	12366142	O, atomic number 8

matches to uncommon vs. common terms. We have experimented with calculating a weighted cosine similarity score using the textual metadata of nodes, which we considered a virtual "document", and to which we compared the text of email message bodies using standard term vector comparison methods [10]. The same type of comparison can also be done using node name term vectors and message subject line terms, and the result of the two comparisons combined to yield a term vector score. Ideally this type of analysis should be dovetailed with the attribute scanning score to capture multi-word or even phrases similarities between node metadata and message text.

We also included as direct evidence any references in email messages to node type (domain ontology class), or a synonym of the type. This required the generation of a mapping of node types in SemanticOrganizer to synsets in WordNet. We generated such a mapping for 252 types of instances in SemanticOrganizer (Table 2). Interestingly, mapping to more than one synset was required for 59 types in order to include all possible meanings for the types. For each node type, t, we used this mapping to compute a synonym score:

$$syn_A = \frac{syn_t}{n} \qquad (2)$$

where n is a recognized noun or noun phrase in the same synset as the type, and m is the number of different synsets containing n. We then distributed these type synonym scores equally over the n instances of each type:

$$syn_t = \sum_n \frac{1}{m} \qquad (3)$$

to yield a synonym score for each node in the network.

We observed that email authors also occasionally referred to a node in SemanticOrganizer by using a hyponym of the node's type. For example, researchers at ASU frequently referred to the node named "Arizona State" of type "institution" as "the University" rather than "the Institution." This was due in part to lack of specificity of the SemanticOrganizer ontology (creating many types that are highly specific renders search and creation of instances more time consuming and cognitively difficult). In order to try and capture such references, we scanned email messages for terms that are hyponyms of terms in the synsets we had mapped to node types. We could then use a measure of "semantic distance" to weight this type of evidence:

$$semanticproximity(synsetA, synsetB) = \prod_{k,m} \frac{1}{k^m} \qquad (4)$$

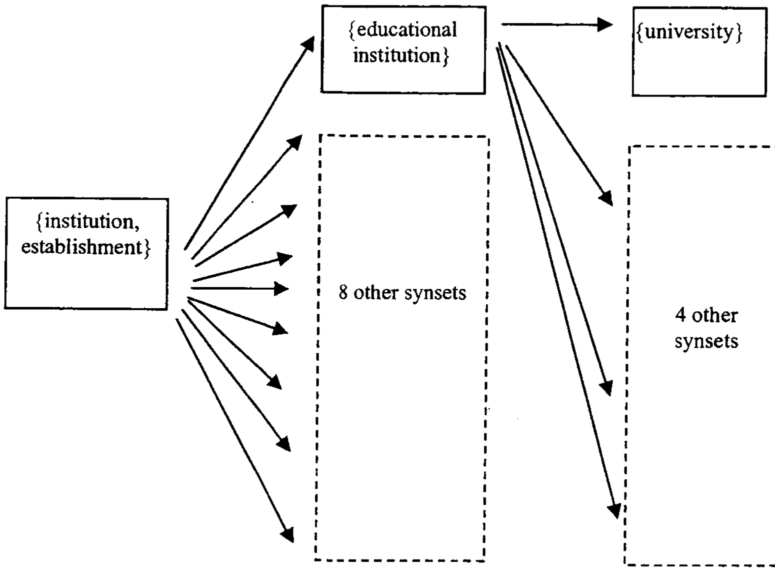


Figure 2. An illustration of the calculation of semantic proximity of the hyponym "university" to the synset containing the terms "institution" and "establishment." The semantic proximity of the term to the node is (1/9)(1/5)2 or 0.0044.

where k is the number of different synsets on each of the m levels of the hyponym tree between synsets A and B (Figure 2).

We calculated semantic proximity to each node type, t, for all nouns and noun phrases found in email, and summed them to yield a type hyponym score:

$$hyp_t = \sum_{t,h} semanticproximity(t,h) \qquad (5)$$

where h ranges over the set of identified noun synsets in the email. After computing these scores for all types of nodes, we again then distributed scores evenly amongst all n instances of a scored type:

$$hyp_A = \frac{hyp_t}{n} \quad . \qquad (6)$$

Finally, we combined all forms of direct evidence from these scores into a single direct evidence score, D_A, for any node A as:

$$D_A = \sigma_A + syn_A + hyp_A \quad . \tag{7}$$

3.2 Indirect Evidence

The binary semantic relations that connect nodes in SemanticOrganizer are a potentially valuable source of knowledge when analyzing email messages. For example, if an email message contains the phrase "water volume taken per sampling time," the use of the term "taken" could refer to some existing link in SemanticOrganizer based on the relation "collected-at(*sample*, *aqueous-site*)." The existing link is part of the context in which the concepts *sample* and *aqueous-site* are understood in the ontology [11]. In other words, the quantity *P(relation r/term t),* if known, could be used to indicate the likelihood that the linked ontology concepts are discussed in an email message. We could estimate *P(r/ t)* in terms of the following probabilities that use conditional probabilities of WordNet synsets:

$$P(r \mid t) = \sum_{s_i \in S} P(r \mid s_i) \bullet P(s_i \mid t) \quad . \tag{8}$$

where S is the set of WordNet synsets. The conditional probabilities could be obtained from semantically annotated, domain-specific corpora, with or without machine learning. However, we did not have such annotated corpora available. Instead, we assumed a uniform distribution when estimating P(s_i|t), so that,

$$P(s/t) = 1/t_{ss} \tag{9}$$

where t_{ss} is the number of synsets containing t. To estimate P(r|s_i), we mapped 262 SemanticOrganizer relation types to WordNet synsets (see Figure 3). We then again assumed a uniform distribution of $P(r/s_i)$ over the set of r_{ss} relations to which s_i was mapped, so that

$$P(r|s_i) = 1/r_{ss} \, . \tag{10}$$

We included in this quantity, evidence from entailing (but not entailed) verb synsets (as defined through entailment and/or hyponym relationships in WordNet). For example, we considered the verb "negotiate" as evidence of the relation "discuss", since negotiation entails discussion.

Substituting these estimates into (8) yields a **relational context probability** for each relation type:

$$P(r \mid t) = \sum_{s_i \in S} \frac{1}{r_{ss} t_{ss}} \tag{11}$$

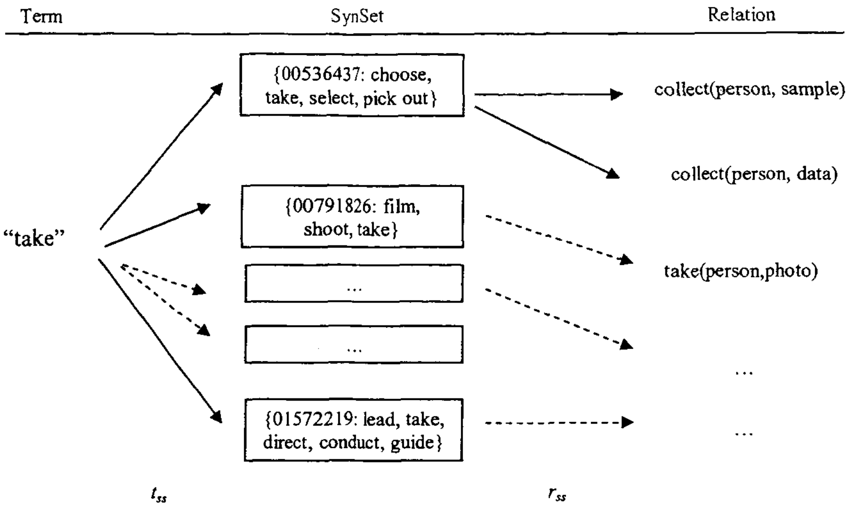

Term
SynSet
Relation
"take"
{00536437: choose, take, select, pick out}
{00791826: film, shoot, take}
...
...
{01572219: lead, take, direct, conduct, guide}
collect(person, sample)
collect(person, data)
take(person,photo)
...
...
t_{ss}
r_{ss}

Figure 3. Estimating P(relation|termEach term recognized in email is a member of t_{ss} synsets in WordNet, each of which we manually mapped to r_{ss} relations in SemanticOrganizer.

which we calculated and summed over the set of all terms, *T*, detected in verb phrases of email message, so that:

$$P(r) \approx \sum_{T} \sum_{s_i \in S} \frac{1}{r_{ss} t_{ss}} \quad . \tag{12}$$

The challenge remained as to how to combine this evidence of binary relations in email messages with the direct evidence of nodes described above. Nodes in SemanticOrganizer can be linked to any number of other nodes (i.e., relations have no cardinality restrictions). For a given link *l* between nodes *A* and *B*, *A* has *k* other links based on the same relation (to other nodes), and *B* has *m* other such links. Assuming equal probability of any of these links given evidence of the relation on which they were based, we estimated

$$P(l_{AB}) \approx \frac{P(r)}{(k+m)} = \frac{\sum_{T} \sum_{s_i \in S} \frac{1}{r_{ss} t_{ss}}}{(k+m)} \tag{13}$$

and used this quantity as an indirect evidence score, I_A for node A.

3.3 Examples

In order to validate the reasoning behind the node scores, have analyzed a set of email messages exchange amongst a group of collaborating astrobiologists. Our goal was first to inspect and informally validate case results of the analysis. For example, consider the following excerpts from a single email message in this domain [emphasis added]:

> "Accordingly, we will be performing biogeochemical **measurements** and collect **samples** for **various** analyses back at our respective **laboratories**.
>
> ...
>
> Key intermediates in the biogeochemical sulfur cycle (B. **Thamdrup**) ...
>
> ...
>
> 5) Survey **measurements** in pond (locations to be determined):
> ? salinity
> ? temperature
> ? dissolved oxygen (Winkler **titration**)
>
> Cores for vertical profiles:
> ? Number of diel periods: 2 (once in **Pond 4 near 5**, once in Pond 5 near 6)
> ? Number of cores: 3 (near or under each of the chambers **used** for the flux **measurements**)"

And the nodes, A, B, and C in the SemanticOrganizer ontology:

	Attribute	Value
A	name	Bo's sulfate reduction
	type	measurement
	data reduced by	Thamdrup, Bo
	measurement device	various
	measured for	EMERG
B	name	Pond 4 near 5
	type	aqueous site
	latitude gps	27o 41.3450' N
	study area	Baja
	approx. water depth	1
	site marker	yes

C	name	SSX Microscopy Lab
	type	laboratory
	institution	NASA Ames Research Center
	lab manager	Blake, David
	lab users	Bebout, Leslie; Blake, David; Kato, Katharine
	lab description	Electron microscopy facility TEM SEM AFM STM
	location	Building 239, Room B30

As shown in Table 3, Node A has a moderate σ because much of the node's meta-data is mentioned in the email ("measurement", "various", "Thamdrup"). It has high *syn* and *hyp* scores because its type and many hyponyms ("sampling", survey", "titration") of the node's type are mentioned. Node B has a high σ because it is mentioned explicitly by name several times in the email, along with references to other attributes ("aqueous site", "Baja"), and to its type (yielding a modest *syn* score). Node C is of interest because contextual information for the node is detected. For example, its type, "laboratory", is linked to persons mentioned in the email through the relation "lab users" that is detected in the email via the term "use" (among other contexts detected for the type).

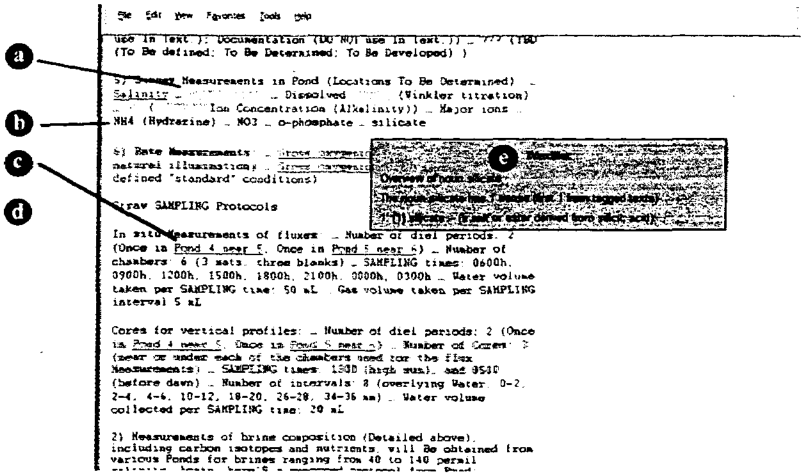
node		σ	*syn*	*hyp*	*I*
A	Bo's Sulfate Reduction	0.32	2.0	0.44	0.33
B	Pond 4 near 5	2.37	0.67	0.0	0.0
C	SSX Microscopy Lab	0.0	2.0	0.0	0.11

Table 3. Example Indirect and Direct Evidence Scores for three nodes in the SemanticOrganizer ontology.

node		σ	*syn*	*hyp*	*I*
A	Bo's Sulfate Reduction	0.32	2.0	0.44	0.33
B	Pond 4 near 5	2.37	0.67	0.0	0.0
C	SSX Microscopy Lab	0.0	2.0	0.0	0.11

3.4 Semantic Email Developer Interface

We have begun to develop and evaluate a user interface for exploring the results of the type of semantic analysis of email messages described above (Figure 4). This interface was initially designed to help develop analysis methods (by indicating the basis for score calculations), although certain features of the interface could prove of value for a semantic email application. For example, through the interface, users can read email messages with embedded links to nodes in the SemanticOrganizer ontology, and see WordNet glosses for nouns and verbs when they place the mouse over terms in the email message.



use In Text): Documentation (DO NOT use In Text.)) _ ??? (TBD (To Be defined: To Be Determined: To Be Developed))

5) [illegible] Measurements in Pond (Locations To Be Determined) _ Salinity _ [illegible] _ Dissolved [illegible] (Winkler titration) _ [illegible] ([illegible] Ion Concentration (Alkalinity)) _ Major ions _ NH4 (Hydrazine) _ NO3 _ o-phosphate _ silicate

6) Rate Measurements: _ Gross [illegible] natural illumination) _ Gross [illegible] defined "standard" conditions)

Straw SAMPLING Protocols

In situ Measurements of fluxes: _ Number of diel periods: 2 (Once in Pond 4 near 5. Once in Pond 5 near 6) _ Number of chambers: 6 (3 sets. three blanks) _ SAMPLING times: 0600h. 0900h. 1200h. 1500h. 1800h. 2100h. 0000h, 0300h _ Water volume taken per SAMPLING time: 50 mL . Gas volume taken per SAMPLING interval 5 mL

Cores for vertical profiles: _ Number of diel periods: 2 (Once in Pond 4 near 5. Once in Pond 5 near 6) _ Number of Cores: 3 (near or under each of the chambers used for the flux Measurements) _ SAMPLING times: 1330 (high sun), and 0530 (before dawn) _ Number of intervals: 8 (overlying Water. 0-2. 2-4. 4-6. 10-12. 18-20. 26-28. 34-36 mm) _ Water volume collected per SAMPLING time: 20 mL

2) Measurements of brine composition (Detailed above). including carbon isotopes and nutrients. will Be obtained from various Ponds for brines ranging from 40 to 140 permil

Overview of noun silicate

The noun silicate has 1 sense (first 1 from tagged texts)

1. (1) silicate -- (a salt or ester derived from silicic acid)

Figure 4. A GUI for presenting results of semantic email analysis. An email message body is shown on the left, below its subject line. Terms in the message body have been styled to indicate results of the analysis: (a) synset terms (e.g., "Temperature", above) mapped to relations and types are highlighted; (b) terms are expanded into parenthetical lists inserted following the term ("pH", "NH4") using a lookup table of organizational acronyms and abbreviations); (c) Exactly matching instances names are hyperlinked to the instance in SemanticOrganizer; (d) terms that provide evidence of types or relations through hyponymy and entailment are highlighted (on mouse-over action, not illustrated here); (e) on mouse-over of a term, details of hyponymy, synonymy, and entail information regarding the term, as well as WordNet information are shown (in this case, the glosses for the term "silicate" are shown).

4 Related Work

Our work echoes methods reported by the ArtEquAKT project [5] which used linguistic knowledge together with domain ontologies to infer knowledge from natural language documents. However the ArtEquAKT system relies solely on WordNet knowledge of the expression of ontological relationships using natural language terms. We feel the wide variation in expressing formal relationships in email and other natural language documents, requires, at a minimum the type of human-generated mapping of these relationships to language that we have described. In addition, the use of specialized terms in many domains limits the use of general thesauri like WordNet to discern meaning in email messages.

The problem we sought to address shares some features with the problem targeted by applications of latent semantic indexing [12, 13]. LSI attempts to find patterns of terms that indicate a semantically important index for documents. Similarly, we seek to find patterns of terms that indicate instances and contextual relationships in the SemanticOrganizer ontology. It may be that a more extensive group of linked instances is a more appropriate index for some email messages (than individual nodes), and LSI could have a valuable role in discovering these types of indexes. There is also some overlap with work to retrieve documents supporting a natural language hypothesis [14], inasmuch as an ontological concept and its relational context may be propositional representations of such hypotheses.

In the area of email message understanding, the results of applying machine learning methods to discern and extract a set of five speech acts (which could be modeled as instances in an ontology) from email appear promising [7]. Of course, machine learning requires annotated test corpora, which are frequently not available. The authors of [7] also investigated the hypothesis that using links between email messages can improve understanding of individual messages themselves (and vice versa). This approach could potentially be of value in the kind of analysis we have reported.

5 Discussion

We report here our attempts to develop an information extraction application to analyze email message for semantic content. We focused on identifying instances from the ontology of our collaborative knowledge management application, SemanticOrganizer. Our methods combine elements of linguistic analysis via WordNet, statistical term models, and probabilistic reasoning, and we present some illustrative examples of their application.

There are several obvious shortcomings to the methods we have described. Both indirect and direct evidence scores have assumed uniform distributions for term senses. With the use of semantically annotated domain-specific corpora, we could instead employ domain-specific probability distributions and likely improve both types of scoring. Also, the relational context probability we have proposed would likely be more accurately calculated if it included a distance measure of terms related to the extracted contexts, with assumption being that terms closer together in email messages are often more likely part of the same context than terms further apart.

We plan to do a more rigorous evaluation of our approach by comparing results to human-annotated message corpora. While evaluating the precision (the frequency with which identified instances are deemed relevant or correct) of the methods we propose could be relatively straightforward, estimating recall could be problematic. Some domains in SemanticOrganizer have hundreds or even thousands of instances; the ability to identify missed references in email messages to instances from such a large space of knowledge could be beyond the capabilities of domain experts. Scoping the evaluation domain to an appropriate set of domain instances will be critical.

The problem of identifying ontology instances in email messages can be generalized to include other types of natural language documents. The types of links between messages and instances we seek to discover with our work could be of value for documents from a myriad of domains, including medical histories, engineering requirements, business plans, etc. The problem can also be viewed as one small aspect of a more general, difficult problem: how should the information in a natural language document alter knowledge in an ontology? Automated analysis of a document could lead to the discovery of new instances, to deleting instances, or to altering properties of instances. Such analyses could be a vital component of systems that automatically understand web documents, and act on their content through semantic web services.

Acknowledgements. We gratefully acknowledge funding support by the NASA Intelligent Systems Project and by the NASA Engineering for Complex Systems Program. Some of this work was supported by the Universities Space Research Association. We thank Brad Bebout and Dave Blake for use of data in illustrations. Our sincere appreciation goes to our colleagues Robert Carvalho, Keith Swanson, Shawn Wolfe, Stephen Rich, Ian Sturken, and Sergey Yentus for their contributions to system development.

References

[1] O. Corcho, A. Gomez-Perez, A. Lopez-Cima, V. Lopez-Garcia, and M. Suarez-Figueroa, "ODESeW. Automatic generation of knowledge portals for Intranets and Extranets," *SEMANTIC WEB - ISWC 2003*, vol. 2870, pp. 802-817, 2003.

[2] O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez, and O. Vicente, "WebODE: an integrated workbench for ontology representation, reasoning, and exchange.," 2002.

[3] R. M. Keller, D. C. Berrios, R. E. Carvalho, D. R. Hall, S. J. Rich, I. B. Sturken, K. J. Swanson, and S. R. Wolfe, "SemanticOrganizer: A customizable semantic repository for distributed NASA project teams," presented at 3rd International Semantic Web Conference (ISWC2004), Hiroshima, Japan, 2004.

[4] D. C. Berrios, R. J. Cucina, and L. M. Fagan, "Methods for semi-automated indexing for high precision information retrieval," *J Am Med Inform Assoc*, vol. 9, pp. 637-52, 2002.

[5] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt, "Automatic ontology-based knowledge extraction from Web documents," *IEEE Intelligent Systems*, vol. 18, pp. 14-21, 2003.

[6] J. Cowie and W. Lehnert, "Information extraction," *Communications of the ACM*, vol. 1, pp. p. 80-91, 1996.

[7] R. Khoussainov and N. Kushmerick, "Email task management: An iterative relational learning approach.," presented at Proc 2nd Conference on Email and Anti-Spam, 2005.

[8] Language Technology Group, "LT POS." Edinburgh, UK: University of Edinburgh, 2003.

[9] G. A. Miller, "WordNet: A Lexical database for English," *Communications of the ACM*, vol. 38, pp. 39-41, 1995.

[10] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, pp. 513-23, 1988.

[11] A. Segev and A. Gal, "Putting Things in Context: A Topological Approach to Mapping Contexts and Ontologies," presented at Proceedings of the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications, Pittsburgh, PA, 2005.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1990.

[13] S. T. Dumais, "Latent Semantic Indexing (LSI) and TREC-2," presented at The Second Text REtrieval Conference (TREC-2), 1993.

[14] O. Glickman, I. Dagan, and M. Koppel, "Probabilistic Classification Approach for Lexical Textual Entailment," presented at AAAI '05, Pittsburgh, PA, 2005.