Fostering Collaboration with a Semantic Index over Textual Contributions

Kenneth Murray, John Lowrance, Douglas Appelt, and Andres Rodriguez

SRI International
{murray lowrance appelt acr} @ ai.sri.com

Abstract

Collaboration is at the heart of many activities required for effective homeland security, from intelligence analysis to policy formation. We are exploring new approaches to facilitating effective collaboration that remove or reduce common barriers and that exploit opportunities to encourage more effective collaboration, including transcending the cognitive biases of the participants. In order to evaluate our approaches we are developing Angler, a web-services tool that supports collaboration among participants on some focus topic. Several challenges arise in helping participants manage their contributions. A semantic index over the participant contributions is used to address these challenges.

Angler as a Collaboration Tool

Collaboration is at the heart of many activities required for effective homeland security, from intelligence analysis to policy formation. We are exploring new approaches to facilitating effective collaboration that remove or reduce its common barriers and that exploit opportunities to encourage more effective collaboration, including transcending the cognitive biases of the collaboration participants and expanding their joint cognitive vision.

To such effect we are building Angler, a web-services tool for supporting collaboration on some focus topic (Rodriguez et al. 2005). Angler facilitates effective collaboration by overcoming some of the common barriers to collaboration. It does so allowing for participants to be geographically distributed and allowing for asynchronous collaboration. This differentiates the Angler process from similar business processes that normally take place in a synchronous, face-to-face manner. Angler facilitates cognitive expansion during collaboration by using divergent (such as brainstorming) and convergent (such as clustering or ranking) thinking techniques.

Angler is organized around the idea of virtual (possibly hierarchical) workshops. Workshops are organized by a facilitator that brings a group of people together to accomplish a knowledge task. Simple workshops usually start in a brainstorming phase and then go through other phases, such as clustering and ranking. The facilitator for the workshop will enforce and manage the timelines. As more and more workshops are stored in the Knowledge Base, a Corporate Memory is formed.

In the brainstorming phase, collaboration participants are asked to contribute thoughts (or ideas) to answer a focused request from a facilitator. Participants submit contributions as brief statements on some aspect of the overall focus topic being considered; each such contributed thought is authored as a small textual document that can be reviewed by other participants. Other participants' thoughts will be incrementally disclosed to a participant, so that he or she can think independently and can benefit from others' ideas. Angler provides a convenient interface for participants to author such thoughts, review and respond to the thoughts of other participants, and organize the growing set of shared contributions in meaningful ways

After sufficient input and review, the facilitator moves the workshop into the clustering phase. The participants begin to organize the thoughts into various clusters. Each cluster suggests one or more candidate ways to coalesce individual thoughts into a coherent theme, and the participants benefit by considering the emerging themes, for example, for scenario planning (Schwartz 1991). This process promotes a rich interchange of ideas and perspectives while removing some of the classic barriers to conventional collaboration, such as the requirement that all participants must be in the same place at the same time. It also allows each participant to express his or her own abstractions and themes (as opposed to group clustering in the same room). Angler then calculates a consensus clustering based on agglomerative clustering (King 1967) techniques and calculates how aligned is the group vision. Finally, there is a consensus and ranking phase, where participants come to understand their differing views and vote on the final names of consensus clusters.

Angler has been used within several workshops (e.g., for scenario planning) and has been shown to facilitate collaboration among participants of those workshops. Table 1 presents (slightly obfuscated versions of) some thoughts contributed during a workshop considering the consequences of a hypothetical political regime change in a country (Country-X) that has nuclear weapons. These are the 23 thoughts related to nuclear security issues. A total of 140 thought contributions were made by the seven participants during this phase of the workshop. Each thought includes a pithy summarizing "Catch Phrase" as well as a concise description of the issue to be considered.

ID	Catch Phrase	Author	Description
1	Amount of fissile material	G	How much fissile material is in Country-X? In what form is it stored?
2	C2 apparatus	D	The nature of the nuclear command and control structure.
3	configuration	B	configuration of arsenal - assembled, disassembled, state of deployment
4	control over nukes	C	Does the new regime have effective control over Country-X's nuclear arsenal and facilities?
5	domestic audience	B	domestic public opinion
6	International Support	E	Level of US and international support for maintaining Country-X’s nuclear security
7	new regime nuke policy	C	Will the new regime alter Country-X's existing nuclear policies concerning deterrence, etc.?
8	Nuclear Command and Control	G	Specific structures and systems; safeguards, command authority, designation
9	nuclear labs and military	B	domestic nuclear and military corporate interests
10	Personnel Reliability	F	Extent of radical Religion-Z leanings of nuclear C2 personnel, and/or level of corruption.
11	Physical security of nuclear arsenal	G	Permissive Action Links, Hardened sites etc.
12	quality and reliability of personnel	B	nature of Religion-Z among nucledar personnel
13	Regional Developments	A	Country-X, Country-Y and stability
14	Resource allocation	D	The inclination of the new leadership to allocate resources to nuclear safety and security measures (or alternatively, force expansion/enhancement).
15	Successor regime	A	Nature of the successor regime and its willingness to maintain strong positive control
16	Technical Reliability	F	Functionality of nuclear safeguard devices.
17	Threat level	E	Who wants Country-X's nukes, either to use or to sabotage?
18	Threat perceptions	D	The nature of the new leadership's threat perceptions and how it impacts nuclear decision making.
19	US contingency plans	C	Does the US have effective contingency plans to seize control of Country-X's nuclear arsenal to prevent it from falling into hostile hands?
20	US inputs	B	US inputs into safety and security of Country-X’s arsenal
21	US intervention	A	US willingness to preempt should instability occur
22	Vulnerability to pre-emption	F	Extent of device core separation from firing technology given events surrounding succession.
23	Willingness to divert funds	E	State of Country-X's economy and willingness of govt to divert funds/budgetary allocation towards nuclear security

Table 1: Example Angler “thought” Contributions

Technical Challenges in Supporting Collaboration

One challenge that arises in collaborative contexts such as Angler involves helping participants manage the often large number of thoughts that are contributed. A rich exchange of ideas among even a relatively small group of collaborators can quickly produce a volume of thoughts that may overwhelm some participants and impair their continuing participation, obscuring “the forest for the trees.” An important observation about such sets of thought contributions is that they are not independent from each other, and semantic relations can be defined among them and used to organize them. Sometimes a thought contributed by one participant may be redundant with a thought contributed by another participant (e.g., from Table 1, thought 8 subsumes thought 2). Sometimes two thoughts share a significant amount of semantic content, and although one is not subsumed by the other the two are good candidates for merging (e.g., from Table 1 thoughts 2, 4, and 8 or thoughts 19 and 21). Establishing such semantic relations among thought contributions within a collaboration thread helps to organize the shared thoughts and helps the participants in reviewing the contributions of others and grasping how some thought contributions may converge with or diverge from other thought contributions.

A second challenge that arises in collaborative contexts such as Angler involves helping participants appreciate how extensive is the coverage of their collective contributions. Sometimes two or more thoughts may be complementary in nature, so that together they cover a natural set of possible considerations. For example, some thoughts may focus on international considerations (e.g., thoughts 6 and 13 from Table 1) while other thoughts focus on domestic considerations (e.g., thoughts 5 and 9 from Table 1). Sometimes a set of thoughts will partially cover a natural set of considerations. Sometimes a set of thoughts will fail to cover any of the considerations within some natural set (e.g., no thought in Table 1 considers treaties). Establishing such properties of the coverage of a set of contributions can reveal to the participants areas that are relatively under-considered and remain candidates for additional attention and may help collaborators overcome their personal and collective cognitive biases. Through spurring considerations of otherwise unexamined horizons during collaborative brainstorming, Angler may address one of the primary deficiencies cited by the 9/11 Commission; that is, "a failure of imagination" (9/11 Commission 2004).

A third challenge that arises in collaborative contexts such as Angler involves helping participants to perceive shared interests or complementing expertise with other participants. The content of each participant's contributions offers clues to that participant's areas of interest and expertise. For example, in Table 1 authors B and F may share an interest in the reliability of nuclear personnel affiliated with Religion-Z (thoughts 10 and 12), and authors A and F may share an interest in preemption and the successor regime (thoughts 15, 21, and 22). Identifying shared interests among participants can facilitate further collaboration between those participants, and it can help a workshop organizer concentrate or disperse interest in a topic while selecting participants.

A fourth challenge that arises in Angler involves helping participants find thought contributions. Given a topic of interest to the user, how can we identify those thoughts that may be relevant to the topic, even though that topic was not explicitly mentioned in the text of the thought? For example, a user might want to find the thoughts relevant to beliefs (e.g., opinions, preferences, religions), even though the word "belief" does not appear explicitly in any thought. Supporting such semantic search helps the participants review and organize a large and growing body of thoughts and helps policy makers review the workshop "thought process" regarding a topic of interest.

We are developing technology that responds to these four challenges by enabling Angler to automatically and semiautomatically establish semantic relations and properties over the collaboration contributions it manages and to identify shared topics of interest among the collaborators. The emerging technology is intended to provide interactive assistance: suggestions (e.g., about merging similar thoughts) are vetted by the human participants.

Overview of Our Technical Approach

We are investigating the use of an ontology (Gruber 1993) for establishing semantic properties of Angler thoughts. We can then use these properties to reason in simple but useful ways about the semantic content of the thought contributions to address the four collaboration challenges described above. For example, a substantial number of the semantic concepts of thought 2 are also included in thought 8, suggesting a high degree of semantic overlap.

Our approach is to use TextPro (Appelt and Martin 1999) to parse the textual elements of Angler thought contributions; this includes both the catch phrases and the descriptions for each thought. We then map the parsed words into concepts defined within the ontology; our ontology is implemented as a network of semantically related concepts. We then index each thought with the set of concepts from the ontology that correspond to words in the parse of the text for the thought. This index of concepts can be viewed as a thought-specific vector in the semantic space defined by the ontology.

Each concept in the ontology is related to other concepts using formally defined relationships that denote properties such as "is a type of," "is an instance of," "is a part of," and "is a sub region of." Our ontology was created manually to cover the content of the example 23 Angler thought contributions and relevant background knowledge, and it was engineered to be extendable to cover the content of new thoughts arising in new collaborations involving new domains. The ontology content has been significantly influenced by the KM Component Library (Barker et al. 2001) and by the Suggested Upper Merged Ontology (Teknowledge). KetL, the formal language used to represent the ontology, has been influenced by CycL (Cycorp) and KIF (Genesereth and Fikes 1992).

Our methodology for developing these applications emphasizes evaluation. Several transformations occur in the representation an Angler thought, and each results in an alternative representation. For each such transformation we want to be able to empirically assess the relative costs for performing that transformation, the relative benefits with respect to the quality of the resulting relations and properties established for the thoughts using the representation produced by the transformation, and the scalability issues, including considering how difficult it is to support the transformation in new domains.

The following sections discuss in detail how we represent Angler thoughts, and how we use that representation to address the four technical challenges for supporting collaboration.

Representing and Indexing Thoughts

To support the simple ways in which our applications will reason over them, the representation of the Angler thoughts are annotated with terms that represent their contents. Figure 1 presents a part of the annotated representation of an Angler thought. The original text is augmented by four types of representation terms: tokens, lexemes, concepts,

```
Angler-Thought-2
   isa : ( Angler-Thought )
   author : (Author-D)
   description-text : ("The nature of the nuclear command and control structure.")
   catch-phrase-text : ("C2 apparatus")
   ketl-concept : (Device  State-Description  Nuclear-Technology  Command-and-Control Structure)
   ketl-ancestor : (Authority  Knowledge State-Description  Agent-Role  Artifact  Structure Interaction  Command
                        Technology  Description  Control  Cognitive-Resource  Mechanism  Nuclear-Technology  Agent
                        Command-and-Control Device Resource)
   textpro-lexeme : ("c2" "apparatus" "nature" "of" "the" "nuclear" "command and control" "structure")
   textpro-token : ("C2" "apparatus" "The" "nature" "of" "the" "nuclear" "command and control" "structure")
```

Figure 1: Representing an Angler Thought

and ancestors; each type of representation term is produced sequentially by a transformation of the thought content.

Tokens are the substrings of the original Angler thought text that TextPro deems to be treated as words. In extracting tokens from the thought text, TextPro handles issues such as punctuation, hyphenation, and apostrophes. The tokens are algorithmically extracted from text.

Lexemes are the standardized versions of words extracted as tokens. In converting tokens to lexemes, TextPro handles issues such as case, tense, plurality, and proper nouns. Lexemes are entries in a manually engineered corpus of canonical word stems.

Concepts are terms defined in the ontology. The concepts of an Angler thought include the concepts associated with the lexemes extracted by TextPro from the thought. The lexemes extracted from the thought texts are mapped into concepts by using a translation table that is constructed manually as part of the knowledge engineering effort that produces the ontology

The ancestors of a concept include those terms in the ontology that are in some way implied by the concept. For example, Weapon is an ancestor of Pistol, and Africa is an ancestor of Egypt. The ancestors of an Angler thought are the unions of the ancestors of the concepts of the thought.

Table 2 summarizes the number of terms required to represent the 23 thoughts for each of the different representation types. The data in Table 2 show that in the representation of these 23 thoughts there is a natural compaction that occurs during the transformations that produce the first three of the four types of representation terms: on average, the 12 tokens appearing in a thought is reduced to 8.5 concepts. However, on average, about four additional ancestors are added to the representation of a thought for each concept. In the presence of a very large corpus of thoughts, the ancestors may be efficiently computed when needed rather than stored for each thought as presented in Figure 1.

	min	max	Total	Avg
Token	4	22	277	12.0
Lexeme	4	22	267	11.6
Concept	5	14	196	8.5
Ancestor	21	76	1003	43.6

Table 2: Representation Use Summary

Effectively, each concept defines a folder that includes the thoughts referencing that term; the folder is created as a consequence of the representation via the slot-inverse of ketl-concept. Similarly, each ancestor defines a folder containing thoughts that reference concepts deemed relevant to that ancestor term. This structure defines a semantic index for the Angler thought contributions.

Having created this semantic index, we can leverage it to reason about the Angler thought contributions in simple but useful ways that address the four technical challenges identified above. The following sections describe in detail these four applications of the semantic index.

Identifying Similar Thoughts

The first application we consider is helping participants identify when a thought (e.g., perhaps a newly authored thought) is semantically related to other thoughts. For a given thought, we use the semantic index to identify all other thoughts that share some ancestor term and so may be semantically relevant. Next we rank the relevant thoughts for semantic similarity. We have experimented with a few different ranking scores; all examples in this paper are scored with the cosine measure

$$\frac{X \bullet Y}{\| X \| \times \| Y \|}$$

where X is the vector of terms from one thought and Y is the vector of terms for the second thought. This is a traditional similarity measure when applied to the terms of a document (Manning and Schutze 1999). We are experimenting with applying this measure to each of the four types of representation data in order to compare the match quality supported by each type. We currently are not weighting terms (e.g., considering the numbers of times a term appears in each thought and the number of thoughts in which a term appears); this is an obvious and important next step in our further work. In considering the ancestors of a thought when computing semantic similarity, we are effectively performing ontology-based query expansion (Brodner and Sang 1996, Navigli and Velardi 2003).

There are 23*22 candidate matches among the 23 thoughts. Table 3 presents the best one or two matches for 14 of the thoughts using the four different types representation terms for the similarity assessment. In Table 3, the left-most column (A) presents the IDs of the thoughts for which the best matches are computed. The remaining four columns each present the best matching thought and the score computed using one of the four types of representation data: tokens, lexemes, concepts, and ancestors. In each column, the left subcolumn (B) is the matching thought and the right subcolumn is the cosine score obtained for that thought. The best and second-best matches are presented for thoughts 8, 9, and 22; for the others the thoughts presented are the best among all the candidate matches. The thoughts in column A were selected to include the more promising matches as computed using the four types of representation data. Specifically, the thoughts in column A include the eight thoughts having the highest scores using the ancestor data, the concept data, and the token data, and the six highest matches using the lexeme data.

One of the most desirable candidate matches is matching thoughts 2 and 8. Note in Table 3 that thoughts 2 and 8 are high-ranked matches using the concept data and the ancestor data, but not using the lexeme nor the token data. The latter types of data support 2 and 18 as mutual best matches. The quality of pairing thoughts 2 and 8 over pairing thoughts 2 and 18 suggests the promise of using the more semantically meaningful terms from the ontology over the tokens and lexemes to estimate similarity.

One of the less desirable matches found among these data involves the matches for thought 16. Using the ancestor data and the concept data, the best matches are thoughts 8 and 2 (themselves deemed semantically similar); these matches seem quite reasonable. However, using the lexeme data and the token data, the best match is thought 10, which seems much less semantically similar to thought 16 than do thoughts 8 and 2. Again, the quality of pairing thought 16 with thoughts 2 and 8 over pairing it with thought 10 suggests the promise of the ontology terms over the tokens and lexemes to estimate similarity.

	ancestors		concepts		lexemes		tokens	
A	B	score	B	score	B	score	B	score
2	8	63.5	8	44.7	18	36.5	18	40.4
3	22	27.5	20	16.7	20	25.2	20	25.2
4	7	73.5	19	56.4	19	49.5	19	49.5
6	20	56.9	20	40.8	20	40.2	20	37.0
7	4	73.5	4	50.2	4	40.0	4	40.0
8 a	16	67.2	2	44.7	9	23.9	13	12.9
8 b	2	63.5	18	27.2	2	22.4	9	12.0
9 a	23	41.8	5	33.8	8	23.9	6	21.0
9 b	5	35.0	2	16.9	6	22.8	12	20.2
10	12	50.4	12	40.5	12	37.0	16	30.3
14	18	75.7	18	28.9	18	34.6	18	37.0
16	8	67.2	2	36.5	10	31.4	10	30.3
18	14	75.7	4	30.2	2	36.5	2	40.4
19	23	58	4	56.4	4	49.5	4	49.5
20	6	56.9	6	40.8	6	40.2	6	37.0
22 a	16	48.7	16	22.6	10	20.2	10	19.4
22 b	8	42.9	21	22.6	15	20.2	15	19.4

Table 3: Selected "Similar Thought" Matches

Another of the more desirable candidate matches is matching thoughts 3 and 22. Only by using the ancestor data is the best match for thought 3 found to be thought 22 (although the matching score for this is relatively low); this match is not found using the other representation types. Various thoughts are found as the best match for thought 22; thoughts 16 and 8 (using the ontology terms) seem more appropriate than do thoughts 10 and 15 (using the tokens and lexemes).

Another match that distinguishes use of the ancestor data involves the best matches found for thought 4. Using the ancestor data, thought 7 is deemed the best match; using other types of representation, the best match found is thought 19. The quality of pairing thought 4 with thought 7 instead of thought 19 again suggests the promise of using the ancestor data to support estimating semantic similarity.

Appraising the Coverage of Thoughts

The second application we have addressed is using the ontology to support helping Angler participants appreciate the coverage of the current set of contributed thoughts. This provides semiautomatic support for the participants to partially overcome their collective biases by guiding new contributions to take up as yet unconsidered topics.

One technique for identifying new topics involves determining which classes do not appear among the ancestors of the existing thoughts, and then presenting them for review in a standard hierarchical browsing interface. By analyzing the classes among the ancestors of the thoughts we can identify those classes defined in the ontology not yet associated with any thought (that is, not a superclass nor a subclass of any class that is an ancestor of some thought) and then organize them hierarchically along the subclass relation. Figure 2 presents a selected part of the browsing hierarchy of unconsidered concepts for the 23 example thoughts.

The concepts in Figure 2 suggest that the existing thoughts might be extended with new contributions to

Agreement
- Formal-Agreement
  - Treaty
  - Contract
- Informal-Agreement

Ally
- International-Ally
- Military-Ally
- Political-Ally

Antagonist
- Opponent

Figure 2: Hierarchy Fragment of Unconsidered Concepts

consider the possible agreements (formal and otherwise) and the allies and opponents of the new regime.

A second technique to identify and convey to the Angler participants what topics have not yet been considered involves identifying attributes mentioned by the current set of thoughts that have alternative values as yet not mentioned. For example, thought 10 suggests considering religious radicals; an alternative to a radical is a moderate, and it may be useful to consider religious moderates as well. Figure 3 presents some examples of new candidate thoughts; each is a variation of some existing thought, formed so that it references an alternative attribute value (e.g., moderate rather than radical, old rather than new, international rather than domestic).

Identifying Collaboration Opportunities

The third application we have addressed involves facilitating subsequent collaboration among subgroups of the Angler participants by identifying candidate areas of interest shared by the participants. This involves identifying terms in the ontology that are relevant to thoughts contributed by different authors. Each such term can be viewed as a possible topic of interest shared by the authors of those collaborations.

Table 4 presents selected results from performing this analysis on the 23 example thoughts. Concepts that suggest possible topics for further collaboration include Personnel, Reliability-Value, and Successor-Role; ancestor topics include Belief and Deter. Thus, it may be beneficial for authors B, F, and D to collaborate further on physiological considerations (e.g., religions, opinions, and perceptions).

Supporting Semantic Search over Thoughts

The fourth application we have addressed involves aiding the Angler participants in finding relevant thoughts. This can be supported by browsing or by semantic search.

- New vs. Old:
  - Angler-Thought-4:
    - description: Does the old regime have effective control over Country-X's nuclear arsenal and facilities?
- Radical vs. Moderate
  - Angler-Thought-10:
    - description: Extent of moderate Religion-Z leanings of nuclear C2 personnel, and/or level of corruption.
- Domestic vs. International
  - Angler-Thought-5:
    - description: international public opinion
    - catch-phrase: international audience
  - Angler-Thought-9
    - description: international nuclear and military corporate interests

Figure 3: Selected Thought Variations Formed with Alternative Attribute Values

Topic	Author	Thought	Relevant Concept
Personnel	F	10	Personnel
	B	12	Personnel
Deter	A	21	Intervention
	C	7	Deter
Reliability-Value	F	10	Reliability-Value
		16	Reliability-Value
	B	12	Reliability-Value
Successor-Role	F	22	Successor-Role
	A	15	Successor-Role
Belief	F	10	Religion-X Preference
	D	14	Preference
		18	Perception
	B	5	Opinion
		12	Religion-X

Table 4: Selected Candidate Collaboration Topics

Browsing is directly supported by the indexing created in the representation of the Angler thoughts: each ancestor accesses a folder that can be used to browse thoughts relevant to that ancestor. Figure 4 presents a hierarchical browsing interface in which each ancestor class presented; the numbers next to each class indicate how many thoughts are indexed by that class. Participants can browse this hierarchy, accessing a thought via any of its ancestor classes. Alternative browsing hierarchies could be formed using other relations (e.g., sub-region, part-of).

- Thing [23]
  - Abstract-Thing [22]
    - Abstract-Object [20]
      - Cognitive-Resource [14]
        - Knowledge [14]
          - Technology [14]
            - Firing-Technology [1]
            - Nuclear-Technology [13]
      - Disposition [5]
        - Mental-Disposition [5]
          - Belief [5]
            - Belief-System [2]
              - Religion [2]
            - Opinion [1]
            - Perception [1]
            - Preference [2]
      - Specification [3]
    - Extension [13]
      - Group [13]
        - Group-of-Agent [7]
        - Organization [2]
          - Legal-Organization [1]
          - Public-Organization [2]
            - Government [1]
            - Government-Organization [1]
              - Military [1]

Figure 4: Hierarchy Fragment of Considered Classes

Alternatively, the user can do a search for thoughts relevant to some set of search terms. Table 5 presents some results of semantic search using single-term queries; each row of the table identifies a query search term (e.g., Religion), the thought that best matches that query (e.g., thought 12), the cosine score for the match (e.g., 65.5), and, to explain the match, the concepts for the selected thought that are relevant to the query term and support the match (e.g., Religion is an ancestor of Religion-Z, a concept of the matching thought).

Representation Extensibility and Reuse

The applications described above identify several potential benefits from maintaining a semantic index for Angler thoughts. Next we consider the expense of supporting this index. Specifically, we consider the representational requirements that might be expected to handle additional thoughts: given a new thought, how many new lexemes and concepts and ancestors might we expect to have to define in order to represent the new thought?

One way to consider these requirements is by identifying the incremental need for additional representation terms for the 23 example Angler thoughts. Table 6 presents the incremental needs for tokens, lexemes, concepts, and ancestors required to represent each of the thoughts.

As indicated by the data in Table 6, the number of new terms required for each thought should be less than the number required by the preceding thoughts because later thoughts can reuse the representation terms required by prior thoughts. Since this assessment is sensitive to the order by which thoughts are encountered, the data in Table 6 summarizes 50 runs, each run considering the 23 thoughts in a random order, and then averaging over those runs, and then presenting the averaged incremental need for each type of representation data for every other thought.

As indicated in Table 6, initially the requirement for the ontology is substantial; however, the number of new concepts and ancestors required for each new thought diminishes quickly. In particular, the burden of providing the background knowledge required to define the ancestors is initially quite high, but it rapidly converges to levels only slightly higher than the concepts (e.g., only one new ancestor for each three new concepts). This suggests that the overhead of defining the hierarchical background knowledge in the ontology may be negligible in the long run because of the reuse of this knowledge across many thoughts.

Query	Thought	Score	Relevant Concept
Religion	12	65.5	Religion-Z
Energy-Device	17	38.5	Nuclear-Device
Enriched-Uranium	1	20.4	Fissile-Material
Asia	13	51.6	Country-X Country-Y Region

Table 5: Selected Results of Semantic Search

Thought	Tokens	Lexemes	Concepts	Ancestors
1	11.3	10.9	8.1	47.7
3	9.3	8.6	7.2	18
5	8.9	8.3	5.1	10.7
7	8.1	6.9	6.4	7.6
9	7.3	7	5.3	7.1
11	7.2	6.7	4.9	6.3
13	6.8	6.0	4.4	5.2
15	6.6	5.9	3.8	5.6
17	6.0	5.6	3.7	4.3
19	6.3	5.7	3.4	4.4
21	6.6	5.7	3.6	3.8
23	6.5	4.8	3.1	3.4

Table 6: Incremental Representation Requirements

Discussion

In spite of the relatively encouraging representation reuse data over the 23 example thoughts, a natural concern is how to support Angler participation considering entirely new domains. One approach is to use as a safety net alternative representation resources, such as WordNet (Miller 1995). For example, when a word is encountered for which no term in the ontology is defined but for which a WordNet synset is defined, the synset and its hyponyms (for ancestors) can be used until the ontology can be extended to cover the gap. We feel that an ontology provides better support for assessing similarity than WordNet (of course, this is an empirically testable hypothesis, see future work), but anticipate that a more graceful degradation in performance might be provided by using WordNet when the current coverage of the ontology is exceeded. We anticipate lightly trained users (vs. knowledge-representation experts) being able to extend the lexicon and the ontology with only periodic offline review of those additions by a more deeply trained editor. This would enable Angler users to optionally extend the representation resources as required to handle as yet unsupported words (and word senses) as they author new thoughts, enabling the semantic index to immediately support its applications without intervention by linguists or knowledge-representation experts.

Our approach of using an ontology to relate distinct terms differs from the more traditional approaches to information retrieval (IR), such as latent semantic analysis (LSA), for two reasons. First, Angler provides an inadequate corpus over which to perform the LSA; in particular, the size of text for each document (e.g., thought) is small in the context of Angler, relative to traditional IR applications (i.e., the 23 example Angler thoughts included just 13.4 words on average). Second, and perhaps more important, is the need for explaining judgments of similarity. Our applications provide interactive support for humans in making decisions about Angler thoughts; it is important to be able to explain to the human user why two documents are deemed similar or why some document was deemed relevant to some search term. Note that in tables 4 and 5 we present the concepts associated with words in the thought texts that support the assessment of semantic similarity. Traditional IR approaches to dimension reduction, such as LSA, often cannot provide meaningful explanations for why two texts are deemed to be semantically similar.

The semantic description of the content of Angler thoughts involves terms rather than logical sentences; that is, it characterizes what concepts (e.g., topics) have been mentioned by a thought rather than trying to capture the relations among those concepts. Consequently, the meaning captured is much less precise than a full propositional account of the thought content. However, we feel this is appropriate for our applications for several reasons. First, the Angler participant is firmly involved; for example, the results of the semantic-similarity analysis are presented to the participant for vetting and possible action. Second, we hypothesize that the term-level analysis is much less brittle than attempting a full propositional account of the thought texts. Finally, we feel that the Angler participants (vs. knowledge-representation experts) are in a better position to extend an ontology with terms (as might be defined in an organization's thesaurus) and simple relations over those terms (e.g., sub-region-of) rather than trying to make extensions to a knowledge base full of quantified first-order logic sentences.

Future Work

Term weighting: One important component of our future work is to add consideration of term weighting to our semantic-similarity measure. This should improve the performance quality on the semantic search and semantic-similarity tasks. Term weighting for tokens, lexemes, and concepts will summarize both how often a term is referenced within each thought, and how many different thoughts reference the term. For ancestors, term weighting may also summarize how "near" the ancestor is to the relevant concepts appearing in the thought (e.g., matching the concept Pistol to Gun should be stronger than matching it to Artifact).

Evaluation: A second important component of our future work involves an evaluation of the matches supported by the different types of representation terms. For example, we can identify when different best-matches are supported by the different representation data (e.g., tokens, lexemes, concepts, and ancestors). Then we can ask a subject-matter expert to rank the alternatives. This should give us a quantitative measure of how much improvement (if any) lexemes provide over tokens, concepts over lexemes, and ancestors over concepts. In addition to the current data types (e.g., tokens, lexemes, concepts, and ancestors), it would be prudent to perform comparative experiments using WordNet synsets instead of the ontology terms. For each among a broad set of candidate transformations and alternative representations we hope to empirically assess the associated costs and benefits.

Phrases (vs. terms): A third direction of future work is extending TextPro and the KetL ontology to handle head-modifier phrases rather than simple terms. For example, "nuclear weapon" would result in a structured term, [Weapon :modifier Nuclear-Energy], rather than a set of independent terms {Nuclear-Energy Weapon}.

Applications beyond Angler: While Angler has provided the original motivation to develop the semantic-index uses described above, we feel that these applications may be usefully applied in other contexts as well. For example, the semantic-similarity analysis may help identify similar argument steps and rationale recorded in SEAS (Lowrance, Harrison and Rodriquez 2001), or it may help identify similar FAQs or their answers in technical help systems. The semantic search may be useful in helping people manage email, and so on. Further investigation should help to reveal the useful limits of these applications.

References

9/11 Commission 2004. 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks Upon the United States. http://www.gpoaccess.gov/911/.

Appelt, D., and Martin, D. 1999. Named Entity Extraction from Speech: Approach and Results Using the TextPro System. In Proceedings of the DARPA Broadcast News Workshop.

Barker, K., Porter, B., and Clark, P. 2001. A Library of Generic Concepts for Composing Knowledge Bases. In Proceedings of the First International Conference on Knowledge Capture.

Bodner, R., and Song, F. 1996. Knowledge-based approaches to query expansion in information retrieval. In Lecture Notes in Computer Science, volume 1081

Cycorp. The Syntax of CycL. http://www.cyc.com/cycdoc/ref/cycl-syntax.html

Genesereth, M., and Fikes, R. 1992. Knowledge Interchange Format Version 3.0 Reference. Stanford University.

Gruber T. 1993. A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220.

King, B. 1967. Step-wise clustering procedures. Journal of the American Statistical Association, 69:86–101

Lowrance, J., Harrison, I., and Rodriguez, A. 2001. Capturing Analytic Thought. In Proceedings of the First International Conference on Knowledge Capture.

Manning, C., and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Miller, G. 1995. WordNet: a lexical database for English Communications of the ACM, Volume 38, Issue 11.

Navigli, R., and Velardi, P. 2003. An Analysis of Ontology-based Query Expansion Strategies, Workshop on Adaptive Text Extraction and Mining. In Proceedings of the 14th European Conference on Machine Learning.

Rodriguez, A., Boyce, T., Lowrance, J., and Yeh, E. 2005. Angler: collaboratively expanding your cognitive horizon. Submitted to the First International Conference on Intelligence Analysis Methods and Tools.

Schwartz, P. 1991. *The Art of the Long View: Planning for the Future in an Uncertain World.* Currency Doubleday.

Teknowledge. SUMO. http://ontology.teknowledge.com