AFRL-RI-RS-TR-2009-122
Final Technical Report
April 2009

SYNERGIST: COLLABORATIVE ANALYST ASSISTANT

Lymba Corporation

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2009-122 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/	/s/
JAMES M. NAGY Work Unit Manager	JOSEPH CAMERA, Chief Information & Intelligence Exploitation Division Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* APR 09	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* Oct 06 – Feb 09

4. TITLE AND SUBTITLE

SYNERGIST: COLLABORATIVE ANALYST ASSISTANT

5a. CONTRACT NUMBER
FA8750-06-C-0201

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)

Munirathnam Srikanth, Marta Tatu, Adriana Badulescu, Guillaume Bailey, Marian Olteanu, and Christine Clark

5d. PROJECT NUMBER
CASE

5e. TASK NUMBER
00

5f. WORK UNIT NUMBER
08

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lymba Corporation
1701 North Collins Blvd., Suite 3000
Richardson, TX 75080-3587

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RIED
525 Brooks Rd.
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
NA

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TR-2009-122

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA# 88ABW-2009-1674

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Abstract: Intelligence Analysts work with a significant amount of information expressed in natural language. The textual information that Analysts work with or generate provide important clues on what is known to them, what are their immediate and long term needs, and what aspects are typically explored or missed in the context of a particular topic. For CASE, Lymba developed Synergist: Collaborative Analyst Assistant, an automated system that understands unstructured content, extracts knowledge from text and collaborates with intelligence analysts to model their needs and to aid with their information discovery. Synergist provides a grid-computing framework for state-of-the-art natural language processing to extract rich semantic information from large document collections. With a hierarchical semantic representation of entity, event, relation and context information text, Synergist is able to capture, represent and organize the knowledge from text in ontologies and knowledge bases.

15. SUBJECT TERMS
Natural language understanding, knowledge extraction, representation and organization, natural language reasoning, prior and tacit knowledge, evidence marshalling, multi-lingual information access, tacit collaboration, natural language question and answer, context

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** James M. Nagy
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	56	**19b. TELEPHONE NUMBER (*Include area code)*** N/A

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 SUMMARY

Intelligence Analysts[1] work with a significant amount of information expressed in natural language. This includes the tasking or problem description they get from decision makers, the queries they generate to search for information, the relevant documents and text snippets they collect in their repositories and the intelligence products they generate to address the information needs of decision makers. The textual information that Analysts work with or generate provide important clues on what is known to them, what are their immediate and long term needs, and what aspects are typically explored or missed in the context of a particular topic. For CASE, Lymba developed **Synergist: Collaborative Analyst Assistant**, an automated system that understands unstructured content, extracts knowledge from text and collaborates with intelligence analysts to model their needs and to aid with their information discovery.

Synergist provides a grid-computing framework for state-of-the-art natural language processing to extract rich semantic information from large document collections. With a hierarchical semantic representation of entity, event, relation and context information from text, Synergist is able to capture, represent and organize the knowledge from text in ontologies and knowledge bases.

Lymba has demonstrated automatic ontology generation in a variety of different settings. This includes National Intelligence Priorities Framework (NIPF) topical ontologies that automatically detected and organized concepts and relations under different NIPF topics and Analyst's prior and tacit knowledge bases (PTKB) from textual documents from their shoebox.

While providing interfaces to search and browse the automatically generated topical ontologies or analyst knowledge base, Lymba also developed prototypes to exploit the acquired knowledge in different applications:

- Synergist uses the formal semantic representation of information to reason on text, to extract structured arguments from text and to marshal evidence that support or refute Analyst's claims or hypotheses.
- Develop EventNet as a linguistic resource that organizes event attributes and properties for use in hypotheses generation and situation understanding.
- Predicting novelty of information to an analyst based on their PTKB.
- Synergist supports tacit collaboration by recommending different resources including documents, answers, users, tags, and topics that can satisfy user's information need.

These capabilities are realized in **Synergist Analyst Suite** which is an **integrated Search-Browse-Discovery platform** for Intelligence Analysis.

[1] Intelligence Analysts and Analysts in this document refers to members of the intelligence community who perform intelligence analysis to assist decision makers understand situations and make decisions.

Human language understanding is essential for Incisive Analysis[2] of actors and their actions in the real world. Synergist's textual analysis and knowledge exploitation capabilities are exposed through web-services that can be accessed by any contractor of an IARPA program. Synergist integrates these services with other CASE applications (e.g. BAE's expertise finder) that complement Lymba's current capabilities.

Lymba has transitioned some of these capabilities as part of the HOSTT project to the intelligence community. In addition, research from Synergist has contributed to Lymba's commercial activities developing text understanding, knowledge organization, and social text analysis software for customers like Northrop Grumman, Oracle and Xerox.

[2] As one of the program offices of IARPA, Incisive Analysis has the goal for its programs to maximize insight from the collected information in a timely fashion.

2.0 INTRODUCTION

For Synergist, Lymba proposed nine (9) technical tasks that included two (2) optional tasks. The following sections describe the activities and accomplishments of the seven (7) tasks performed during the CASE program. Due to the cancellation of the CASE program, Lymba did not start its work on the ‘Geospatial Ontologies and Reasoning’ and the optional ‘Collaboration’ task. These tasks were to be explored in the later years of the project. In addition to a brief description of the objective of a task, the description correlates the work done to the activities identified in the project’s statement of work. The report will discuss each technology area with research performed and evaluations conducted.

3.0 EXTRACTING RICH KNOWLEDGE FROM TEXT

The objective of this task was to perform natural language processing on large document collections, efficiently extract knowledge in the form of entities, events, general concepts, relations and context, and build representations that yield well to reasoning on text and providing information access. NLP is computationally-intensive and Lymba has developed a distributed grid-computing framework to provide these capabilities.

3.1 Distributed NLP Framework

Unlike web-search engines that require keywords in documents and at best document structure information, text understanding involves a number of steps to capture, represent and link information at lexical, syntactic and semantic levels. Lymba has developed a solution that is based on distributed grid-computing for processing large document collections to extract semantic information. The grid framework is based on ZeroC's Internet Communications Engine (ICE) [www.zeroc.com]. Nodes in the grid run the NLP pipeline and other NLP applications. This includes natural language question-answering (systems that understand a natural language question and return text snippets as answers to those questions), automatic ontology building (ref. Section 4.0), and resource recommendation (ref. Section 6.3) and new nodes can be dynamically added to the grid to support increased loads.

Figure 1 presents the different components that constitute Lymba's NLP pipeline. Spanning the lexical, syntactic, and semantic layers of information extraction from text, the pipeline also includes a cross-document information fusion module that correlates information from different documents and organizes them in the extracted knowledge base. In the Synergist project, Lymba developed and/or improved on the following components: **Concept/Temporal Tagging**, **Semantic Parsing**, **Context Detection**, and **Event and Relation Extraction**.

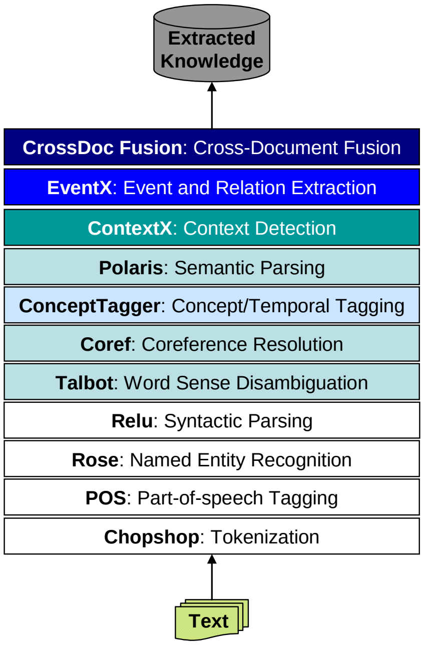

Extracted Knowledge
CrossDoc Fusion: Cross-Document Fusion
EventX: Event and Relation Extraction
ContextX: Context Detection
Polaris: Semantic Parsing
ConceptTagger: Concept/Temporal Tagging
Coref: Coreference Resolution
Talbot: Word Sense Disambiguation
Relu: Syntactic Parsing
Rose: Named Entity Recognition
POS: Part-of-speech Tagging
Chopshop: Tokenization
Text

Figure 1: Lymba's NLP Pipeline

3.2 Identifying Concepts in Text

Lymba developed methods for detecting, capturing and representing concepts in text (Section 4.1 describes Lymba's semantic representation of information extracted from text). Text understanding requires the concepts detected in text be mapped to an underlying ontology. Lymba used WordNet as the underlying ontology. Concept extraction includes detection and normalization of temporal expressions.

3.2.1 Concept Detection

Lymba uses WordNet[3]-based concept detection methods using ***lexico-syntactic and semantic rules*** for identifying words and phrases as concepts. Machine learning algorithms for disambiguating word sense defined by the synsets[3] of WordNet are used to identify and associate concept occurrences in text to their corresponding synsets in WordNet. Lymba's concept detection methods ranged from the detection of simple nominal and verbal concepts to more complex named entity and phrasal concepts. Concepts that do not appear in WordNet are identified as new concepts and classified later in automatic ontology and knowledge base generation by Jaguar.

Lymba ***evaluated its automated word sense disambiguation algorithm in the 4th International Workshop of Semantic Evaluation (SemEval 2007)*** coarse-grained all-words sense disambiguation task and was ranked second (2nd) in performance. We experimented with different syntactic features and features from Extended WordNet Knowledge Base (XWN-KB). New sense clusters are used in more accurate concept detection.

3.2.2 Temporal Expression Detection and Normalization

Lymba developed a new temporal expression detection framework with a pattern-based approach using temporal patterns and hand-coded rules for tagging. The methods ***detect absolute data*** (*5 January 2008*) ***and times*** (*5:00 pm*)***, partial*** (*5 January*)***, relative*** (*last Friday*) and ***fuzzy*** (*a few weeks*) ***temporal expressions in text***.

Temporal normalization associates temporal data structures for time expressions that facilitate robust representation and manipulation of time expressions. Lymba developed a number of back-off steps to identify reference time for events described in text. These steps included detecting the reference time from text using the first absolute time in content, the document/file time and document processing time. All temporal expressions are normalized based on the reference time. Temporal expression detection and normalization was extended to include durations. Initial evaluation achieved a score, comparable to the state-of-the-art, on temporal tagging of 93% precision, 92% recall.

[3] WordNet (http://wordnet.princeton.edu) is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

3.3 Semantic Relations

Semantic relations as output by **Polaris**, Lymba's ***semantic relations extractor***, are defined as connections of a certain type of *meaning* between words or concepts. They are similar to the relations in conceptual graphs, and also to theta-roles or thematic roles in linguistics, or to semantic roles in the semantic role labeling task in natural language processing. Lymba has developed a list of 30 semantic relations that are useful in linguistics and knowledge representation. These semantic relations can be identified in text, and also used as elements of knowledge bases.

As an example of semantic relations, the sentence "*I went to the park yesterday because I saw hot air balloons taking off from there*," contains the following semantic relations as shown in Table 1:

Table 1: Semantic relations in the example sentence

AGENT(I, went)	*STIMULUS(hot air balloons taking off from there, saw)*
LOCATION(to the park, went)	*PROPERTY(hot, air)*
TEMPORAL(yesterday, went)	*ISA(hot air balloons, balloons)*
CAUSE(I saw hot air balloons taking off from there, I went to the park yesterday)	*AGENT(hot air balloons, taking off)*
EXPERIENCER(I, saw)	*LOCATION(from there, taking off)*

3.3.1 Semantic Relations Hierarchy

We conducted studies toward defining and rearranging our set of semantic relations into a more useful ***hierarchy of semantic relations***. This allows applications to use semantics at different levels of granularity, and systems to output them at higher precision. We determined ***semantic mappings from various resources***, such as PropBank[4], and FrameNet[5], to our semantic relation set, to enable us to train Polaris using such resources and improve its performance.

3.3.2 Semantic Relations Extraction from Text

Lymba improved semantic relation extraction from text with new patterns and methods for covering a larger amount of text. This includes an expanded ***list of syntactic patterns*** used for detecting ontological relations like appositions, coordinations, and lists. New classifiers were developed for detecting and linking semantic relations between nominals. These approaches were evaluated in SemEval 2007 [7]. Other improvements include ***additional sentential patterns*** for interrogative sentences and other frequent patterns around noun phrases. For example, the pattern that looks at the sequence of noun phrase, followed by a verb and a noun phrase detects semantics that links the *report* to *information* in the sentence: *The report contains important information*.

[4] M. Palmer, P. Kingsbury, D. Gildea. *The Proposition Bank: An Annotated Corpus of Semantic Roles*, Computational Linguistics 31 (1): 71–106, 2005.

[5] C. Fillmore and C. F. Baker. *Frame semantics for text understanding*. Proceedings of WordNet and Other Lexical Resources Workshop. NAACL, 2001.

We started work on ***complex nominal bracketing/paranthetization*** by gathering text statistics and using them for determining the correct association of the words from a complex nominal. For examples, in the compound "*former American President John F. Kennedy*" the correct bracketing is "*(former (American President)) (John F. Kennedy)*", given by the higher probability of the n-grams "*John F. Kennedy*" and "*American President*" and "*former American President*" in text. We developed the capability to detect and tag relations between multi-word concepts that are part of a complex nominal.

3.3.3 Machine Learning and Explicit Classifiers

Semantic relations are detected by first using syntactic patterns to identify the participants of a relation and then using explicit and machine learning based classifiers to identify the semantic relation. We improved the verbal and nominal classifiers and the evaluation benchmarks. Part of our focus was on improving the extraction of semantic relations like ***IS-A, Part-Whole, Cause*** (ref. to Table 1 for examples) and other ***ontological relations*** for their importance to automatic generation of high accuracy ontologies (ref. Section 4.2) and topical indexes (ref. Section 8.1) and building an EventNet for hypotheses generation(ref. Section 7.0).

We expanded our training data from different sources including FrameNet, WordNet and EXtended WordNet (XWN) glosses, and others. We also used techniques like cross-validation (using portions of labeled training data to test automatic labeling) and bootstrapping (learn from automatic labeling of unlabeled data) for improving the number of examples and the quality of the classifiers.

Improvements to Polaris include additional syntactic and semantic features for existing classifiers, use improved resource annotation mapping and bootstrapping the number of training examples. Polaris uses a hybrid rule-based and machine learning approach to relation detection. Lymba experimented and evaluated maximum entropy classifiers. They provided speed improvements and comparable performance to SVM and decision tree classifiers.

Considering the complexity of the semantic relation detection task, Polaris performs well on different types of corpora including the cable data corpus (63.90% F-measure[6]), TreeBank financial corpus (50.40% F-measure), and GlassBox intelligence reports and document (50.28% F-measure).

3.3.4 Temporal Relation Identification

We enhanced the temporal relation identification capability by introducing more lexical rules based on temporal relations in VerbOcean, a publicly available verbal resource for linguistic research. VerbOcean is a broad-coverage semantic network of verbs and automatically constructed from Google. We extracted all verb pairs in VerbOcean that contains temporal relations and converted them into Decision Tree rules that can be used by temporal relation identification module in the system.

[6] F-measure is a single measure that combines precision (P) and recall (R) values given by 2PR/(P+R).

3.4 Co-reference Resolution

Lymba's co-reference detection module ***identifies nominal and pronominal anaphora and links them to antecedents*** they refer to. Building a profile based on these with-document co-reference links, the intent is to build entity co-references across documents and incorporate them in the knowledge bases. We also worked on entity co-references across documents developing methods to represent entity profiles in a document and compare them. We started working on interfaces for displaying disambiguated entities and drill down to text snippets and documents about entities.

The effort was to perform entity-centric discourse analysis to identify and resolve entity references in text. Lymba experimented with different approaches for pronominal and nominal co-reference resolution in text including variations of Hobbs algorithm[7] and Lappin and Leass' Resolution of Anaphora (RAP) algorithm[8] for pronominal co-reference resolution. While these methods addressed only pronominal co-reference, Lymba developed machine learning approaches to pronominal and nominal co-reference resolution by using contextual and semantic features of anaphora. We evaluated our co-reference component on the MUC-6 and MUC-7 combined corpora (referred to as MUC) and ACE-2 datasets. In our Machine learning approach, window size plays an important part in improving the recall and hence the F-measure of the system. We obtained F-measures ranging from 45.36% (for size3) to 51.81% (for size 20) on the MUC corpus and 44.44% (for size 3) to 51.71% (for size 20) on the ACE-2 corpus.

[7] J. Hobbs. *Resolving pronoun references*, Lingua, 44:311-338, 1978.

[8] S. Lappin, H. J. Leass. *An Algorithm for Pronominal Anaphora Resolution*, Computational Linguistics, 2004

3.5 Event Extraction

Lymba has built an ***event detection and extraction*** module as part of the CASE project. The component detects candidate event concepts in text, identifies their event class (e.g. reporting, perception, aspectual, etc.) and builds event structures that represent the attributes and participants of an event. In addition temporal expressions in text are tagged and normalized, where possible, to their absolute times. The temporal expressions tagged and normalized include absolute (e.g. *Friday, October 1st, 1999, the winter of 1999*) and fuzzy dates and times (*two days earlier; three past five; 11 in the morning*) and durations (*an 8-month period of time; the next 45 years; a few days*). This results in the tagging and linking of temporal expressions and events as illustrated in the following example: (*[t1 Over the summer], Anheuser competitors [e1 offered] more and deeper discounts than industry observers have [e2 seen] [t2 for a long time]: DURING(e1,t1); BEFORE(e2,e1)*). The addition of this rich information (events and event structures, normalized temporal expressions and event-event and time-event relations) of textual knowledge as an additional layer results in a much more expressive representation of text and enables text understanding. Lymba experimented with the use of this representation and reasoning on them to detect and link event-time and event-event relations. The current performance of the event extraction is 91% precision and 92% recall.

Semantic relations for an event concept are organized in ***event structures***. The event structures will include other normalized attributes like temporal and spatial information associated with the event and other attributes of events that are useful, such as event class, tense, aspect, whether the event is a “main event” in the sentence, and more.

While ***event co-reference*** is similar to entity co-reference resolution, it is more sensitive to context and needs more semantic input in resolution. Our primary approach is to use in-house and external semantic parsers to generate semantic features from a large but adjustable context in documents. Taken these features as input, along with other event features extracted by our system, we experimented and extended two widely cited statistical text similarity measures to compute an *identity score* for each pair of events.

3.5.1 Event-Time and Event-Event Relations

We developed an event extraction component to link temporal expressions to events and tasks for linking events. This work was evaluated in the SemEval 2007 – TempEval challenge (http://timeml.org/tempeval).

We also focused on solidifying that work into a robust NLP tool for event extraction. A number of feature extraction rules have been developed to extract features and attributes of events to detect and assign appropriate temporal relations between them.

We also extended our capability to assign ***temporal attributes*** of main events to assign time-stamp for each event identified in text. The assumption is that each event in text must occur or hold in a particular time frame (a time point or interval). Most of time this temporal aspect of events are implicitly expressed or totally hidden in natural text, and our goal is to recover this temporal knowledge by assigning an explicit time for each event identified by our system. There are a number of practical applications for this research (e.g. temporal QA and summarization). Obtaining precise time-stamps for events is important for document understanding.

In addition to processing events in news-style text, we extended our capabilities to other domains targeted at particular applications. We developed a ***Use Case for the Terrorism Intelligence Analysis domain***: An initial set of 37 intelligence analysis reports from the “Sign of the Crescent” case study by Frank Hughes have been collected and preprocessed. There are good reasons for choice of this domain: 1) intelligence analysts are sensitive to temporal and eventual aspects of activities in text, since in their analysis they usually seek answers to such questions as “At what time, where and what did they do? How are these activities connected? 2) Effective analysis of intelligence data is crucial for national security 3) Developing such a use case would also be a good opportunity to evaluate our system performance.

3.5.2 Evaluation of Event Extraction

We enhanced the TimeBank corpus with additional annotation of assigning explicit time to events. Two annotators have completed an annotation of 62 documents (31 by each) from the TimeBank corpus. We evaluated the system's accuracy when assigning time to events at three levels of time: Year, Month and Day, as well as the ability to identify major activities (Main events) from text.

We performed a series of experiments for **evaluating temporal and event information extraction** capabilities of our system, and improved the system based on performances at different stages of evaluation. For time expression extraction, we obtained comparable performances to the state-of-the-art: 0.972 & 0.959 of recall and precision respectively (F-measure 0.965). Temporal normalization performed at 0.905 precision. Event-time linking for anchoring temporal expressions in a sentence to corresponding events performed at 0.866 precision. All these tasks are evaluated on a portion of TimeBank 1.2 articles. For temporal relation resolution, our main effort was to introduce our reasoning system to enhance temporal relation identification capability by utilizing a temporal reasoner to boost training data. We trained multiple machine learning models for fine- and coarse-grained temporal relations and experimented multiple machine learning algorithms, including Support Vector Machine and Maximum Entropy.

Our primary findings in those experiments were:

1. temporal closure does help improve the system's ability to resolve temporal relations, but in a minor way in our experimental setting, improving about 2 points for accuracy;
2. in terms of machine learning algorithms, Maximum entropy is slightly better than Support Vector Machine;
3. The system did not perform well on learning the original 13 fine-grained relations defined in TimeBank corpus (adopted from Allen's temporal algebra[9]), which is due to the nature of vagueness and implicitness of natural languages. We are looking at alternate methods for improving its performance.

Evaluating the performance on coarse-grained relations (Before, After and Overlap, as defined in SemEval 2007 workshop), we get a best combination of 0.940 and 0.612 of precision for time-event relation and event-event relation respectively, which are better than the best scores reported in SemEval 2007 evaluation effort.

[9] J. F. Allen, *Maintaining Knowledge about Temporal Intervals*, CACM, 1983.

3.5.3 Event Timeline Creation

Lymba performed experiments on identifying and normalizing time expressions, associating time attributes to events and generating a timeline of events. The focus here was to identify the starting and ending points of the *DURATION* from text using different kinds of strategies, therefore turning this *DURATION* into an Interval. On an annotated dataset, Lymba evaluated the identification of starting and end times for events detected in text.

3.6 Context Detection and Tagging

We focused on identifying several types of context which include report, belief, volitional, conditional, temporal and spatial. For example, detecting report context involves recognizing and associating statements or CONTENT to its SOURCE. *[Clinton]:SOURCE informed the two parties that [in the absence of an agreement, a public announcement would be made, with the apparent objective of applying pressure to both sides, particularly the Israelis]:CONTENT.*

We use a *syntax-based approach for detecting data contexts*. Signal concepts (e.g. communication verbs and nouns provide clues about report contexts) identify candidate contexts that are filtered by a matching process against a set of syntactic tree regular expressions. A successful match identifies the source to the boundaries of the context. The source of a context is the entity that reports, beliefs, or wants the statement to occur associated with the time or location of the statement. The boundary of a context identifies the statement being reported, believed, or desired, etc. The knowledge representation module uses the source and the content components of the context as well as the type and the signal of the context to generate an accurate representation of the text which conveys the source of the context as a precondition for the validity of the contextual content. For the report context shown above, this translates into if *Clinton's report is true, then in the absence of an agreement, a public announcement would be made, with the apparent objective of applying pressure to both sides, particularly the Israelis*.

We detect **report contexts**. For example consider the sentence *"After all, it was intended to be a scientifically rigorous report," the judge said*. The context of the information in quotes is as reported by the judge. Our detection strategy is to create a list of syntactic patterns that match report context in parse text. The indicators for the contexts were verbs and nouns whose lexicographic information given in WordNet 2.0 is "communication". Initial set of syntactic patterns included 20 patterns with semantic restrictions (the listener of the judge’s statement cannot be a named entity of type date or time).

For **belief contexts**, we used concepts that indicate cognitive processes. The set of syntactic-semantic patterns we use for identifying the scope and the source of the context includes 36 patterns, which overlap only partially with the ones we developed for report contexts.

We also implemented algorithms for detecting and tagging **desire context** (example – John hopes that [his brother will win the lottery]). We are adapting the current context detection tool for "desire" contexts. Desire context indicators include feeling-related nouns ("sentiment", "urge"), emotion-related verbs ("want", "hate") and adjectives which are values of the noun indicators ("afraid" is a value of "fear"). Because of WordNet's sparseness with respect to synset-synset relations across part-of-speech, we added an additional set of adjectives. The initial selection was done based on the definition of the adjective and it was followed by manual filtering. To increase the robustness of the context detection tool, we also added a set of rules which make up for incorrect parse trees.

Report and belief context detection and tagging algorithms have been enhanced with additional constraints to ensure the correctness of some of their context elements (for example, recipient of report context). The benchmark we created for the evaluation of the context detection includes, so far, 22 sentences with 30 contexts detected.

4.0 AUTOMATICALLY CONSTRUCTING KNOWLEDGE BASES

Objective: Develop automatic tools for representing and organizing the information extracted from text or other resources into knowledge bases. Transform WordNet into a very large, general-purpose knowledge base. Develop methods for re-using knowledge by merging knowledge bases or federating them together on an ad-hoc, needs-driven basis. Allow analysts to weight the contribution of each KB in the federation.

4.1 Layered Knowledge Representation

Answering/Reasoning complex questions/information requires a great deal of natural language processing (NLP). For this purpose, Lymba Corporation has pioneered a deep knowledge representation of text; a hierarchical representation consisting of several layers as shown in Figure 2. Each layer in the hierarchy builds upon the previous layer by bringing new information. The base is the lexico-syntactic layer which reveals the text's syntactic structure. The semantic layer adds semantic relations between the concepts identified by the previous layer. The context layer identifies temporal, spatial and subjective contexts and adds this information to the previous layers. Next, events and event properties are identified and this brings a new level of abstraction to the knowledge representation since events tend to dominate the meaning of text. Last, relations between events are established with the goal of determining the text coherence at the highest level.

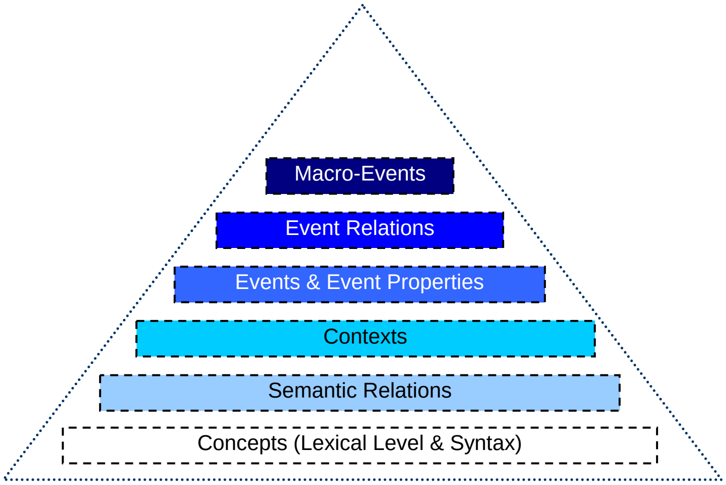


Figure 2: Lymba's hierarchical knowledge representation of natural language texts.

In order to demonstrate the depth of our proposed representation, let us consider the following text snippet *Gilda Flores's kidnapping occurred on January 13, 1990. A week before, he had fired the kidnappers*. Figure 3 displays the multi-layered representation of these two sentences. All concepts mentioned in the text constitute the bottom layer of the representation. Named entity classes (e.g., *Gilda Flores* = *human*) and normalized values of temporal expressions (e.g., *a week before January 13, 1990 = 01/06/1990*) are captured within this layer of lexical information. The syntactic parse trees of these two sentences are also included in the bottom layer (not shown in Figure 3). The next level of knowledge is dedicated to semantic relations identified between the text's concepts (e.g., ***theme*** (*Gilda Flores*, *kidnapping*)). Co-referring concepts are also stored in this layer of representation (*he* = *Gilda Flores*). Contexts are identified and represented as the third layer of information, followed by events with their properties and relations (e.g., *kidnap*, ***theme***(*Gilda Flores*), ***agent***(*kidnappers*), ***during***(*01/13/1990*), ***cause***(*fire*, *kidnap*)). Lastly, the high-level event that semantically combines the two sentences is derived.

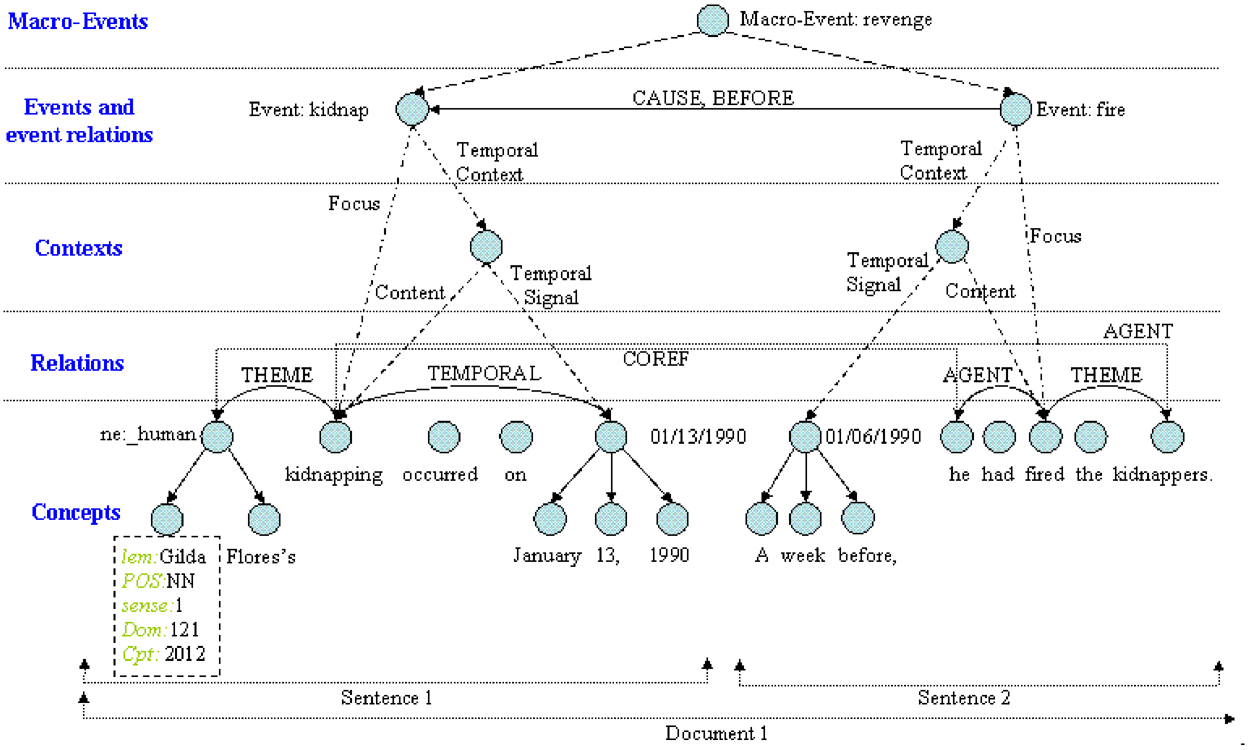


Figure 3: Example of a hierarchical knowledge representation

We note that many of the layers shown in our knowledge representation capture a text's semantic information. This rich knowledge encoded in Lymba's representation is required for any successful NLP automated system, including QA engines. The hierarchical knowledge representation of input text (documents or questions) is the foundation of all succeeding NLP modules.

4.2 *Automatic ontology and knowledge base generation from text*

Improvements to Jaguar; generating topical ontologies; National Intelligence Priorities Framework (NIPF) ontologies; evaluating ontologies; refer to papers; knowledge bases from text; ontology browser;

Jaguar automatically builds domain-specific ontologies from text [1]. The ontology/knowledge-base created by Jaguar includes the following constituents:

- Ontological Concepts: basic building blocks of an ontology
- Hierarchy: structure imposed on certain ontological concepts via transitive relations that generally hold to be universally true (e.g. ISA, Part-Whole, Locative, etc)
- Contextual Knowledge Base: semantic contexts that encapsulate knowledge of events via semantic relations
- Axioms on Demand: capture assertions about knowledge and are useful for reasoning

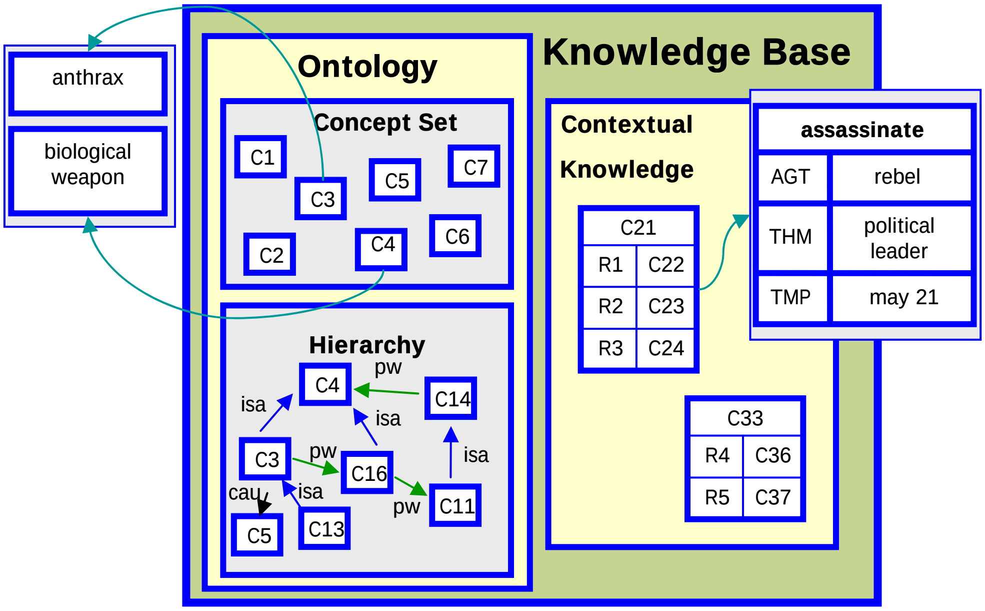


Figure 4: Example Jaguar knowledge-base containing concepts, hierarchy and contextual knowledge.

Figure 4 shows an example Jaguar knowledge-base containing concepts, hierarchy and contextual knowledge. The input to Jaguar includes a document collection (Text, MS Word, PDF and HTML web pages, etc.) and a seeds file containing the concepts/keywords of interest in the domain. Jaguar's ontology creation involves complex text processing using advanced Natural Language Processing (NLP) tools, and an advanced knowledge classification/management algorithm. A single run of Jaguar can be divided into the following two steps:

1. Text Processing
2. Classification/Hierarchy Formation

In **Text Processing**, the first step is to extract text from the input document collection and then filter/clean-up the extracted text. The text files then go through a set of NLP processing tools (named-entity recognition, part-of-speech tagging, syntactic parsing, word-sense disambiguation, coreference resolution, and semantic parsing (or semantic relation discovery)). The concept discovery module then extracts the concepts of interest using the input seeds set as a starting point and growing it based on the extracted NLP information.

The classification module forms a hierarchical structure within the set of identified domain concepts via transitive relations that generally hold to be universally true (e.g. ISA, Part-Whole, Locative, etc). Jaguar uses well-formed procedures to impose a hierarchical structure on the discovered concepts set with WordNet [3] as its upper ontology.

4.2.1 Improvements to Knowledge-Base/Ontology Generation from Text

- Designed a KB representation called a cluster to capture the semantics and events corresponding to a single discourse (currently, a sentence in text).
- Provided facility to access the clusters in a KB, and also ways to search the clusters for information regarding particular concepts.
- Researched and developed methods for extracting concept properties from text. The plans include steps to use the named entities detected in text and develop an initial application of semantic calculus to propagate and extract properties from text. The semantic relations extracted from text, in conjunction with semantic calculus, are used to identify the properties and their values from text for different concepts in text.
- Experimented with axiom extraction and propagation (up the ISA hierarchy) for the Natural Disasters ontology. The procedure was similar to that performed on the XWN-KB. This process requires converting the said ontology from a noun-only ontology into a full-fledged KB with verb concepts and the appropriate AGT, THM relations. The axioms were found to be NOT as rich as those for XWN-KB as the ISA hierarchy is very shallow compared to that of XWN-KB, where propagation of axioms led to much richer axioms. More experiments are required to develop methods for building a richer KB from text.
- Designed a methodology for creating Domain KBs on top of Lymba's General Purpose (Common sense) KB (XWN-KB). Domain KBs require enhanced concepts in XWN-KB with domain specific properties and axioms. In addition they have to be created from documents relevant and rich in domain concepts and properties. We experimented with the idea of using screened wikipedia documents on specific topics (e.g. Natural Disasters)
- Implemented methodologies for intelligent enhancing of domain ontologies. We ran several experiments including one on the existing Natural Disasters Domain ontology, which was built automatically from news articles on the domain. An enhanced ontology was constructed using screened Wikipedia documents (concepts like volcano, flood, quake, famine, etc.). The results were interesting. We found many noun-centric relations as against those extracted from news articles on the domain.
- Improved the rules/methods for detecting concepts expressed in text. This includes better processing of head-words in concepts, named entities and collocations, certain PP attachments, handling possession, conjunctions and hyphenation in parse trees.

- Analyzed some of the ontologies generated for the NIPF topics and made improvements in identify concepts, determining semantic relations and organizing concepts. These improvements are reflected on the rich ontologies generated by Jaguar.
- Investigated ontology creation scenarios involving augmentation of an existing ontology with information from additional documents. We added methodologies into Jaguar so that, when an existing ontology is provided, the concepts from it (in the hierarchy) are now added to the seed word set, and the semantic relations from the existing ontology are used in the concept extraction procedure, and this results in the creation of an improved seeded ontology.
- Added a capability to associate each created ontology concept with a named-entity tag, which is extracted when the document is processed through our tools pipeline. Also, Jaguar can now deal with natural-order based semantic relations or head-ordering based semantic relations.
- Improved our concept extraction/creation methodology by creating concept normalization technique to hold the value of the concept's lexeme in a normalized form for common comparisons; we made sure that newly created concept sets the appropriate lexeme and uses the normalized name from the existing concepts for any comparisons.
- Implemented a framework in Jaguar to extract the class/instance information during the ontology creation process in Jaguar. We used the Named-Entity information along with certain sanity rules to classify a particular concept to be a class or an instance in the ontology. Our concept extraction module makes a distinction between domain concepts that represent concept classes versus those that represent instances of those concept classes. This separation of concept classes from their instances is one of the crucial points required for the creation of rich knowledge. By default, when concepts are extracted from the text, the system considers all the extracted concepts to be classes. After all the knowledge has been extracted, the system re-processes all the extracted concepts (hierarchical, non-hierarchical and orphans) and uses the semantic relation (mainly ISA relations) and named-entity class information for each concept to identify if that concept needs to be re-classified as an instance.
- Implemented an automatic seed-concept extraction algorithm in Jaguar. Jaguar requires a seed-concepts set or a seed-ontology as input to create a domain specific ontology. Our efforts focused on automatically extracting these seed concepts from the document collection based on concept occurrence statistics (probability of concept's occurrence in the domain collection versus its occurrence probability in the real-world) and concept linking through semantic relations (adaptation of Google's PageRank for Semantic-Relations).
- Extended the automatic seed concepts extraction algorithm in Jaguar to implement an automatic skeleton ontology extraction algorithm. Jaguar requires a seed-concepts set or a seed-ontology as input to create a domain specific ontology. Our efforts focused on automatically extracting important seed concepts and semantic relations between these concepts from the document collection based on: Table-Of-Contents structure; concept occurrence statistics (probability of concept's occurrence in the domain collection versus its occurrence probability in the real-world); and concept linking through semantic relations (adaptation of Google's PageRank for Semantic-Relations).

4.2.2 Automatic Generation of Extended WordNet Knowledge Base for Text Understanding

The project/activity aims at capturing the semantics present in the WordNet synset gloss entries – the textual description of the synset concept.

- Designed and implemented a working XWN-KB as an upper ontology, on which any number of domain specific Knowledge Bases can be built.
- Each WordNet gloss entry is processed using Lymba's NLP tools include part-of-speech tagging, syntactic parsing, word sense disambiguation and semantic parsing. Knowledge representation derived from such processing of gloss entries is used to build XWN-KB.
- Automatic processing of gloss entries is checked manually for quality as part of the XWN-KB generation.
- Finished hand checking the semantic relations for 35000 noun glosses (out of a total of about 80000 glosses)
- As an application of XWN-KB, the lexical chains building component was enhanced to use the XWN-KB to form lexical chains between components to aid in reasoning from text. Initial evaluation indicates that the lexical chains generated using the semantics captured in XWN-KB are more meaningful than those generated by using gloss co-occurrence relations from WordNet.
- A Visualization tool has been developed in Java, to navigate and search concepts in the XWN-KB and the semantic information associated with each of them
- Lymba manually annotated the gloss entries from eXtended WordNet (XWN) for semantic relations. This annotated data set was processed to update XWN-KB – a knowledge base derived and extended based on semantics extracted from gloss entries for different concepts. While a number of semantic relations are extracted and propagated through IS-A hierarchies, we selected and generated three kinds of axioms from this knowledge base. Patterns and statistics derived from each occurrence of a given concept in the gloss entries in WordNet are used to generate these axioms. The axioms relate concepts that are AGENT-OF and THEME-OF a given word. These axioms are propagated through the noun hierarchy to get rich set of axioms with confidence measures associated with them.
- Implemented more kinds of axioms (apart from the SVO ones) for the XWN-KB. The notable ones among these are the INS (instrument) and CAU (cause).
- XWN-KB Browser: Incorporated axiom trace (tracing every axiom back to the gloss/sentence it comes from). The XWN-KB browser was demonstrated in the March PI meeting.
- The XWN-KB supported the glosses (definitions) for WordNet noun concepts only. This has been extended to gloss entries of Verb concepts generating a richer semantic network of concepts and relations from WordNet glosses. Generated and experimented with axioms for the verb concepts in the XWN-KB.

4.2.3 Ontology Creation Capability using the Distributed, Asynchronous LymbaGrid Framework

During the course of this year, we have created a distributed, asynchronous LymbaGrid framework to distribute our various NLP capabilities. The capability to create ontologies from the pre-defined & user-defined document collections using the ICE-based LymbaGrid framework was added to this framework. We created a distributed, asynchronous service to create ontologies from the documents documents provided by the users or those listed in a PowerAnswer answer-list using the ICE-based LymbaGrid framework. Using such a framework, the clients can request for an ontology-creation job based on the documents provided by the users or those listed in the answers received from the previous QuestionAnswering Job. The ontology-creation job needs to marshal all the sub-tasks required to process such a request and to fire-off all the required sub-tasks through Document Processing Workers and the Jaguar Worker. The status updates from the various workers is handled by the job and the resulting ontology is returned to the client when available.

To test such a LymbaGrid based ontology creation capability, we have benchmarked the process (varying the NLP pipeline) on several document collections including 3 NIPF domain collections. We have successfully run the LymbaGrid based ontology creation process on 32 NIPF domain collections among several other ontology modeling test domains.

4.2.4 Automatically Building of Ontologies for National Intelligence Priorities Framework (NIPF) topics

There are around 33 NIPF topics and the intelligence community is interested in organizing its activities around these topics. The ability to define ontologies for these topics and enable their use in the IC by classifying documents under those topics is viewed to be of significant interest to the IC. We built ontologies for NIPF topics based on their descriptions. A team of two annotators collected documents from the web (based on topic descriptions), and generated a set of seed words (domain-related words/concepts) for each topic. The documents were processed using NLP tools and master ontologies for all the topics were generated. The remainder of this section presents the details of this ontology creation process.

We browsed through all the NIPF topic descriptions and for each topic, we collected 500 documents from the web (Weapons topic an exception with 50 Wikipedia documents) and manually verified their relevance to the corresponding topic. We then use Jaguar to create an ontology, for each identified NIPF topic. Jaguar builds each ontology with rich semantic content extracted from the corresponding document collection while keeping the manual intervention to a minimum. These ontologies are fine-tuned to contain the level of detail desired by an analyst.

We first extract text from the input NIPF document collections and then filter/clean-up the extracted text. The NIPF text input to Jaguar comes from all possible document types, including MS Word, PDF and HTML web pages, and is therefore prone to having many irregularities, such as incomplete, strangely formatted sentences, headings, and tabular information. The filtering mechanism of Jaguar is a crucial step that makes the input acceptable for subsequent NLP tools to process it.

For each NIPF topic, Jaguar is provided with a seed set containing on average 51 concepts of interest. The seed sets are used to determine the set of sentences of interest in a topic's document collection. The sentences selected based on the topic's seeds go through the entire set of NLP processing tools (listed previously). The NLP processed data files are then passed through the concept discovery module, which identifies noun concepts in sentences which are related to the NIPF topic target words or seeds. Each processed sentence is scanned for noun phrases, and targeted noun concepts are added for subsequent processing into the ontology's hierarchy. The resulting data structure is processed and used to populate one or many semantic contexts, groups of relations or nested contexts which hold true around a common central concept. The seed set is then augmented with concepts that have hierarchical relations with the target words or seeds. The entire process is repeated n number of times (n=3).

The extracted NIPF topic noun concepts and semantic relations are fed to the classification module to determine the hierarchical structure. Certain hyponymy relations discovered via classification contain anomalies (causing cycles) or redundancies. Hence, we run them through a conflict resolution engine to detect and correct inconsistencies. An NIPF topic hierarchy is created link by link (relation by relation) and follows a conflict avoidance technique, wherein each new link is tested for causing inconsistencies before being added to the hierarchy.

Although single runs of Jaguar yield rich NIPF ontologies, Jaguar's real power lies in providing an ontology maintenance option to layer ontologies from many different runs. Jaguar can merge disparate ontologies or add new knowledge by using the aforementioned conflict resolution technique. The merge tool merges the two ontologies' concept sets, hierarchies (using conflict resolution), and their knowledge bases (set of semantic contexts). Merging is useful for distributed or parallel systems where small chunks of the input text may be processed on some portions of the system and then subsequently merged. It also provides a foundation for future work in contextual reasoning and epistemic logic. The resulting rich NIPF knowledge bases can be viewed at many different levels of granularity, providing an analyst with the level of detail desired.

4.2.5 Jaguar Ontology Creation, Editing and Benchmarking – Additional ontologies edited

Lymba investigated a benchmarking methodology to provide feedback to the ontology creation process in Jaguar and the underlying tools used by Jaguar. We have currently created several gold-standard ontologies capturing three semantic relations (ISA, CAU & PW). The gold-standard ontologies represent the human understanding of concepts and their relations in the domains of interest:

1. 17 different NIPF topics
2. 6 Business Domain topics:
    a. Finance
    b. Investment
    c. RealEstate
    d. Intellectual Property
    e. Mergers Aquisitions and Buyouts
    f. Banking

This set includes both topics of interest to the intelligence community as well as other general topics of interest to commercial applications. These gold-standard ontologies where created by multiple annotators resulting in several ontologies for the same domain. The annotators worked in isolated as well as collaborative environments. We have benchmarked the performance of Jaguar on these gold-standard ontology entries and this has given us a good indication of the good/bad performing features in our NLP pipeline and Jaguar.

4.2.6 Jaguar's NIPF Ontologies Evaluation

We evaluated the quality of Jaguar's NIPF ontologies by comparing them against manual gold annotations [6]. Viewing an ontology as a set of semantic relations between two concepts, the annotators:

1. Labeled an entry *correct* if the concepts and the semantic relation are correctly detected by the system else marked the entry as *Incorrect*
2. Labeled a *correct* entry as *irrelevant* if any of the concepts or the semantic relation are irrelevant to the domain
3. From the sentences *added new entries* if the concepts and the semantic relation were omitted by Jaguar

The annotation rules provide feedback on the automated concept tagging and semantic relation extraction and also are used for computing precision (Pr) and coverage (Cvg) metrics for the automatically generated ontologies. The equations below capture the metrics defined by Lymba to evaluate Jaguar ontologies. Nj(.) gives the counts from Jaguar's output and Ng(.) correspond to counts in user annotations.

$$\Pr(Correctness) = \frac{N_J(Correct) + N_J(Irrelevant)}{N_J(Correct) + N_J(Incorrect) + N_J(Irrelevant)}$$

$$\Pr(Correctness + \mathrm{Re}\,levance) = \frac{N_J(Correct)}{N_J(Correct) + N_J(Incorrect) + N_J(Irrelevant)}$$

$$Cvg(Correctness) = \frac{N_J(Correct) + N_J(Irrelevant)}{N_G(Correct) + N_G(Irrelevant) + N_G(Added)}$$

$$Cvg(Correctness + \mathrm{Re}\,levance) = \frac{N_J(Correct)}{N_G(Correct) + N_G(Added)}$$

Table 2: Performance of Jaguar's automatic topical ontology generation from text

Number of Annotators	Topic	Precision		Coverage	
		Correctness	Correctness+ Relevance	Correctness	Correctness+ Relevance
3	Weapons	0.610090	0.501499	0.702424	0.657122
1	Missiles	0.533867	0.485364	0.793775	0.777747
2	Illicit Drugs	0.471938	0.274506	0.801422	0.701122
1	Terrorism	0.388788	0.291019	0.822285	0.776206

Table 2 presents our initial evaluation results for 4 NIPF topics using the ISA, PART − WHOLE and CAUSE relations only. These measures are averaged over the results for different annotators. The first column in Table 2 identifies the number of annotators for each topic. Jaguar obtained the best precision for Weapons with 2620 concepts in 2562 evaluated relations (total of 3473 concepts in 3508 relations); followed by *Missiles* with 5982 concepts in 5881 evaluated relations (total of 7873 concepts in 8573 relations); IllicitDrugs with 5107 concepts in 5213 evaluated relations (total of 7935 concepts in 10677 relations); and finally Terrorism with 7929 concepts in 8306 evaluated relations (total of 11638concepts in 13711 relations).

4.2.7 Development of Jaguar-on-the-Web

We completed our effort to provide an easy-to-use, web-browser based user-interface for clients to upload documents, create/manage collections and create/manage ontologies. The ***Jaguar-on-the-Web*** user interface uses the LymbaGrid based ontology creation/management process in the back-end. Currently, Jaguar-on-the-Web interface allows users to create ontologies from predefined document collections or users can create/maintain their own document collections and build ontologies from these customized collections. A user has the option to choose and build an ontology from a pre-defined seeds-file, or to upload and specify their own seeds-file. The seeds-file can be a text file (with a list of keywords, one keyword per line) or can be an owl/lymba-xml format ontology file. Jaguar-on-the-Web allows users to browse/download the created ontology in owl/lymba-xml/pace-xml format. We have also added the capability for the users to view the complete ontology creation process logs in a live mode or in a batch-mode.

4.2.8 Exporting Jaguar Ontologies

In world of applications that can greatly benefit by using the knowledge stored in an ontology/KB, it is very important to create/export ontologies in formats/standards that are widely accepted (e.g. Resource Description Framework (RDF) and Web Ontology Language (OWL)). Jaguar can create/export ontologies in our proprietary XML format as well as the OWL and RDF n-triple formats. To provide database driven ontology support for applications, we created our own RDBMS table schemas for storing/querying Jaguar ontologies/KBs in databases and also create tools to export our ontologies/KBs into third party database system including Oracle's 11g Spatial RDF database systems. We have successfully managed to populate Oracle's RDF 11g database (which has support for RDF and OWL) by converting a Jaguar ontology into RDF N-triple format and then populating the RDF database with the N-triples. We have created tools to export Jaguar ontologies into relational databases, which can be then used by search solutions, reasoners, multiple users in collaborative editing environment, etc. Thus, we completed the task for loading/manipulating/storing ontologies in any normal RDBMS.

4.2.9 Ontology/Knowledge-Base Search

We initiated a task to provide a KB search service for any application, over some interface. This was implemented over a socket (dumb) client / (smart) server architecture, with requests and responses being transferred via xml messages. We designed and implemented a “Message interface” - XML parsers both at the client and the server side. The server loads an ontology into memory and provides a socket interface to requests. The client currently is a php application (HTML embedded scripting language) requesting from the server such things as (a) "Give me the entire ISA hierarchy of the KB", or (b) "Give me all concepts related to concept X via any relation type", etc. The client provides a web based GUI to explore the KB. We enhanced the search facility to support different kinds of queries, including:

- Get the ISA hierarchy for a Knowledge Base
- Get any concept's details (relations to other concepts)
- Get all the sentences (in KB representation) in which a given KB concept occurs
- Get relation tracing (sentence from which a given KB relation is derived)
- Relation Filtering (get concepts related to concept c via specific relations only)
- Ontology selection (selecting the ontology to query)

The search interface includes support for querying multiple ontologies (searching a concept in multiple ontologies at the same time).

We further enhanced the Query Interface to support tracing any piece of information (concept, relation, or semantic cluster) to the source documents and sentences that it was derived from. We associate and maintain statistics about concepts in the KB and use in KB Viewer/Browser. We also provided a Query Interface for the XWN-KB (eXtended WordNet Knowledge Base). It provides a transparent search interface to the user for searching regular WordNet relations (that are currently stored in the XWN-KB), and those (semantic relations) coming from the enhanced XWN-KB. It supports Lymba's lexical chains system in using the XWN-KB relations.

We enabled support for multiple ontologies on the same socket, instead of having to assign each ontology to a different port. Each query message includes the (unique) ID of the ontology to which the request must be forwarded, so as to forward the query to the appropriate ontology.

Ontology/Knowledge-Base Visualization and Editing Tool Using Search Interface: We created a web-based Ontology Viewer using the previous described socket-based Query Interface.

- Initial development uses the KB query/search support (connects to an KB/Ontology using a socket connection; and displays concept details)
- Enhanced to support display of Semantic Clusters (groupings of information that represent an entire sentence and/or a frame of reference)
- Enhanced the KB Viewer with editing and update capabilities
- Enhanced the KB Viewer with Editor to enable users to browse their PTKB and correct/update them
- Development of capabilities to update/delete nodes, relations, hierarchies in the KB.
- Reduced the observed hierarchy size for large ontologies by truncating concept hierarchies that have already appeared under a different concept. This issue arises because a concept can, at times, appear under more than one parent. It is unnecessary to display the entire sub-tree under 2 or more parents of the same concept.

Jager – A Database-driven Ontology/Knowledge-Base Visualization and Editing Tool: The population of Jaguar ontologies into an RDBMS has helped us to start efforts to develop a robust ontology editor. We have completed implementing an initial, completely functional version of Jager (pronounced "yeager"), a web-browser based application that provides scalable, multi-user, collaborative editing of Jaguar ontologies stored in an RDBMS like mySQL. It is based on the Django framework and written in a mix of Python, HTML and Javascript. Jager supports the following operations:

- Loading and saving/exporting of ontologies in different formats
- Adding, deleting or moving concepts, relations and concept sub-trees
- Search concepts in the ontology
- Support for multi-user collaborative editing of ontologies
- Display source sentences/documents for concepts, relations extracted from text
- Ontology Merging
- Ontology filtering based on concept list

4.2.10 Ontology-driven document classification

As an application of ontologies, we developed text classification algorithms that use the NIPF topic ontologies to classify documents in the IC domain. Models generated from features from topic ontologies are used for classifying new documents. The NIPF dataset was used to train Maximum Entropy classifiers. The documents were represented as concepts and were used to generate TF-IDF[10] feature vectors. The trained classifiers were used to process the tag|Connect[11] data in order to determine the set of topics a tag|Connect document belonged to. This information was used to update the Lucene[12] index to included two additional fields for each document, a topic field and a concepts field. This new index was used in the Synergist system to browse the tag|Connect documents based on concepts in the NIPF ontologies.

BAE's expertise finder uses the NIPF classification of documents to model and recommend experts for a given user question.

[10] TF-IDF is a weighting scheme employed in vector space approaches for information retrieval that incorporates the term frequency (TF) and inverse document frequency (IDF) values.

[11] tag|Connect is social bookmarking tool developed by General Dynamics. It enables users to bookmark documents and associate descriptive terms to indicate user's interest in the documents. The tags facilitate users to 're-find' documents as well as enable others in the social network to discover new resources.

[12] Lucene is a open source keyword-based document indexing and retrieval system that can index large document collections and support keyword-based search of the content.

5.0 REASONING AND ARGUMENTATION FROM TEXT

Objective: Use Lymba's rich knowledge representation to construct epistemic structures (Kripke structures[13]) that allow multi-agent and contextual reasoning. Develop a technology to construct axioms on demand from knowledge bases. Develop a theory and prototype for Argumentation Reasoning that, like human reasoning, involves assumptions, exceptions, uncertainties, and counter-arguments.

5.1 Reasoning in Context

Lymba envisioned a contextual reasoning mechanism, which prevents the assertion of contextual information without having fulfilled the conditions imposed by the context (without ensuring, for example, that the location of the contextual event is identical to spatial constraints of the locative context). More specifically, given the statement *It rains in Alaska*, the contextual reasoning module will guarantee that the *raining* event will be inferred only when the location is set to *Alaska*.

Given the layered knowledge representation detailed in Section 4.1, and, more specifically, Lymba's solution for representing contexts, no additional measures need to be taken to enforce the represented contexts. All logical predicates specified in the antecedent of the logical implication of the context ensure that the consequent will be considered by the prover in its search for a proof only when the contextual conditions are met. However, most of the information conveyed by a natural language text is not explicitly stated in the text. Consider the following sentence:

It was an Acela Express, number 176, scheduled to depart 30 minutes later on its return trip to the Big Apple.

Human interpretation of the sentence will be that the train *number 176 left the station 30 minutes later* despite the fact that the text mentions that it was only *scheduled to depart* at that time. In order to accommodate this type of reasoning, we implemented new types of axioms in ***COGEX***, Lymba's logic prover. This includes high-penalty default axioms that enable the prover to infer consequents by assuming some or all of the contextual conditions are met, unless there is evidence to the contrary. This is similar to 'what-if' analysis where one attempts to generate all possible assertions if a set of assumption are satisfied. The system generates an axiom for each context enabled with default reasoning. For the above example, the axiom assume_CTXT(c1) → planning_CTXT(c1) will enable COGEX to assume that the train left on time, according to the schedule and hence the past tense in the sentence can enable us to infer that the train left the station 30 minutes later. Formally, these axioms have the following format:

assume_CTXT(c1) → LF(source) & signalClass_predicate & contextType_predicate.

[13] Clarke, Grumberg and Peled: *Model Checking*, The MIT Press, 1999.

5.2 Reasoning with Events and Temporal Expressions

Given the rich knowledge representation derived from text, Lymba's natural language reasoner was augmented to exploit the new semantic information extracted from text. With respect to temporal expressions, COGEX builds, on demand, axioms that describe temporal calculus rules. These use an interval-based representation of identified temporal expressions and link related temporal concepts. A time-instant is represented by a 7-tuple for (year, month, date, hour, min, sec, timezone). The default for timezone is the local timezone and this is assumed in the following examples. The interval-based representation of *October 2008* (identified by x1) is given by *Time(BeginFn(x1),2008,10,01,00,00,00) & Time(EndFn(x1),2008,10,31,23,59,59)*. The *BeginFn* and *EndFn* predicates indicate that the time value is the beginning and ending times, respectively for what x1 identifies in text. The rule '*October 2008* entails *2008* is represented by the axiom:

Time(BeginFn (x1),2008,10,01,00,00,00) & Time(EndFn(x1),2008,10,31,23,59,59) →
Time(BeginFn(x2),2008,01,01,00,00,00) & Time(EndFn(x2),2008,12,31,23,59,59) &
INCLUDES_SR(x2,x1).

More complex examples and axioms can be expressed using this representation. *September 2008 is immediately-before 1st October 2008* is represented by the time-interval representation and the semantic relation (*I-BEFORE_SR*).

Time(BeginFn(x1),2008,09,01,00,00,00) & Time(EndFn(x1),2008,09,30,23,59,59) →
Time(BeginFn(x2),2008, 10,01,00,00,00) & Time(EndFn(x2),2008,10,01,23,59,59) & I-BEFORE_SR(x1,x2)

This represents I-BEFORE(*September 2008, 1st of October 2008*). Any temporal relations generated by the application of this type of temporal axioms can be later combined with any other temporal relations as COGEX searches for an entailment between a text and a hypothesis.

Reasoning about time, time intervals and their relationship to events they constrain requires not only a temporally enhanced knowledge representation, but also a knowledge base of temporal reasoning axioms. Lymba created a pool of 94 temporal axioms that link each temporal relation with its inverse, for example, *before_sr(x1,x2) ↔ after_sr(x2,x1)* and define the temporal relation resulting from the combination of two temporal relations (e.g. *before_sr(x1,x2) & before_sr(x2,x3) → before_sr(x1,x3)*. These axioms were derived from Allen's interval algebra[9].

5.3 Using Reasoning to Identify Event-Event Temporal Relations

We experimented with three possible interactions between the machine learned models for temporal relation resolution and the temporal reasoning engine:

(1) Expanding the training data using the temporal closure generated by the reasoning engine. Once we manually resolved the temporal inconsistencies from the TimeBank training data (used to train our event relation extraction modules), COGEX computed its temporal closure using the set of temporal axioms detailed in Section 5.2. This process increased the temporal module's training data by 3.5 times.

(2) Validating the automatically identified temporal relations for a test set by checking for temporal inconsistencies and replacing the lowest confidence relation, which creates an inconsistency. We implemented in our reasoning system, a function which outputs the temporal inconsistencies found by the prover which are then analyzed and the lowest confidence relation is replaced by either the relation generated by the prover for that pair of events or by the next highest confidence relation automatically identified for that pair of events

(3) Automatically aligning the temporal relations identified for a test collection: learned models output top *n* - in terms of confidence - temporal relations for a given pair of events; the reasoning engine selects one relation as final based on the quality of the temporal closure it generates. We implemented a function which finds the best temporal closure, given the set of temporal relations that produces it. We investigated several scoring functions, which make use of the confidence of the relations that are closed, confidence assigned by the machine learned models and the size of the temporal closure derived from the relations

The results and discussion on these experiments can be found in [3].

5.4 Argumentation-Based Reasoning from Text

Lymba has developed an initial argumentation detection module that identifies the argument structure of an input document. Our approach uses both NLP and reasoning. In an initial detection phase, we use a rich set syntactic-semantic tree patterns to label text snippets as *claims*, *grounds*, or *rebuttals*. For instance, the following are some of the syntactic patterns for grounds (G) and claims (C):

In conclusion/Therefore/Hence/Thus, C
Subsequently/Consequently/Evidently/Apparently, C
The upshot/conclusion/suspicion is C

C, as/since/because G
G confirms/corroborates/demonstrates/indicates/proves/ shows/supports/establishes/implies C
G is reason/proof/evidence of/for C

on account of/in view of/by virtue/reason/cause of G
owing/thanks/due to G
First/First of all/Secondly/Next/Last of all/Finally, G

All syntactic tree regular expressions were derived by analyzing more than 500 files of Wall Street Journal OpEd corpus. In the linking phase, the system inputs the detected claims, grounds, and rebuttals into Lymba's reasoning module and tests for logical relationships in surrounding text.

Furthermore, we synthesized the argumentation schemes from Perelman's New Rhetoric[14] and proposed methods of detection based on semantic relations and content for reciprocity (we should treat two situations which are counterparts to each other the same), justice (we should treat the same kind of situation in the same way as we did before) and waste (we argue against stopping something because all our efforts would be wasted) argumentation schemes. For the purpose of prioritizing the argumentation schemes that we identified in Perelman's work, we performed an initial analysis of the Glassbox data, which will provide statistics about the schemes used by analysts in their reports.

Automated methods for detecting argument structures in text will enable machine intelligence to be used in representing and analyzing arguments in intelligence reports. Also, argumentation-based reasoning can be used in analyzing the claims and grounds that analysts generate and maintain in systems like BAE's Polestar and Analysis of Competing Hypotheses (ACH). ACH provides an unbiased methodology for evaluating multiple competing hypotheses for observed data.

[14] C. Perelman. *The new rhetoric: A theory of practical reasoning*. In P. Bizzell and B. Herzberg (Eds.), The rhetorical tradition, 2001.

6.0 ANALYST PRIOR AND TACIT KNOWLEDGE TOOLS AND APPLICATIONS

Objective: Develop technologies to capture and represent the prior and tacit knowledge of analysts from their textual products, placing it into Prior and Tacit Knowledge Bases (PTKBs). Use PTKBs to detect novel information, identify domain experts, and generate models about the analyst, her documents, and tasking. Also develop tools for merging analysts' individual prior and tacit knowledge bases into a group or organization knowledge base, and allow analysts to search and use each others' knowledge bases. Extend and enhance knowledge transfer and analyst productivity with social networking technologies.

6.1 Generating User Knowledge Models from Text

Exploiting known user-document associations, Lymba has designed various models of user knowledge and created automatic tools that derive them from natural language texts. The development, experiments and evaluation of user knowledge models from text is based on social bookmarking data. Users bookmark documents and associate words/phrases as tags to provide their description of the document. The (user, document, tag) triple provides sufficient information to model users and other resources in the social media.

Given the rich semantic information that Lymba's NLP pipeline extracted from natural language content, we indexed in a user's knowledge base all event structures identified in the user's documents. For each event structure, we consider the focus of the event (verbs or action nouns), the time attribute of the event (time interval automatically associated with the event focus which estimates its absolute temporal location), all its semantic relation dependents (participants in the event, agent and theme, as well as other attributes, location, manner, instrument, etc.), and all of their semantic relation dependents (attributes of the event's agents, themes, instruments, etc.). We note that a user's knowledge base makes use of coreference links found by Lymba's coreference resolution module ('he' will not be indexed as the agent of 'go', but 'john' for a document, which contains 'John is a good boy. He went to the market yesterday.'). Our implementation of user knowledge models stores the user knowledge bases into optimized relations databases (MySQL) for fast and scalable access.

However, within social bookmarking settings, users associate tags with their bookmarked documents. These labels denote the user's specific interest in the document. Thus, we developed user models by exploiting (user, document, tag) associations.

We start with methods to induce tags from the content of a document. This process is based on projecting the knowledge bases (automatically generated ontologies – Section 4.0) derived from documents relevant to a tag or user to the *tag space*. This provides a description of the document in terms of tags. Users can also be represented in the tag space if we consider their set of tags as well as their document's representation in the tag space. Each tag space representation is weighted set of tags. The tag's weight within a document or user model can be computed based on the tag's frequency or its probability. The impact of this variation in the tag-weighting scheme was measured as part of the RDE NIST evaluation (Section 9.2).

However, we varied the document and user model representations within social bookmarking settings to consider not only the tag space, but also the concept space (the set of all concepts identified in a document's content). We explored the following types of models:

(1) Representations in the concept space (no projection to the current tag space is performed).

(2) Because models generated in the concept space are large and often include information not directly linked to the overall topic/interest of the bookmarking data (tag space), we use the information gain measure of a concept given the existing tag space to reduce the document and user representations in the concept space.

(3) By projecting the models built using the information gain measure to the tag space, new representations of document and users were created.

We note that Lymba developed a novel tag/concept conflation method, which normalizes the set of tags used to bookmark documents as well as concepts identified in document content. This normalization process includes the following steps: (1) tokenization, (2) spell checking, (3) capitalization restoration, (4) lemmatization, (5) abbreviation and acronym expansion, and (6) synonymy conflation. By employing this procedure, we address the tag space sparseness problem. All representations described above use the normalized tag space.

6.2 Novelty detection

Given a document subset and a user, the intent of this task is to identify and score a document based on the novelty of information to the given user. For this purpose, we use the user knowledge models described in Section 6.1 to denote what is known to the user. When measuring the novelty of a document against a user's knowledge base, we compare the information extracted from the document with the one stored in the user's knowledge base. The granularity of the novel information can vary:

(1) *concept level novelty*: we restrict the information to be checked to only certain types of concepts (named entity, event focus, noun or verb concepts, etc.) or to important concepts (identified using statistical measures);

(2) *semantic relation novelty*: we exploit the set of semantic relations identified in the document and, once we map them to the user's knowledge base, we report new semantic relations between novel concepts as well as new relations that link known concepts;

(3) *event structure novelty*: we determine known events, new properties and relations of events.

Below, we show the algorithm used to compute the degree of novelty of a text snippet (document, paragraph, or sentence) for user $\mathbf{U_i}$.

- For each event structure $\mathbf{e_j}$ identified in **text**,
    1. Determine whether $\mathbf{e_j} \in \mathbf{KB(U_i)}$ ($\mathbf{U_i}$'s knowledge base)
        - Find known events with similar focus → $\mathbf{\{e_{jk}\}}$
        - Search among $\mathbf{\{e_{jk}\}}$ for $\mathbf{e_j}$'s best match
            - Temporal attributes → temporal overlap of events
            - Event arguments → common arguments, novel arguments, conflicting arguments
            - Score each $\mathbf{e_{jk}}$ based on its structural similarity with $\mathbf{e_j}$
    2. If a matching known event is found : $\mathbf{novelty(e_j) = 1 - match(e_{jk},e_i)}$, where $\mathbf{e_{jk}}$ is the event with the maximum score
    3. If no match is found $\mathbf{(e_j \notin KB_i)}$: $\mathbf{novelty(e_j) = 1}$
- **novelty(text,KB(U$_i$)) = average-novelty-of-events**

By extending our initial novelty detection module to support novelty at the granularity of a paragraph, answers returned for a given question can be ranked using their novelty scores. Thus, the passages returned as answers to a user's query in the PACE Research Environment in Synergist (Figure 6) can be automatically evaluated for novelty and a second score can be returned in addition to the relevancy score.

Our *evaluation of the novelty detection* component used the TREC 2002-2004 Novelty Track datasets (http://trec.nist.gov/data/novelty.html). TREC Novelty track aims to identify novel sentences within an ordered list of sentences relevant to given topics. We note that this task does not explicitly include a user. We cast the problem posed by the TREC Novelty track organizers into a user-based novelty detection by creating a temporary user which does not possess any knowledge at the beginning of the task. This temporary user gains small pieces of knowledge in each iteration of the system through the ordered list of relevant sentences. At step **i**, the temporary user's knowledge base includes sentences *1*, ..., *i-1*. Sentence *i* is checked against these sentences, is assigned a novelty score and then added to the temporary user's knowledge base. Learning a threshold score for novelty, a sentence is deemed novel if its score exceeds the threshold, else it is deemed redundant. For the TREC 2003 and 2004 Novelty track data (50 topics each - http://trec.nist.gov/data/novelty.html), the average F-measure of our system is 79% and 60% respectively (for 2004 data, precision = 0.466, recall = 0.899, maximum F-measure = 0.96; baseline had an average F-measure of 0.577 and the best performing system was at 0.622 average f-measure.

6.3 Resource Modeling and Recommendation

In Section 6.1, we detailed the different models that we built for documents and users from social bookmarking data: (user, document, tag) triples. These rich representations can be exploited in order to produce recommendations. Given the three types of resources available in a social bookmarking setting (users, documents, and tags), various recommendation pairs can be made as presented in Table 3.

Table 3: Possible recommendations within bookmarking systems

Recommending → for ↓	**User**	**Document**	**Tag**
User	Users with shared interests	Documents of interest to user	Tags of interest to user
Document	Users who would be interested in this document	Document similarity – clustering documents in tag space	Tag recommendation: Other tags that can be associated with the document
Tag	Users who would be interested in this tag	Documents relevant to this tag	Tag Clustering and enhanced tag description

A significant information access problem in the intelligence community is to enable resource finding the user who would make the most use of it. Lymba’s resource modeling and recommendation system facilitates “*resource finding user*” capability. Lymba has implemented recommendation strategies for most of possible combinations shown in Table 3. For example, the system can recommend tags for a given document by first processing the document content to extract concepts, map the concepts to the existing tag space and identify a subset of those tags that were not manually assigned to the document as tag recommendations for that document. A similar procedure can be used for ‘tag recommendations for a user’. For tag recommendations for a (user, document) pair, we return a ranked list of tags part of the user’s model as well as the document’s representation.

For settings where the social bookmarking activity is guided by a task (described by a short textual description), we introduced the task in the process of recommending documents/tags. Each task was introduced in our system as a document whose content included the few sentences that describe the task. The document and user models are not affected by these 'new documents'. The document/tag recommendation procedure changed to take into account the task model in addition to the user/document models previously used.

We performed *an initial evaluation* of the system that does not rely on human input. We used the Monterrey data (TagConnect, MIIS2 experiment) for this purpose. For each user stored in this dataset, we split his or her list of documents (each user is associated with the set of documents that they tag) into a set with the first 40% of the documents (the list of documents is ordered in chronological order) used to build the user model and the a set with the remaining 60% of the documents. Once the 'known' set of documents is indexed into the user's knowledge base, Lymba's RDE makes recommendations based on the user model generated based on these documents. The remaining 60% of the user documents (as collected during the Monterrey experiment) are used to score the document-for-user recommendations returned by Lymba's RDE.

For each list of document-for-user recommendations, we computed the precision, recall, F-measure and average precision of the documents returned as relevant. In addition to these measures, we compute the time that the user would have gained by using the recommendation system (we compute the difference between the time of the recommendation and the actual timestamp of the bookmark - the moment when the user tagged the document).

We note that the set of documents that were not tagged by a user include both relevant and not-relevant documents. The set of 'future' documents (60% of the user documents in our split of the data) includes documents that the user found to be of interest, but not all relevant documents are included in this set. Our document recommendations include relevant documents that the user did not happen to encounter during the experiment.

Furthermore, Lymba participated in the *ECML PKDD Discovery Challenge 2008 Tag Recommendation task – RSDC'08 Challenge* (http://www.kde.cs.uni-kassel.de/ws/rsdc08). For this challenge, we used the textual content associated with each bookmark/bibtex (title, URL, author, description, etc.). We experimented with different tag recommendation methods by restricting the document and user models to the tag space created from the training data and by relaxing the tag recommendations to include the concepts present in the textual content associated with each bookmark. Our submission was scored to 0.19325 (f1-measure) - *the best performing system for the tag recommendation task.*

User evaluation performed under the guidance of NIST for tag and document-recommendations is described in Section 9.2

7.0 EVIDENCE MARSHALLING AND HYPOTHESES MANAGEMENT

Objective: Marshal evidence for and against analyst assertions or hypotheses using advanced question-answering and textual entailment technologies. Develop methods to help combat analysts' premature fixation on an early hypothesis by presenting alternative points of view. Develop methods for hypothesis generation using a combination of data-driven and prior/tacit knowledge approaches. Anticipate analysts' future information needs and proactively retrieve relevant data.

7.1.1 Evidence Marshalling

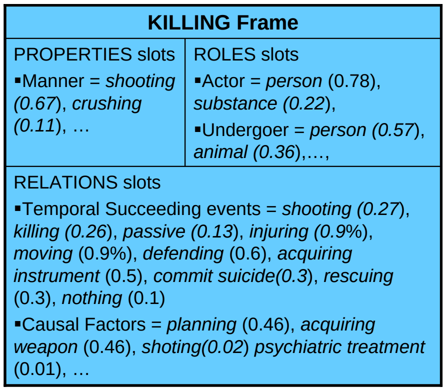
Under this task Lymba developed Wolverine as an evidence marshalling system that given a statement or hypothesis return evidences qualifying them to support or refute the hypothesis as well as support the marshalling of similar evidences for a given (hypothesis, evidence) pair:

- Text snippets relevant to a given hypothesis are selected using answer selection methods employed in natural language question answering
- Semantic representation of hypothesis and relevant text snippets are generated using natural language processing. Reasoning on this semantic representation, Wolverine is able to qualify with confidence scores if given evidence supports or refutes given hypothesis. Confidence scores are estimated based on the likely of the evidence entailing the hypothesis or its negation in COGEX, Lymba's resolution-based logic prover,
- Integrated evidence marshalling capability with BAE's polestar system to support 'find more evidences' feature. Users can identify a particular evidence they have collected for a given hypothesis and request for other similar evidences

7.1.2 EventNet

Given a situation, hypothesis generation corresponds to identifying what entities and events have led to the current situation as well as identifying its possible futures. With the focus on events, Lymba developed and demonstrated EventNet as a linguistic resource for modeling events, their properties and relations that can help in situation understanding and prediction problems.

EventNet is a new generalized network of events built on top of WordNet represented by frames linked to the corresponding WordNet concepts and connected by properties. The frames include slots corresponding to different properties (attributes, states, roles, relations) that an event has, each property having different values with different probabilities of occurrence in text. Lymba is building this data resource from the knowledge extracted from existing resources and/or open or domain-specific texts processed using the NLP tools, is clustered (using classes of nouns or verbs, clusters of adjective or adverbs, set or intervals of numbers, etc), generalized (using WordNet hierarchies), and stored in a relational database (using MySql). For example, Figure 5 shows a part of the frame for *Killing* with knowledge from Wikipedia. The Killing event includes property slots such as manner, roles of the participants of the event as well as relations with other events.



KILLING Frame

PROPERTIES slots

- Manner = *shooting (0.67*), *crushing (0.11*), …

ROLES slots

- Actor = *person* (0.78), *substance (0.22*),
- Undergoer = *person (0.57*), *animal (0.36*),…,

RELATIONS slots

- Temporal Succeeding events = *shooting (0.27*), *killing (0.26*), *passive (0.13*), *injuring (0.9*%), *moving* (0.9%), *defending* (0.6), *acquiring instrument* (0.5), *commit suicide(0.3*), *rescuing* (0.3), *nothing* (0.1)
- Causal Factors = *planning* (0.46), *acquiring weapon* (0.46), *shoting(0.02*) *psychiatric treatment* (0.01), …

Figure 5: Portion of the EventNet frame for Killing

This knowledge organization can be used in a number of applications including situation and event understanding, hypotheses generation, and reasoning. The EventNet resource can help an analyst understand different types of events, situation in which a particular type of event can take place, how an event happened, and what may happen next. For example, the knowledge from Figure 5 can help generate hypotheses about the actors and undergoers of a killing event (e.g. very likely to be people), what can happen after a killing event (likely to be another shooting or killing), and why an event happened (likely to have been a planning activity or acquisition of a weapon).

Given two events from a given situation description, EventNet can identify the set of possible events that may have occurred between them in time or due to causality. Evidences supporting or refuting the occurrence of those events provides better situation understanding.

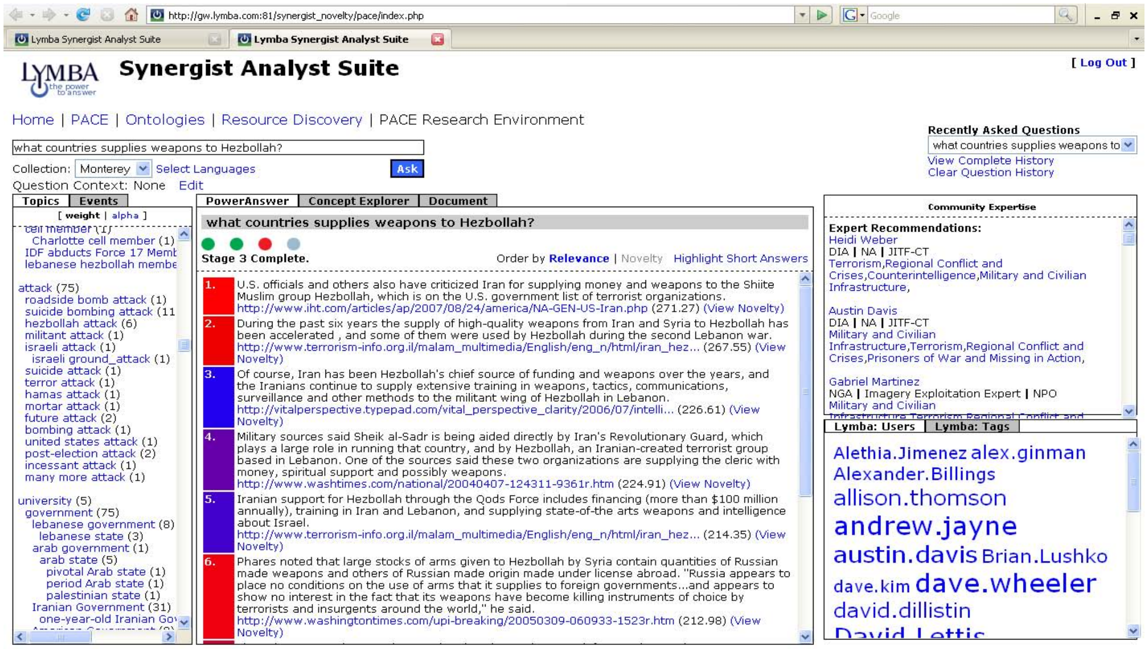
8.0 SYNERGIST AND MULTI-LINGUAL INFORMATION ACCESS

Objective: Lymba developed **Synergist Analyst Suite** as a integrated web-based application that provides analysts the ability to manage the different tasks assigned to them, search and discover information in the context of their tasks, and provide tacit collaboration through recommendations of documents, tags and users. These capabilities are realized with the ability to process and extract semantic information from text in multi-lingual settings. Lymba's METRE machine translation system is used for this purpose and it supports Arabic, Chinese, Korean, Farsi and a number of European languages.

8.1 Synergist system

Synergist Analyst Suite enables users to login and manage the different tasks they are working on. Through the PACE Research environment, a user can pose a natural language question to a multi-lingual document collection and get different types of relevant information. Figure 6 illustrates the "Augmented Pull" capability of Synergist where in addition to answers presented in the PowerAnswer tab, the left pane presents an hierarchy of related topics/concepts in the documents relevant to the given query. While the answers are ranked for their relevance, color-coding of the rank of the answer provides visual clues on the novelty of the information to the user.

The right panes provide expertise recommendation from BAE systems and Lymba's recommendation of users and tags in the Social media. BAE's expertise recommendations are based on the documents relevant to given questions and their NIPF categories automatically assigned by Synergist. Lymba's user and tag recommendations are based on the model Synergist generates for the user and the tags associated with relevant documents for the given question.


http://gw.lymba.com:81/synergist_novelty/pace/index.php
Google
Lymba Synergist Analyst Suite
Lymba Synergist Analyst Suite
LYMBA the power to answer
Synergist Analyst Suite
[Log Out]
Home | PACE | Ontologies | Resource Discovery | PACE Research Environment
Ask
Topics
Events
[weight | alpha]
PowerAnswer
Concept Explorer
Document
Order by Relevance | Novelty Highlight Short Answers
Recently Asked Questions
View Complete History
Clear Question History
Community Expertise
Lymba: Users
Lymba: Tags

what countries supplies weapons to Hezbollah?

Collection: Monterey Select Languages

Question Context: None Edit

cell member (1)
Charlotte cell member (1)
IDF abducts Force 17 Memb
lebanese hezbollah membe

attack (75)
roadside bomb attack (1)
suicide bombing attack (11
hezbollah attack (6)
militant attack (1)
israeli attack (1)
israeli ground_attack (1)
suicide attack (1)
terror attack (1)
hamas attack (1)
mortar attack (1)
future attack (2)
bombing attack (1)
united states attack (1)
post-election attack (2)
incessant attack (1)
many more attack (1)

university (5)
government (75)
lebanese government (8)
lebanese state (3)
arab government (1)
arab state (5)
pivotal Arab state (1)
period Arab state (1)
palestinian state (1)
Iranian Government (31)
one-year-old Iranian Gov

what countries supplies weapons to Hezbollah?

Stage 3 Complete.

1. U.S. officials and others also have criticized Iran for supplying money and weapons to the Shiite Muslim group Hezbollah, which is on the U.S. government list of terrorist organizations. http://www.iht.com/articles/ap/2007/08/24/america/NA-GEN-US-Iran.php (271.27) (View Novelty)

2. During the past six years the supply of high-quality weapons from Iran and Syria to Hezbollah has been accelerated , and some of them were used by Hezbollah during the second Lebanon war. http://www.terrorism-info.org.il/malam_multimedia/English/eng_n/html/iran_hez... (267.55) (View Novelty)

3. Of course, Iran has been Hezbollah's chief source of funding and weapons over the years, and the Iranians continue to supply extensive training in weapons, tactics, communications, surveillance and other methods to the militant wing of Hezbollah in Lebanon. http://vitalperspective.typepad.com/vital_perspective_clarity/2006/07/intelli... (226.61) (View Novelty)

4. Military sources said Sheik al-Sadr is being aided directly by Iran's Revolutionary Guard, which plays a large role in running that country, and by Hezbollah, an Iranian-created terrorist group based in Lebanon. One of the sources said these two organizations are supplying the cleric with money, spiritual support and possibly weapons. http://www.washtimes.com/national/20040407-124311-9361r.htm (224.91) (View Novelty)

5. Iranian support for Hezbollah through the Qods Force includes financing (more than $100 million annually), training in Iran and Lebanon, and supplying state-of-the arts weapons and intelligence about Israel. http://www.terrorism-info.org.il/malam_multimedia/English/eng_n/html/iran_hez... (214.35) (View Novelty)

6. Phares noted that large stocks of arms given to Hezbollah by Syria contain quantities of Russian made weapons and others of Russian made origin made under license abroad. "Russia appears to place no conditions on the use of arms that it supplies to foreign governments...and appears to show no interest in the fact that its weapons have become killing instruments of choice by terrorists and insurgents around the world," he said. http://www.washingtontimes.com/upi-breaking/20050309-060933-1523r.htm (212.98) (View Novelty)

what countries supplies weapons to

Expert Recommendations:

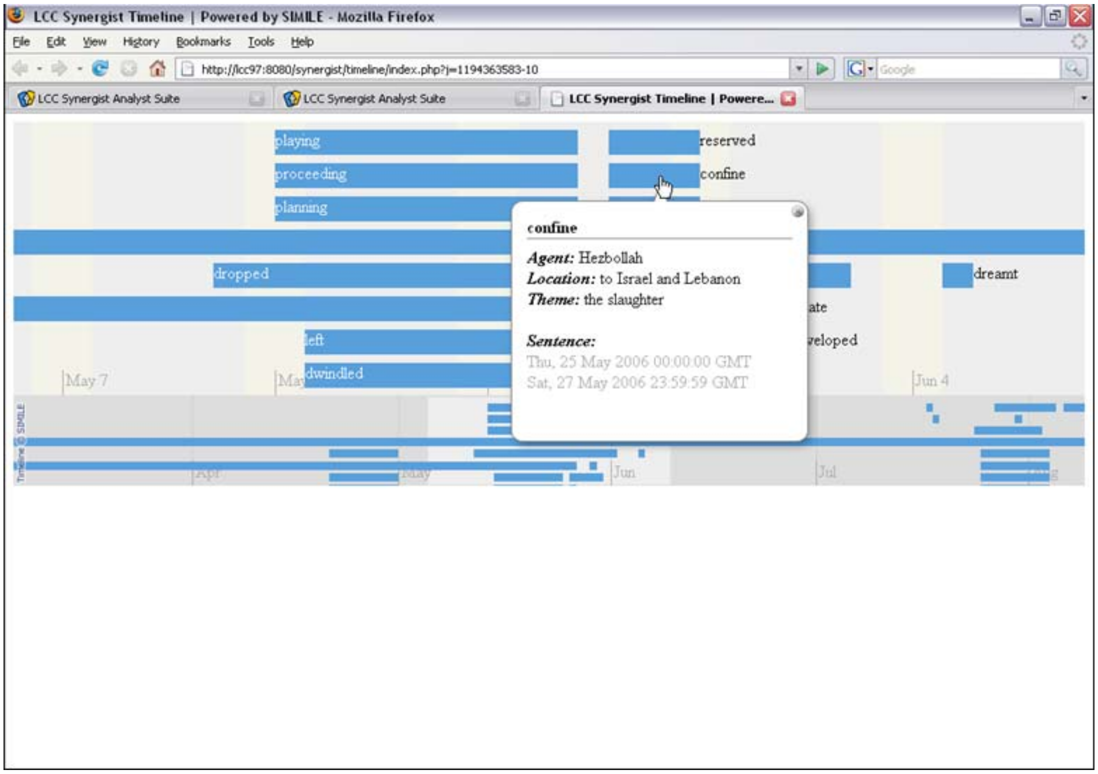
Heidi Weber
DIA | NA | JITF-CT
Terrorism,Regional Conflict and Crises,Counterintelligence,Military and Civilian Infrastructure,

Austin Davis
DIA | NA | JITF-CT
Military and Civilian Infrastructure,Terrorism,Regional Conflict and Crises,Prisoners of War and Missing in Action,

Gabriel Martinez
NGA | Imagery Exploitation Expert | NPO
Military and Civilian Infrastructure,Terrorism,Regional Conflict and

Alethia.Jimenez alex.ginman Alexander.Billings allison.thomson andrew.jayne austin.davis Brian.Lushko dave.kim dave.wheeler david.dillistin David.Lettis

Figure 6: Synergist Analyst Suite: PACE Research Environment


LCC Synergist Timeline | Powered by SIMILE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://lcc97:8080/synergist/timeline/index.php?j=1194363583-10
Google
LCC Synergist Analyst Suite
LCC Synergist Analyst Suite
LCC Synergist Timeline | Powere...
playing
proceeding
planning
reserved
confine
dropped
dreamt
ate
left
veloped
dwindled
May 7
May
Jun 4
Timeline © SIMILE
Apr
May
Jun
Jul

confine

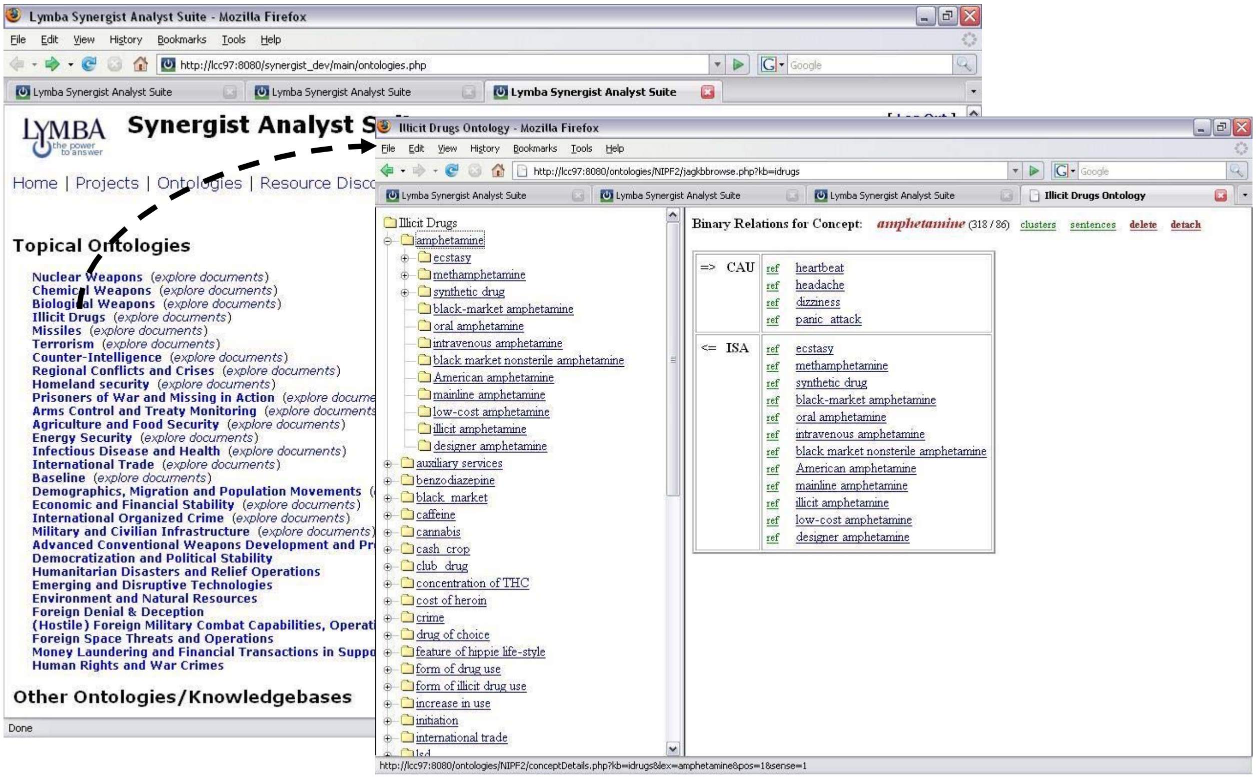
Agent: Hezbollah
Location: to Israel and Lebanon
Theme: the slaughter

Sentence:
Thu, 25 May 2006 00:00:00 GMT
Sat, 27 May 2006 23:59:59 GMT

Figure 7: Synergist Analyst Suite: Event Timeline Browser

In addition to the topical index, the left pane includes the option to display the events extracted from the relevant documents. With time attributes associated with events detected in text, users can visualize the events in a timeline to obtain better situation understanding. Figure 7 presents a snapshot of the timeline browser that can show events from a document or a subset of documents in a timeline view.

Synergist includes an ontology browser (ref. Figure 8) to navigate topical ontologies. The NIPF ontologies automatically generated from topic descriptions and domain-specific document collections can be browsed using this interface. The topical ontologies are used for classifying documents to assign their NIPF category. Synergist includes interfaces to browse document collections using NIPF ontologies.

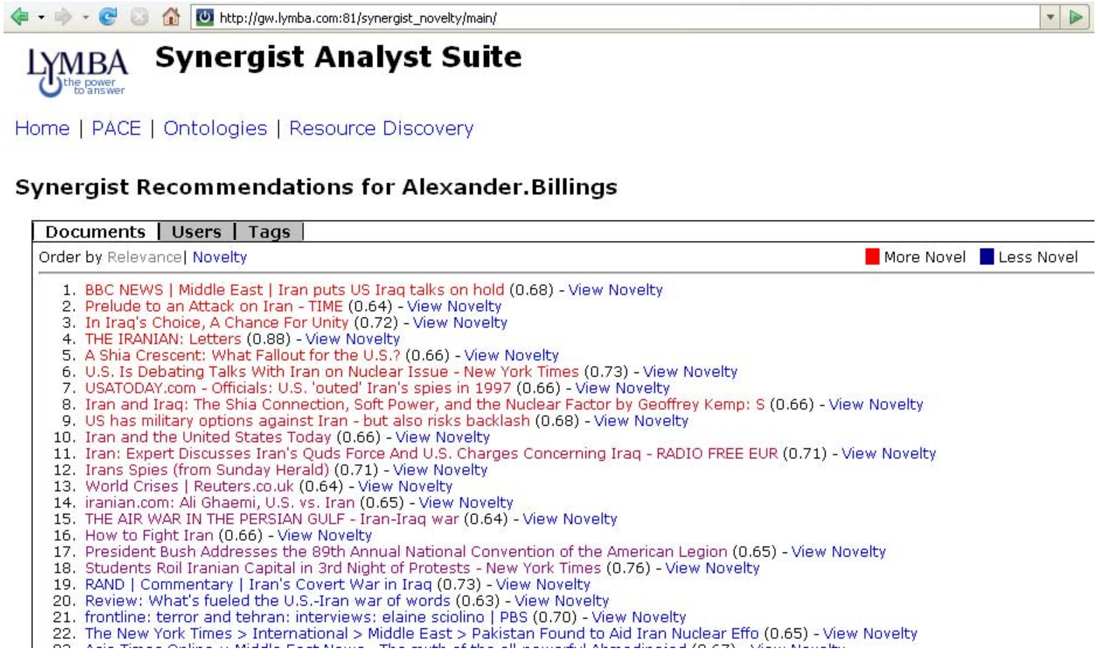

Lymba Synergist Analyst Suite - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://lcc97:8080/synergist_dev/main/ontologies.php
Lymba Synergist Analyst Suite
LYMBA the power to answer
Synergist Analyst S
Home | Projects | Ontologies | Resource Disco
Topical Ontologies
Nuclear Weapons (explore documents)
Chemical Weapons (explore documents)
Biological Weapons (explore documents)
Illicit Drugs (explore documents)
Missiles (explore documents)
Terrorism (explore documents)
Counter-Intelligence (explore documents)
Regional Conflicts and Crises (explore documents)
Homeland security (explore documents)
Prisoners of War and Missing in Action (explore docume
Arms Control and Treaty Monitoring (explore documents
Agriculture and Food Security (explore documents)
Energy Security (explore documents)
Infectious Disease and Health (explore documents)
International Trade (explore documents)
Baseline (explore documents)
Demographics, Migration and Population Movements (
Economic and Financial Stability (explore documents)
International Organized Crime (explore documents)
Military and Civilian Infrastructure (explore documents)
Advanced Conventional Weapons Development and Pr
Democratization and Political Stability
Humanitarian Disasters and Relief Operations
Emerging and Disruptive Technologies
Environment and Natural Resources
Foreign Denial & Deception
(Hostile) Foreign Military Combat Capabilities, Operati
Foreign Space Threats and Operations
Money Laundering and Financial Transactions in Suppo
Human Rights and War Crimes
Other Ontologies/Knowledgebases
Done
Illicit Drugs Ontology - Mozilla Firefox
http://lcc97:8080/ontologies/NIPF2/jagkbbrowse.php?kb=idrugs
Illicit Drugs Ontology
Illicit Drugs
amphetamine
ecstasy
methamphetamine
synthetic drug
black-market amphetamine
oral amphetamine
intravenous amphetamine
black market nonsterile amphetamine
American amphetamine
mainline amphetamine
low-cost amphetamine
illicit amphetamine
designer amphetamine
auxiliary services
benzodiazepine
black market
caffeine
cannabis
cash crop
club drug
concentration of THC
cost of heroin
crime
drug of choice
feature of hippie life-style
form of drug use
form of illicit drug use
increase in use
initiation
international trade
lsd
Binary Relations for Concept: amphetamine (318 / 86) clusters sentences delete detach
http://lcc97:8080/ontologies/NIPF2/conceptDetails.php?kb=idrugs&lex=amphetamine&pos=1&sense=1

=> CAU	ref ref ref ref	heartbeat headache dizziness panic attack
<= ISA	ref ref ref ref ref ref ref ref ref ref ref ref	ecstasy methamphetamine synthetic drug black-market amphetamine oral amphetamine intravenous amphetamine black market nonsterile amphetamine American amphetamine mainline amphetamine illicit amphetamine low-cost amphetamine designer amphetamine

Figure 8 Synergist: NIPF ontologies and Topical Ontology Browser

Figure 9 illustrates the resource recommendation screen for users. User activity of bookmarking and tagging documents is used to model their knowledge and interests and the resource recommendation system returns documents, tags and users that match or are relevant to their user models. In addition to relevance, the returned documents can be ranked for their novelty to the user.

Figure 10 presents the user interface of Lymba's resource discovery engine that adds to an existing social bookmarking and tagging system the ability to recommend resources – users, documents, and tags against a given user based on their user model or for a given (document, user) pair. Viewing users, documents and tags as different dimensions for browsing and analyzing information, the interface enables users to navigate through the links among users as well as among tags and documents.


http://gw.lymba.com:81/synergist_novelty/main/
LYMBA the power to answer
Home | PACE | Ontologies | Resource Discovery
Documents | Users | Tags

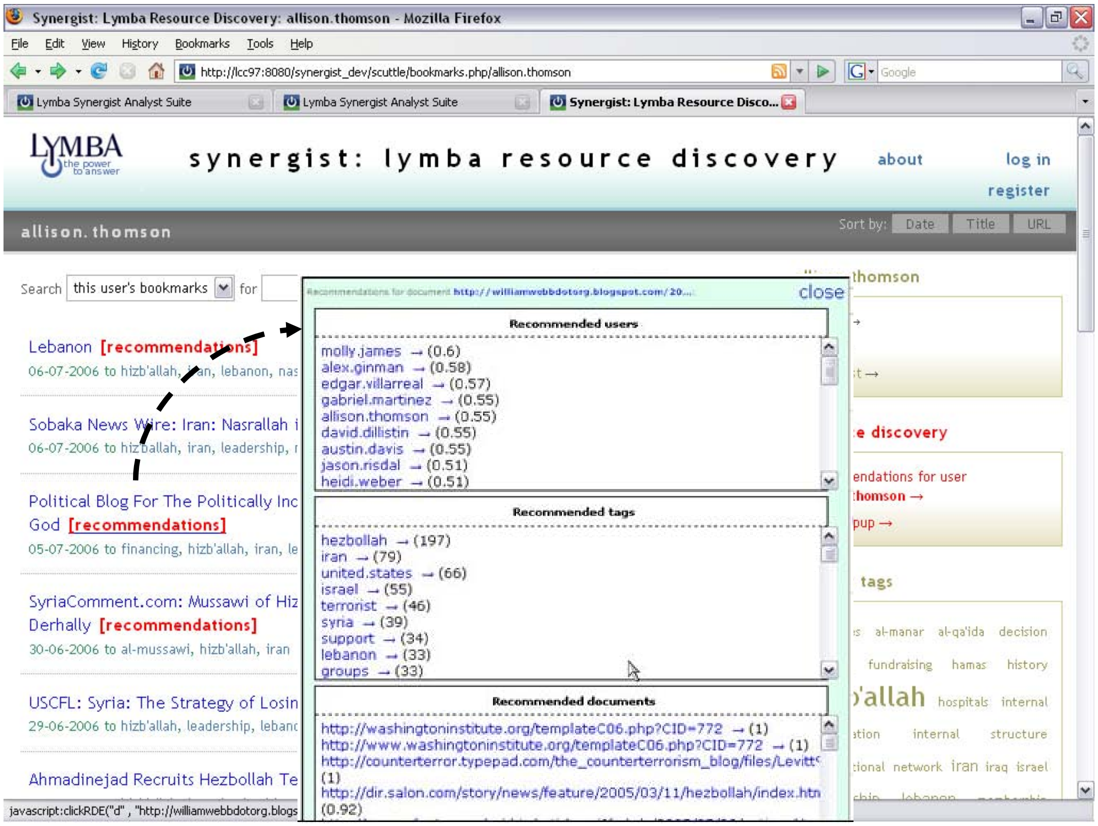
Synergist Analyst Suite

Synergist Recommendations for Alexander.Billings

Order by Relevance| Novelty 🟥 More Novel 🟦 Less Novel

1. BBC NEWS | Middle East | Iran puts US Iraq talks on hold (0.68) - View Novelty
2. Prelude to an Attack on Iran - TIME (0.64) - View Novelty
3. In Iraq's Choice, A Chance For Unity (0.72) - View Novelty
4. THE IRANIAN: Letters (0.88) - View Novelty
5. A Shia Crescent: What Fallout for the U.S.? (0.66) - View Novelty
6. U.S. Is Debating Talks With Iran on Nuclear Issue - New York Times (0.73) - View Novelty
7. USATODAY.com - Officials: U.S. 'outed' Iran's spies in 1997 (0.66) - View Novelty
8. Iran and Iraq: The Shia Connection, Soft Power, and the Nuclear Factor by Geoffrey Kemp: S (0.66) - View Novelty
9. US has military options against Iran - but also risks backlash (0.68) - View Novelty
10. Iran and the United States Today (0.66) - View Novelty
11. Iran: Expert Discusses Iran's Quds Force And U.S. Charges Concerning Iraq - RADIO FREE EUR (0.71) - View Novelty
12. Irans Spies (from Sunday Herald) (0.71) - View Novelty
13. World Crises | Reuters.co.uk (0.64) - View Novelty
14. iranian.com: Ali Ghaemi, U.S. vs. Iran (0.65) - View Novelty
15. THE AIR WAR IN THE PERSIAN GULF - Iran-Iraq war (0.64) - View Novelty
16. How to Fight Iran (0.66) - View Novelty
17. President Bush Addresses the 89th Annual National Convention of the American Legion (0.65) - View Novelty
18. Students Roil Iranian Capital in 3rd Night of Protests - New York Times (0.76) - View Novelty
19. RAND | Commentary | Iran's Covert War in Iraq (0.73) - View Novelty
20. Review: What's fueled the U.S.-Iran war of words (0.63) - View Novelty
21. frontline: terror and tehran: interviews: elaine sciolino | PBS (0.70) - View Novelty
22. The New York Times > International > Middle East > Pakistan Found to Aid Iran Nuclear Effo (0.65) - View Novelty

Figure 9: Synergist: Document Recommendations for User indicating Relevance and Novelty


Synergist: Lymba Resource Discovery: allison.thomson - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://lcc97:8080/synergist_dev/scuttle/bookmarks.php/allison.thomson
Google
Lymba Synergist Analyst Suite
Lymba Synergist Analyst Suite
Synergist: Lymba Resource Disco...
LYMBA the power to answer
synergist: lymba resource discovery
about
log in
register
allison. thomson
Sort by: Date Title URL
Search this user's bookmarks for
close
thomson
t →
e discovery
endations for user
thomson →
pup →
tags
s al-manar al-qa'ida decision
fundraising hamas history
'allah hospitals internal
ation internal structure
tional network iran iraq israel

Lebanon **[recommendations]**
06-07-2006 to hizb'allah, iran, lebanon, nas

Sobaka News Wire: Iran: Nasrallah i
06-07-2006 to hizballah, iran, leadership, r

Political Blog For The Politically Inc
God **[recommendations]**
05-07-2006 to financing, hizb'allah, iran, le

SyriaComment.com: Mussawi of Hiz
Derhally **[recommendations]**
30-06-2006 to al-mussawi, hizb'allah, iran

USCFL: Syria: The Strategy of Losin
29-06-2006 to hizb'allah, leadership, leband

Ahmadinejad Recruits Hezbollah Te

javascript:clickRDE("d" , "http://williamwebbdotorg.blogs

Recommendations for document http://williamwebbdotorg.blogspot.com/20...

Recommended users

molly.james → (0.6)
alex.ginman → (0.58)
edgar.villarreal → (0.57)
gabriel.martinez → (0.55)
allison.thomson → (0.55)
david.dillistin → (0.55)
austin.davis → (0.55)
jason.risdal → (0.51)
heidi.weber → (0.51)

Recommended tags

hezbollah → (197)
iran → (79)
united.states → (66)
israel → (55)
terrorist → (46)
syria → (39)
support → (34)
lebanon → (33)
groups → (33)

Recommended documents

http://washingtoninstitute.org/templateC06.php?CID=772 → (1)
http://www.washingtoninstitute.org/templateC06.php?CID=772 → (1)
http://counterterror.typepad.com/the_counterterrorism_blog/files/Levitt
(1)
http://dir.salon.com/story/news/feature/2005/03/11/hezbollah/index.htn
(0.92)

Figure 10: Synergist: Resource Discovery

8.2 Machine translation for multi-lingual evidence marshalling

Metre is LYMBA's Machine Translation (MT) product. It consists of a high-performance extensible Statistical MT engine and various tools to support the full translation pipeline: web document parsing, tokenizer, word segmentation, morphology analyzers, capitalizers, detokenizer, and statistical training tools. Metre is highly portable (written in Java), distributed and can use load-balancing. Metre is a very high-performance translation engine (it can translate 50-150 thousand words per minute). Metre is also extensible; extensions can be used to enhance translation for special languages (i.e.: agglutinative languages such as Korean or Japanese, synthetic languages such as Arabic), perform special handling of Out-Of-Vocabulary words (i.e.: transliteration of Arabic, Chinese or Korean words) and perform enhancements with grammatical constraints. The current set of foreign languages handled by Metre includes Arabic, French, Spanish, Italian, Portuguese, German, Chinese, Korean and Farsi.

For Korean, we encountered the following issues:

- Morphology – Korean is an agglutinative language. As a consequence, a word in English maps on many words in Korean, in different forms, such as: Subject form, Direct object form, possessive form, etc. We solved the issue by running morphological analyzers to decompose the words into morphemes and treat each morpheme as a token in translation. We experimented with different variants: keep suffixes marked, keep suffixes unmarked, drop suffixes.
- Syntax – Korean is a SOV (Subject-Object-Verb) language, in contrast with English, which is SVO. This creates problems for both the training phase (automatically discovering word alignment on the training parallel corpus) and in the translation phrase (it's harder to determine the right word order when translating from Korean to English). We solved the problem by parsing the Korean text using a Korean parser and performing clause restructuring, on both the training data and the input data during translation.
- Training collection size – The availability of parallel text for Korean is limited. We solved the problem by improving the tools used for discovery of web resources (with automatic search of words in English and their translation in Korean) and by approaching other sources, such as translation dictionaries and film subtitles.

For Farsi, we encountered the issue of limited training resources. In order to increase the training data for Metre, we developed tools that automatically extracts parallel sentences from news article collection that cover similar topics, therefore also cover a similar set of events.

9.0 RESULTS - EVALUATION, INTEGRATION, AND INSERTION

Objective: Develop processes and technologies for evaluating the technologies proposed in the other tasks. Plan for and execute integration efforts, insertions and evaluations into the analyst environment. Participate in system integration with other contractors. The following identified internal integration efforts of Lymba.

9.1 CASE Integration Experiments

Lymba participated in all three program wide integration experiments (IE) to demonstrate the capabilities developed under the project as well as integrate these capabilities with other CASE contributors:

9.1.1 Integration Experiment 1 (IE1)

Lymba integrated Synergist with BAE's Polestar system to enable search and evidence marshalling. User questions are answered by Lymba's PowerAnswer natural language question answering system. Users browse the results and select text snippets to add to their Polestar shoebox. Synergist uses web-services to post the text snippets to Polestar that adds the text to the user's shoebox and enables user to search for additional or similar evidences. Using Polestar's UI, users can select snippets and request for more evidences from Synergist's Evidence Marshalling system.

9.1.2 Integration Experiment 2 (IE2)

Lymba integrated synergist with BAE and General Dynamics. The experiment demonstrated searching and natural language question answering on social bookmarked data. The social media generated using GD's tag|Connect were used for this purpose. Users can search for answers, documents and other resources that satisfy their information need in a unified interface. Augmented pull is realized by returning important concepts from relevant documents in the concept explorer, community expertise from BAE's expertise finder system and relevant users and tags in the social network from Lymba's resource discovery engine. The integrated system generated Analyst Log Events (ALEs) and recorded them through Analyst Logging Service (ALS) hosted by Oculus.

9.1.3 Integration Experiment 3 (IE3)

IE3 focused on user modeling and information triage. Lymba extended its integration with BAE to filter and customize results based on user models generated from their tagging activities. Through web-services Lymba provides text understanding and tuning or personalization of information access services. In addition, Lymba demonstrated novelty of information in Synergist using Analyst Knowledge Models.

9.2 Resource Discovery Engine - NIST Evaluation

NIST evaluated tag and document recommendation systems from Lymba based on modeling user's knowledge and interests. The controlled experiment was performed using Scuttle social bookmarking software. Users tagged, reviewed and judged documents from NIMD Glassbox experiments for tag and document-recommendations. Unlike other social bookmarking data, the dataset generated during this evaluation covers two different types of document recommendations: document-for-user and document-for-document recommendations. Users tagged relevant documents for the domains of interest on the first day and used the same dataset to discover other related documents on day 2 for targeted tasks in their domains.

The evaluation was performed in the month of August 2008. The experiment was based on documents from NIMD Glassbox experiments. Tasks from the two domains – *Syria* and *Russian Bioweapons* – were used. Earlier NIMD experiment cycles 6, 9, and 10 involved analysts performing tasks in these topics and judging documents for their relevance to their tasks. 404 documents from the Glassbox collection were mapped to the original Glassbox tasks and split among the six subjects of the experiment. The split ensured that every subject was exposed to documents from a single domain. For Day 1, each task's documents were evenly distributed to all subjects assigned to a domain. Each Syria subject received around 60 documents. Each Russia subject received around 70 documents for the first day of the experiment.

Lymba performed a number of ***post-experiment evaluations*** with different methods for tag and document recommendations and ***demonstrated the viability of the dataset being a good resource for evaluating recommendation systems***.

The experiment demonstrated in social media, ***document content can be exploited to suggest tags for documents*** leading towards common descriptions of content through reuse of tags. Lymba's approach to combine information from user models and document content performed better than the baseline Scuttle system as well as the content based system. Modeling the user, their knowledge and interests are essential to provide good tag recommendations.

Lymba ***demonstrated two types of document recommendations*** and experimented with different types of features to model resources and evaluate their performance and utility. While feature selection provides good set of descriptors to perform both document-for-document and document-for-user recommendations, incorporating task information does provide the context to return documents relevant to user's immediate information needs.

With 5 users, we observed many differences between the relevance judgments assigned to document recommendations. While some users requested a number of document recommendations of both types in the given time period, a few only requested a couple of recommendations. Similar variations were observed in judging the returned documents for relevance to their task. Other variations observed in the data are the differences in the domains and tasks assigned. The dataset can be analyzed and evaluated in the context of these variations.

User experience with tag recommendation was very positive. However, they preferred document-for-document recommendation over document-for-user recommendation. In user-initiated information discovery experiments for a given task, users would like to satisfy their immediate needs through interactions with a search or recommendation system. The 'more like this' option provides this capability. However, Document-for-user would be most relevant in information routing or filtering settings. In our evaluations, different approaches performed well for these recommendation settings.

9.3 Insertion into the Analyst Environment – HOSTT

Hydra Open Source Terrorism Toolkit, is an integrated foraging and sensemaking framework that operates on multi-lingual and multi-media collections. This project is collaboration between Oculus, HNC/Fair Isaac, BBN and Lymba. We customized our text understanding and natural language question answering system to scale up to support large multi-lingual document collection. The distribution framework for NLP and its applications was hardened and customized for HOSTT. Lymba delivered the software with installation and documentation and is also hosting the service for demonstrations and evaluation by intelligence agencies operating on open source data.

9.4 Commercialization Activities

The research and development performed in part in the Synergist project is contributing to the different commercial activities of Lymba. Semantic technologies are strategic to the next stage of Internet and enterprise evolution. Independent market analysis (including Oracle) estimates the semantic technology marketplace will grow at the rate of 40% percent per year through the year 2010, at which point it will be a $52B industry worldwide, with about half the market in US. The estimate of public and private sector R&D semantic technology investment from 2008 to 2010 will approach $8.5B worldwide. The graphs below illustrate the projected growth in the global Semantic Technology market.

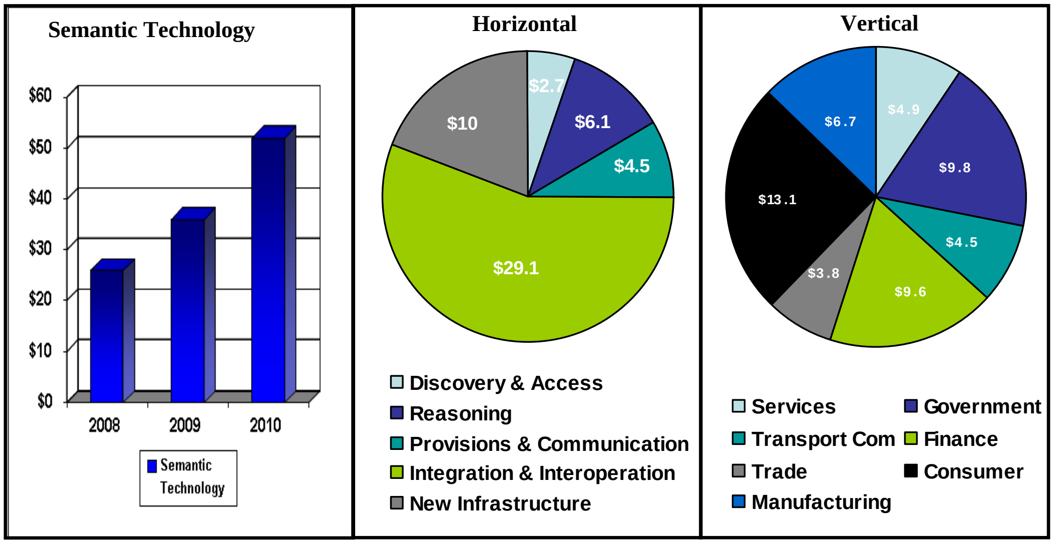


Figure 11: Semantic Technology Market to 2010 ($B) – information provided by TopQuadrant

The actual size of the market is difficult to ascertain, as the technology is becoming ubiquitous. For example, a major semantic initiative in Europe named SUPER (Semantics Utilized for Process Management within and between Enterprises) is being managed by SAP and IBM Research and is focused on integrating semantic technology with business process management, improving tools to shift control of business processes from IT professionals to business experts and to support processes of higher complexity. The difficulty becomes how to measure the revenue generated from the semantic layer of features and benefits. As SAP integrates the technology into its ERP offering, what are the revenue implications? The same holds true for many applications that include the technology: CRM, Customer Support, HR, Litigation Analysis, Market Sentiment, etc. Consequently, while there is agreement that the market for semantic technology is growing rapidly and will be very large, there is considerable debate as to the accuracy of the $52B projections. That being said, a closer look at TopQuadrant's projections provides insights to the market and helps shape Lymba's initial addressable market. The market has been divided into five categories:

- *Discovery and Access*. Sense-making, content mining, moving from documents to knowledge-centered processes, intelligent information access, and social networking will be growth areas.
- *Reasoning*. Semantic technologies enable intelligence, real-time auditing and compliance, simulation-based "virtual" product design, engineering and manufacturing, virtual data centers, adaptive logistics, and supply chain optimization. New application categories will have huge economic benefits.
- *Provision and Communication*. Representing the knowledge about things separately from content and media files will spawn new categories of enterprise publishing, especially relating to product lifecycle management, professional publishing, and business information services.
- *Integration*. By far, semantic integration will be the largest category of service and software during this decade. The estimate for this segment is $29.6B. Approximately $300B is spent each year on system integration and it is estimated that at least 50% of that is caused by semantic disparity across data sets. The projection is that by 2010, 20% of the semantic interoperability issue will be addressed through semantic technology solutions.
- *New Semantic Infrastructure*. The emergence of semantic web services, context and situational computing, semantic grid, pervasive computing, and large-ontology reasoning engines will include new operating systems and hardware categories.

Lymba's technology competence was built by assisting Intelligence Analysts to find answers to questions and to do so in a natural language format, allowing access to passages that would have gone undetected with just keyword search. Consequently, it is uniquely positioned to be successful in each of the defined categories. However, the technology is best leveraged in the first two categories, *Discovery and Access and Reasoning*.

The government insertion activities were identified in Section 9.3 above. The following identifies three distinct but related vertical markets that Lymba is focusing its commercialization efforts.

9.4.1 Customer Service

The initial target is the Global 2000 and comparable government agencies; large complex organizations with multiple requirement for Lymba's PowerAnswer product suite. Our initial research indicates that customer care for both external customers and employees has become an extremely high priority for corporate decision makers. Currently, the quality of response in call centers and help-desks hinges on the aptitude and training of the agent. Agents are supported with automated tools to display some relevant information about the caller but then have to rely on their ability to access the right information from disparate data sets to, in fact, answer a caller's question. Lymba's offering can field questions in natural language and support the agent with specific answers and other relevant knowledge. The solution extends to the entire call center market, which is comprised of 55,000 call centers, including enterprise customer service. The top 10% have the greatest need, potential returns, and appropriate resources. While the potential returns can support license fees in excess of $1M, Lymba is estimating an annual license fee of $450K. Consequently, the addressable market is approximately $2.5B. Below, we describe companies in this market with whom Lymba has existing or is negotiating a contract to customize language understanding and question-answering capabilities.

9.4.1.1 Northrop Grumman

Lymba recently won a contract and executed a project with Northrop Grumman to deliver relation extraction and ontology building in support of an initiative (Masada) to provide a system that monitors message traffic, identifies entities of interest, extracts relationships between these entities, and then contextualizes these relationships within a set of ontologies that define the domain of interest. Lymba is contracted to tailor Polaris, the semantic relation extraction engine, to support more than 200 customer defined relationships. The project consists of three phases, and in phase one Lymba anticipates delivering the first 30-40 relationships. The customization consists of writing rules to map existing relationships to Masada relationships, write and/or use semantic calculus rules to compose the Masada relations, and to build patterns for new relationships. Lymba will also deploy Jaguar to build and/or enhance ontologies from the customer data that will be used to contextualize/filter relationships that are extracted.

9.4.1.2 Xerox

Lymba’s deep semantic information extraction and automatic ontology generation capabilities are being used to extract and organize information in certain domains like manufacturing and construction. Xerox document repositories for scanned documents and documents automatically generated by Optical Character Recognition (OCR) will use these capabilities to associate metadata to documents and enable semantic search. Lymba has demonstrated the ability to use concept extraction to identify manufacturers and products, and relation extraction for identifying IS-A, Part-Whole, Make-Produce relations in text.

9.4.1.3 Empathic Software Systems

Empathic Software Systems provides understandable and efficient information management and electronic medical record-keeping (EMR) for small to medium-sized Behavioral Health Clinics and Practices. For each patient, there is a patient history. Additionally, when a patient comes in for a visit, the therapist takes notes about their physical and emotional state and symptoms. The Diagnostic and Statistical Manual of Mental Disorders (DSM) contains disorders and the symptoms associated with them. The DSM should be searched automatically based on patient history and the list of symptoms in order to return the most likely diagnosis as well as to suggest what symptoms remain for a diagnosis to be complete. Lymba has developed the DSM classification system as well as a social network for a therapist that provides the search, browsing and tag/user recommendation interfaces of synergist.

9.4.2 Content and Media Companies

News companies, training and certification companies, publishers, marketing firms, any company that owns or aggregates large volumes of unstructured or semi-structured content is currently seeking strategies to re-purpose the content to create new revenue opportunities. Their competitive advantage hinges on new ways to discover and access the content; PowerAnswer can provide a significant layer of advantage. While there are thousands of content companies, Lymba has identified an initial subset of 500 companies, many of which have subscription revenues, not just advertising revenues, to become the initial target market. The pricing model will demand more flexibility and will include licensing fees and revenue sharing contracts. However, the target annual revenue of $450K per customer remains constant and creates another $225M market opportunity. As importantly, it creates a platform for Lymba to explore a mobile solution as an extension to the on-line PowerAnswer market. Below, we present two companies to which Lymba has made preliminary product demonstrations with their data.

9.4.2.1 SkillSoft

SkillSoft enables organizations to maximize business performance through a combination of comprehensive e-learning content, online information resources, and flexible learning technologies and support services. SkillSoft's courseware is localized into 18 languages with products and services sold in 65 countries. Their e-learning environment provides students with search and browse access to the books they subscribe to.

9.4.2.2 Newspaperarchive.com

Newspaperarchive.com, the largest historical newspaper database online, contains tens of millions of newspaper pages from 1759 to present. Every newspaper in the archive is fully searchable by keyword and date, making it easy to quickly explore historical content. Focusing on genealogy, newspaperarchive.com would use Lymba's NLP tools to process census and other genealogy data to generate inputs for their genealogy application, use PACE to search and browse genealogy results and ontologies to represent topics from different eras/timeframes.

9.4.3 Partnership with Oracle

Lymba is in the process of entering into a strategic partnership with Oracle with the goal of providing Oracle customers with a tool that speeds up the creation of databases and knowledge bases directly from textual documents.

Relational databases provide methods for storing large amounts of data and query them efficiently. While direct querying of databases is not provided to end users, applications anticipate and encode all semantics that a user would expect to get from the data. Oracle and other RDBMS companies are moving towards capturing the semantics in separate databases (RDF databases) and enabling the interpretation of the data using ontologies. For example, semantic matching with ontologies enables users and applications to search a database of clinical trials for instances of cancer patients without explicitly enumerating all forms of cancer in their query. A newly discovered variation of a particular cancer can be included at the appropriate place in the ontology and applications developed using this ontology will automatically incorporate this new information in subsequent analysis.

10.0 REFERENCES

Lymba has developed a number of prototypes and added these capabilities to the integrated Synergist Analyst Suite environment. Lymba's technical contributions include publication and presentations listed below.

[1] M. Balakrishna and M. Srikanth, *Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF)*, Conference on Ontology for the Intelligence Community (OIC), 2008.

[2] M. Tatu, M. Srikanth and T. D'Silva, *RSDC08: Tag Recommendations using Bookmark Content*, In the Proceedings of Workshop on Resource Discovery Challenge, ECML/PKDD, September 2008. Best performing submission to the tag recommendation task of Discovery Challenge, ECML/PKDD 2008 (http://www.kde.cs.uni-kassel.de/ws/rsdc08/).

[3] Marta Tatu and Munirathnam Srikanth. 2008. *Experiments with Reasoning for Temporal Relations between Events*. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 857–864, Manchester, UK, August. Coling 2008 Organizing Committee.

[4] Marta Tatu and Dan I. Moldovan. 2007. *COGEX at RTE 3*. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 22–27, Prague, June. Association for Computational Linguistics.

[5] D. Moldovan, M. Srikanth and A. Badulescu, *Synergist: Topic and User Knowledge bases from Textual Sources for Collaborative Intelligence Analysis*, CASE PI Conference, Sep 2007.

[6] C. Min, M. Srikanth, and A. Fowler. *A Hybrid Approach to Temporal Relation Identification*. In the Proceedings of SemEval-2007 - 4th International Worshop on Semantic Evaluations.

[7] A. Badulescu and M. Srikanth. *LCC-SRN: LCC's SRN System for SemEval 2007 Task 4*. In the Proceedings of SemEval-2007 - 4th International Worshop on Semantic Evaluations.

[8] A. Novischi, M. Srikanth and A. Bennett. *LCC-WSD: System Description for English Coarse Grained All Words Task*, In the Proceedings of SemEval-2007 - 4th International Worshop on Semantic Evaluations.

11.0 ABBREVIATIONS/ACRONYMS

ACE	Automatic Content Extraction (evaluation)
ACH	Analysis of Competing Hypotheses
ALE	Analyst Log Event
ALS	Analyst Logging Service
BAE	British Aerospace Engineering
CASE	Collaboration and Analyst/System Effectiveness
COGEX	Lymba’s logic Prover
CRM	Customer Relations Management
ECML	European Conference on machine learning
ERP	Enterprise Resource Planning
EventNet	Linguistic resource for modeling events their properties and relations that can help in situation understanding and prediction problems
GD	General Dynamic
HOSTT	Hydra Open Source Terrorism Toolkit
HR	Human Relations
HTML	Hypertext Markup Language
IARPA	Intelligence Advanced Research Projects Activity
IBM	International Business Machine
IC	Intelligence Community
ICE	Internet Communications Engine
ID	Identification
IE	Integration Experiment
IT	Information Technology
Jager	web-browser based application that provides scalable, multi-user, collaborative editing of Jaguar ontologies stored in an RDBMS like mySQL
Jaguar	Tool to automatically build domain-specific ontologies from text
KB	Knowledge Base
METRE	Machine translation system from Lymba
MS	MicroSoft
MT	Machine Translation
MUC	Message Understanding Conference
NIMD	Novel Intelligence from Massive Data
NIPF	National Intelligence Priorities Framework
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OWL	Web Ontology Language
PACE	Power Answer Concept Exploration
PDF	Portable Document Format
PKDD	Principles and practice of knowledge discovery in databases
PTKB	Prior and Tacit Knowledge Base
RAP	Resolution of Anaphora (algorithm)
RDBMS	Relational Database Management System

RDE	Resource Discovery Engine
RDF	Resource Description Framework
RSDC	Relational Software Development Conference
SAP	Special Access Program
SemEval	Semantic Evaluation (International Workshop)
SUPER	Semantics Utilized for Process Management within and between Enterprises
SVM	Support Vector Machines
TempEval	Temporal Evaluation- links temporal expressions to events and tasks
TF-IDF	Term frequency and inverse document frequency values
TimeBank	Corpus assign explicit time to events
TREC	Text Retrieval Conference
UI	User Interface
VerbOcean	Publicly available verbal resource for linguistic research
XWN-KB	Extended WordNet Knowledge Base
XWN	Extended WordNet