

FOR Govt Use Only  
[Proposal Number]

\_\_\_\_\_  
\_\_\_\_\_

## VOLUME 1 - Technical / Management Details

<b>Organization / Company</b>	<b>The Trustees of Columbia University in the City of New York</b>
<b>CAGE Code</b>	<b>1B053</b>
<b>DUNS / CEC Number</b>	<b>049179401</b>
<b>TIN Number</b>	<b>13-5598093</b>
<b>Type of Business</b>	<b>Institute of Higher Education</b>
<b>Proposal Title</b>	<b>Fusing Rich Information Extracted from Multiple Media and Languages to Generate Contextualized Complex Answers</b>
<b>System Design Perspective Category (Check ONLY ONE Box) (See Section 5.1)</b>	<input checked="" type="checkbox"/> 1. End-to-End System <input type="checkbox"/> 2. Component Elements <input type="checkbox"/> 3. Cross Cutting / Enabling Technologies
<b>If "Component Elements" Category Selected Above (Check ALL that Apply) (See Section 5.1.2)</b>	<input type="checkbox"/> Question Understanding and Interpretation <input type="checkbox"/> Determining the Answer <input type="checkbox"/> Formulating and Presenting the Answer <input type="checkbox"/> Other (Identify):
<b>Data Strategy (See Section 5.2.3)</b>	<input type="checkbox"/> Focused Data Strategy <input checked="" type="checkbox"/> Diverse Data Strategy <input type="checkbox"/> Other (Identify):

BAA 03-06-FH - VOLUME 1 - Technical / Management Details (CONTINUED)

<b>Team Members / Type of Business</b>	<b>Columbia University University of Colorado</b>
<b>Principal Investigator(s) Name(s)</b>	<b>Vasileios Hatzivassiloglou</b>
<b>Mail Address</b>	<b>Department of Computer Science 1214 Amsterdam Avenue, 450 CS Building New York, NY 10027</b>
<b>Phone Number</b>	<b>212-939-7110</b>
<b>Fax Number</b>	<b>212-666-0140</b>
<b>E-mail Address</b>	<b><a href="mailto:vh@cs.columbia.edu">vh@cs.columbia.edu</a></b>
<b>Administrative Contact Name</b>	<b>Patricia Welch</b>
<b>Mail Address</b>	<b>Office of Projects and Grants 1210 Amsterdam Avenue 254 Engineering Terrace, Mail Code 2205 New York, NY 10027</b>
<b>Phone Number</b>	<b>212-854-6851</b>
<b>Fax Number</b>	<b>212-854-2738</b>
<b>E-mail Address</b>	<b>phw1@columbia.edu</b>

<b>Proposal Duration</b>	<b>2 years</b>
<b>Cost - Year 1</b>	<b>\$ 989,882</b>
<b>Cost - Year 2</b>	<b>\$ 989,455</b>
<b>Total Cost</b>	<b>\$ 1,979,337</b>

# **Fusing Rich Information Extracted from Multiple Media and Languages to Generate Contextualized, Complex Answers**

AQUAINT Phase II Proposal

Columbia University (Primary Contractor)  
University of Colorado at Boulder (Subcontractor)

Technical Point of Contact: Vasileios Hatzivassiloglou ([vh@cs.columbia.edu](mailto:vh@cs.columbia.edu))

## **Volume I: Technical and Management Details**

### **PART I: Summary of Proposal**

#### **A. Innovative Claims**

We propose to continue our development of an integrated system for answering complex questions: questions that require interacting with the user to refine and clarify the context of the question. Our focus is on the generation of long, complex answers to open-ended questions, where answers require semantic analysis, summarization, and fusion of information from multiple sources in different modalities (speech, text, databases, web pages) and multiple languages (English, Chinese, Arabic).

The unique contribution of our research includes the following innovations:

- Provision of an environment for cooperative interaction with the user, allowing clarification of hard questions, and understanding context and background knowledge.
- Development of a framework for fusion of multiple pieces of information from diverse sources (TREC/AQUAINT/CNS, Google, databases, speech, Chinese, Arabic), resulting in a coherent response that highlights complex relations such as comparison, perspectives, contradiction, and opinions.
- Combination of deep linguistic knowledge-sources and annotations of semantic role, event structure, word meaning, opinions, and question-type specific facts with sophisticated state-of-the art statistical and machine learning methods.

Our unique integration of technologies that include information fusion, language generation, speech recognition, dialogue modeling, and Arabic and Chinese processing, together with sophisticated statistical machine learning algorithms applied to rich linguistic knowledge about events, opinions, contradictions and semantic structure, will allow us to build an end-to-end system to extend significantly the range of information gathering resources available to analysts.

## B. Brief Summary of the Technical Rationale

We propose to continue research and development of an end-to-end system that can respond to hard questions with complex, long answers. In Phase I, we developed components for generating responses to opinion, definition, biography, and event questions. In Phase II we will extend each of these components addressing issues such as change in opinion over time, identification of networks of events, definitions in the context of historical or cause-effect information, and generation of biographical descriptions from different viewpoints. We will also address new question types, such as questions about contradictions or comparison of perspectives. All extensions and new question types will be integrated in our end-to-end QA system.

In order to achieve this goal, we will carry out research in four main areas, outlined below. A key characteristic of all aspects of our research is the integration of principled, linguistic representations and approaches with sophisticated, robust statistical techniques.

**Extraction of Rich Semantics from Documents.** In order to answer factoid questions, it is often sufficient to rely on simple term-indexing and named-entity processing of documents to find answer snippets. But answering hard questions requires more sophisticated semantic analysis of documents. We will apply three kinds of sophisticated semantic analysis to annotate documents with rich information, and extract important pieces of information. These include robust semantic role parsing which assigns a shallow meaning structure to a sentence, extraction of a network of events composed of smaller atomic events, and extraction of opinions, differentiating between subjective and factual, positive and negative and holder of the opinion and opinion held.

**Unified, multi-strategy response generation for complex, long answers.** In Phase I, we experimented with different strategies for different question types. In Phase II, we will merge response strategies and apply the combined set of strategies across question types, allowing us to leverage the use of top-down plans, bottom-up summarization and clustering, and integration of semantic facts drawn from data with textual fragments pulled from text. This central answer generation module will resolve redundancies, merge compatible facts, highlight contradictions and organize the paragraph-length answer to maximize cohesion.

**Integrating diverse data modalities.** A sophisticated question answerer must be able to combine information from structured and unstructured text, web pages, or audio documents (speech recordings), as well as text in foreign languages. We will develop the capability to query each of these diverse data modalities. We will combine information from multiple text sources such as the TREC, AQUAINT and CNS collections; Google; structured and semi-structured data sources such as databases, tables, and web pages; spoken-language documents via speech recognition, and we will provide support for text documents in Chinese and Arabic. All possible sources will be queried to return answer candidates and returned information from all of the sources will be integrated into a single response.

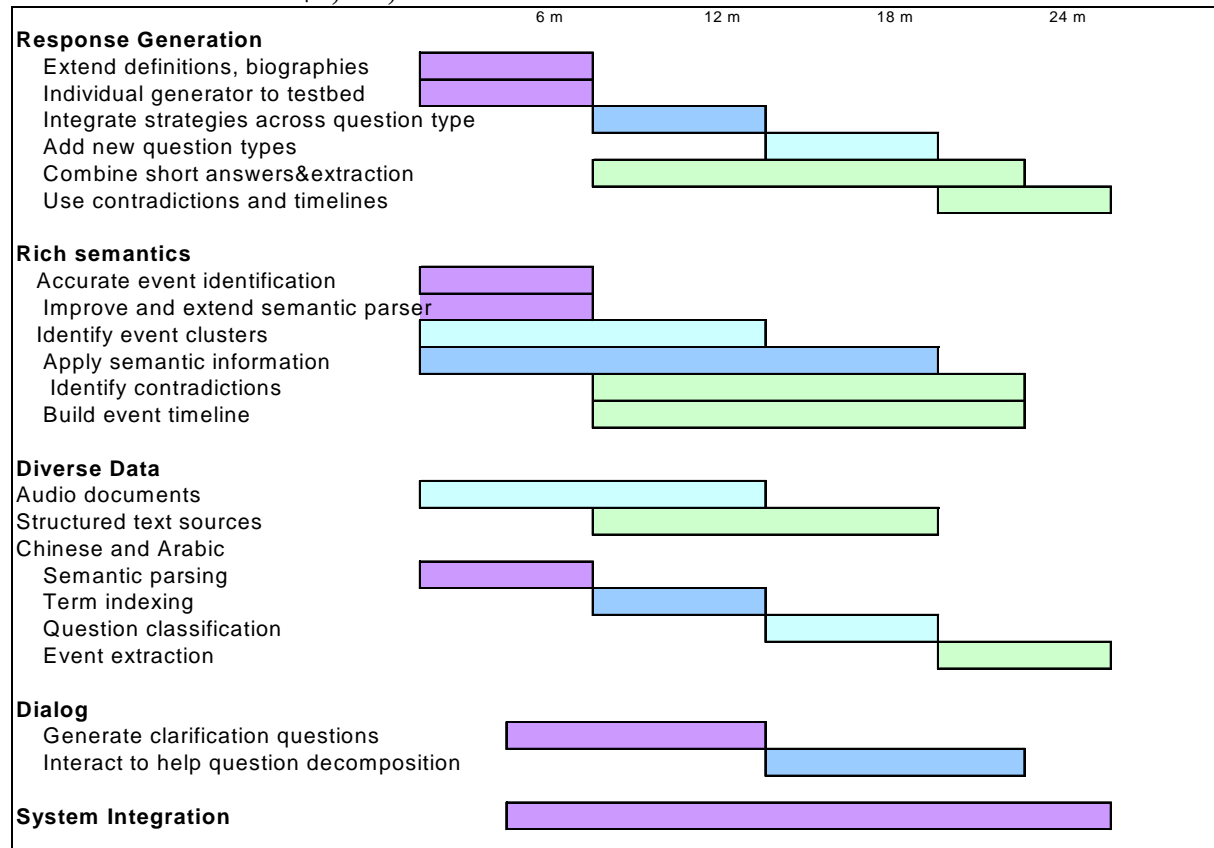
**Going beyond the single question-answer pair.** Our focus to date has been primarily on answering single questions and there has been little interaction between successive questions. As part of Phase I, we implemented a first prototype dialog system which maintains key words and focus across a sequence of questions. In Phase II a large part of our effort will be in developing and using context information across a sequence of queries and answers. Analysis of new questions will be influenced by the prior questions in the set, enabling treatment of the entire set as a scenario providing local context. We will also extend our system to generate clarification questions. Our system will implement new techniques for decomposing complex questions into a sequence of simpler ones and tracking information across successive questions the user asks. We will rely on dialog interaction with the user to help guide question decomposition.

### C. Schedule and Milestones

The objective of the proposed work is to extend and expand the work we did under Phase I to improve the technology for answering questions that require complex answers. We will also extend our coverage of sources and languages. Our specific tasks under Phase II are:

1. **Extraction of rich semantics**     **\$689,474**
  - a. Semantic role parsing
  - b. Identifying and annotating opinions
  - c. Event extraction
  - d. Contradictions, perspectives and timelines
2. **Generating complex answers**     **\$372,746**
  - a. Definitions
  - b. Biographies
  - c. Unified answer fusion
  - d. Evaluation of long answers
3. **Diverse data modalities**     **\$415,194**
  - a. Multilingual support for Chinese and Arabic
  - b. Spoken language (audio) documents
  - c. Semi-structured data
4. **Integrated system**     **\$501,923**
  - a. Interactive dialog over multiple questions
  - b. Integration of independent modules
  - c. Overall project coordination

**TOTAL**                     **\$1,979,337**



## D. Summary of the Deliverables

The primary deliverable of this effort is an end-to-end multimodal question-answering system designed specifically for questions that necessitate complex answers. More specifically, the system engages in either spoken or written dialog with a user, classifies the user’s request by type, consults a variety of spoken or written sources to identify information relevant to the user’s question, and generates a coherent paragraph length answer to the question. The following items represent the core components of this system:

- **Biography, definition, event, and opinion question processing systems** – these systems classify questions falling into these categories, analyze source materials with respect to these categories, account for the current context and generate responses appropriate to each type.
- **Event detection software** – this component will detect events in source materials and generate semantic representations of these events suitable for use in question-answering. It will facilitate the processing of questions involving related events as part of a single timeline, or similar events involving the same or related participants.
- **Unified answer generation** – this module will resolve redundancies, merge compatible facts, highlight apparent contradictions and organize the paragraph-length answer to maximize cohesion. This will allow us to leverage generation strategies across all question types, even those not currently covered by our system.
- **Semantic parsers** – along with a fully functional semantic parser for English, we will develop and deliver prototype semantic parsers for Arabic and Chinese. These parsers will provide a level of semantic markup that is directly useful for answering simple fact-based questions, as well as for the creation of our deeper event-oriented representations.
- **Interactive dialog management system** – A system handling both spoken and written inputs, tracking the dialog focus, generating clarification and follow-up questions.

## E. Organizational Chart of all Anticipated Program Participants

Participant	Organization	Role	Year 1	Year 2
Vasileios Hatzivassiloglou	Columbia University	Principal Investigator	40%	40%
Kathleen McKeown	Columbia University	Key Personnel / co-PI	15%	15%
Rebecca Passonneau	Columbia University	Significant Contributor	20%	20%
Hong Yu	Columbia University	Contributor	50%	50%
Sasha Blair-Goldensohn	Columbia University	Contributor	63%	63%
Elena Filatova	Columbia University	Contributor	63%	63%
Andrew Schlaikjer	Columbia University	Contributor	100%	100%
Daniel Jurafsky	University of Colorado	Key Personnel / co-PI	15%	15%
James Martin	University of Colorado	Key Personnel / co-PI	27%	27%
Wayne Ward	University of Colorado	Key Personnel / co-PI	25%	25%
Kadri Hacioglu	University of Colorado	Contributor	42%	42%
Mona Diab	University of Colorado	Contributor	54%	54%
Graduate Assistant 1	University of Colorado	Contributor	63%	63%
Graduate Assistant 2	University of Colorado	Contributor	63%	63%

**Vasileios Hatzivassiloglou** is an Associate Research Scientist in the Center for Computational Learning Systems and the Department of Computer Science at Columbia University. He received a Ph.D. in Computer Science from Columbia University in 1998. His research interests include statistical natural language processing, machine learning, lexical semantics, summarization, intelligent information retrieval, and bioinformatics. He has authored more than 50 research papers in international journals and conferences, and has served as program committee member for all major conferences in the field of natural language processing, including ACL, NAACL, HLT, COLING, EMNLP, and the joint ACM/IEEE conference on Digital Libraries. He has given numerous invited talks, including at the University of Pennsylvania, IBM, John Hopkins University, University of Tokyo, Rockefeller University, and the National Institute of Statistical Sciences, and has served on advisory task forces convened by the NSF, the CIA, and the European Union.

**Kathleen R. McKeown** is Professor and Chair of Computer Science at Columbia University. She received a Ph.D. in Computer Science from the University of Pennsylvania in 1982. Her research interests include text summarization, natural language generation, multi-media explanation, digital libraries, concept-to-speech generation and natural language interfaces. As a result of her work in generation and multimedia explanation, Prof. McKeown received an NSF Presidential Young Investigator Award in 1985, a NSF Faculty Award for Women in 1990, was elected a AAI Fellow in 1995, and was selected as Outstanding Woman Scientist by the New York Association of Women in Science in 2000.

**Rebecca J. Passonneau** has been consulting in computational linguistics in the New York/New Jersey area for the past three years with institutions such as AT&T Research Labs, ETS, ETS Technologies, the Computer Science Department at Columbia University, and the Columbia University Libraries. She received a Ph.D. in Linguistics from the University of Chicago in 1985, has worked in a range of industry and academic research settings, and has published widely in the areas of discourse, reference resolution, and semantic and pragmatic interpretation of temporal expressions.

**Daniel Jurafsky** is an associate professor of linguistics, computer science and cognitive science at the University of Colorado, Boulder. He received the Ph.D. in Computer Science from the University of California at Berkeley in 1992, and came to Colorado in 1996 after spending 4 years at the International Computer Science Institute. His research focuses on statistical models of human and machine language processing, especially automatic speech recognition and understanding, computational psycholinguistics and natural language processing (including semantic parsing and question answering). He received a National Science Foundation CAREER award in 1998 and the MacArthur Fellowship in 2002. He has served on various editorial boards (Computational Linguistics, Computer Speech and Language), academic executive boards (SIGPHON, SIGNLL, NAACL) and corporate technical advisory boards (Ask Jeeves, Proofpoint). His most recent book, with James H. Martin, is the widely used textbook “Speech and Language Processing”.

**James H. Martin** received a B.S. in Computer Science from Columbia University and a Ph.D. in Computer Science from the University of California, Berkeley. He then joined the faculty of the University of Colorado, Boulder, where he is now an associate professor in the Department of Computer Science, and Director of Academic Programs in the Institute of Cognitive Science. His

research interests include computational semantics, metaphor, machine learning, and information retrieval. He has published numerous papers and two books including the recent text, *Speech and Language Processing*, with Daniel Jurafsky.

**Wayne Ward's** background is in computational models of perception and understanding. His master's thesis presented new findings concerning the *precedence effect* (the suppression of echo by the auditory system). His doctoral thesis was a computational model of the signal processing that occurs in human binaural localization. From 1986-1998, working as faculty in the School of Computer Science at Carnegie-Mellon University, he conducted research projects in the area of spoken dialogs with machines to accomplish real world tasks. From 1996-1998, Dr. Ward led the technical development team for the SmartVoice system, a project to use speech understanding technology to extract information from medical dictations. In 1998 Dr. Ward co-founded the Center for Spoken Language Understanding, along with Ron Cole, John Hansen, Daniel Jurafsky and James Martin. This center conducts research in a wide variety of topics in human language technology.

## **PART II: Detailed Proposal Information**

### **A. Innovative Claims**

We propose to continue our development of an integrated system for answering complex questions, and questions that require interacting with the user to refine and clarify the context of the question. For example, a question such as "How did Ahmed Qurei become the new Prime Minister?" might apparently have a short, specific answer that requires identification, comparison, extraction and validation of information across sources, leading to the following answer: "Arafat appointed him to the post on September 8, 2003." A longer, more relevant answer requires even more complex integration of one or more information-seeking tasks; for example: a biography highlighting his previous governmental posts; an event timeline indicating the events that led to the resignation of the previous prime minister, Abbas; a summary of opinions for and against the new appointment. We focus on questions that have long, complex answers, necessitating semantic analysis, summarization, and fusion of information from multiple sources in different modalities (speech, text, databases, web pages) and multiple languages (English, Chinese, Arabic).

The unique contributions of our research include the following innovations, which are directly responsive to the AQUAINT Phase II goals:

**1. Question Answering as Part of a Larger Information-Gathering Process:** We will answer complex questions that require cooperative interaction with the user, de-composition into component questions, and understanding context and background knowledge. Specific innovations include the use of interactive, spoken dialog to allow the user to become an active participant in finding information through clarification dialogues; speakers can ask follow-up questions, pursue a line of thought, and clarify their intentions when the system misinterpreted a question. Our linked event representation will allow us to detect potentially related information and propose alternative directions to the user.

**2. Evaluating, Validating and Presenting an Answer:** We will answer questions whose long, complex answers require integrating multiple pieces of information from diverse sources. Specific innovations include: new evaluation methodology, significant expansions of our earlier

work on long-answer questions such as opinion, definition, biography, and event questions; new types of long-answer questions, including contradictions and timelines, as well as new unified methods to extract key facts and merge them to produce coherent answers.

**3. Combining Knowledge-based, Statistical and Linguistic Approaches to QA:** We will answer questions by integrating deep linguistic knowledge-sources and annotations with sophisticated state-of-the art statistical and machine learning methods. Specific innovations include new tools for extraction of rich information from documents, including significantly extended robust semantic role parsing which assigns a shallow meaning structure to a sentence, extraction of a network of events composed of smaller atomic events, and extraction of opinions differentiating between subjective and factual, positive and negative and holder of the opinion and opinion held, all based on combining large linguistically-annotated databases with sophisticated statistical classifiers.

**4. Accessing, Retrieving and Integrating Diverse Data Sources:** We will answer questions by extracting and organizing knowledge from a wide variety of diverse data sources. Specific innovations include the combination of information from multiple text sources such as the TREC, AQUAINT and CNS collections; Google; structured and semi-structured data sources such as databases, tables, and web pages; spoken-language documents via speech recognition; partial support of text documents in Chinese and Arabic. All possible sources will be queried to return answer candidates, and the resulting information will be fused into a single response. Our approach implements the *Diverse-Data* strategy, as defined in section 5.2.3.2 of the BAA.

In summary, our unique integration of technologies such as information fusion, language generation, speech recognition, dialogue modeling, and Arabic and Chinese processing, together with sophisticated statistical machine learning algorithms applied to rich linguistic knowledge about events, opinions, contradictions, semantic structure, and question-types, will allow us to build a system which significantly extends the range of possible question types and responses available to analysts, and seamlessly fuses these to generate a response.

## B. Technical Rationale

### *B.1. Introduction*

During the past two years, research in question answering has seen significant growth, leading to new approaches ranging from the purely statistical to hand-built knowledge models and to large improvements in the performance of question answering systems. However, most current efforts in question answering, even within the AQUAINT program, are still directed towards short-answer questions requiring a single fact, or a list of facts, in response. This is a justified strategy since answering such questions remains non-trivial, and many approaches to more complex questions rely in part on the answering of simpler questions (e.g., by question decomposition). Many questions, however, are open-ended, requiring fusion of multiple facts, descriptions or opinions to create a response. Our team at Columbia University and the University of Colorado at Boulder<sup>1</sup> has, from the beginning of AQUAINT Phase I, focused on questions requiring more complex answers. We developed techniques for responding to several kinds of open-ended

---

<sup>1</sup> Consisting of Dr. Vasileios Hatzivassiloglou, Dr. Kathleen McKeown, and Dr. Rebecca Passonneau at Columbia, and Dr. Daniel Jurafsky, Dr. Wayne Ward, and Dr. James Martin at Colorado.

questions which are best answered by integrating information from multiple sources to form a coherent, multi-sentence response. These include biographical questions about people, definitional questions about entities, concepts, and organizations, questions with multiple and uncertain answers (including opinions), and questions where the answer varies according to time and source (questions about events are a special such case where information is inherently linked to time). In addition, we explored shallow, statistical semantic parsing as a means of obtaining a more detailed analysis of questions (and candidate answers); we introduced new models in our semantic parser raising its accuracy to 87% and are currently integrating the parser's semantic labels in our broader question analysis techniques.

As an example, consider a scenario revolving around the question such as "How did Ahmed Qurei become the new Prime Minister?" A relevant answer might integrate one or more information-seeking tasks such as an event timeline indicating the events that led to the resignation of the previous prime minister, Abbas; biographies of Abbas and Qurei; a summary of opinions for and against the new appointment. Figure 1 shows examples of output we aim to generate (biography) as well as current output of system components (opinions, definition).

*Target output for **What do you know about Mahmoud Abbas?***

Mahmoud Abbas, also known as Abu Mazen, was the Palestinian Prime Minister. According to the World Socialist Web Site, he is a businessman and an advisor to the reactionary rulers of Qatar. According to the U.S. administration, Abu Mazen was described as the only credible Palestinian peacemaker. The protests during a PLC meeting described Abbas as a traitor. He has been said to be a longtime Arafat comrade and a fellow founder of the Palestine Liberation Organization faction Fatah. For the World Socialist Web Site, Mahmoud Abbas had little popular support. In 1993, he has a key role in the secret talks that led to the failed Oslo Accord signed.

Abu Mazen offered hope for a U.S.-backed step-by-step peace plan (see the Mideast road map). He experienced a power struggle with Arafat (according to CNN) and he refused to try to dismantle Hamas and other violent Palestinian groups (according to Yahoo! News). On September 6th, 2003, he submitted his resignation.

*CUAQ output for **What do people think of Abbas' resignation?***

Foreign leaders said Palestinian Prime Minister Mahmoud Abbas' resignation Saturday deals a serious blow to Middle East peace efforts, and blamed both Palestinian infighting and Israeli attacks for the latest crisis.

British Foreign Secretary Jack Straw said the resignation did not "put the peace process back to square one, but it is a further difficulty, a huge tragedy that the Palestinians should be so divided.

*CUAQ output for **What are the Oslo Accords?***

The Oslo accords are the foundation on which peace negotiations between Israel and the Palestinians are based. The accords laid out the long-term goals to be achieved, including the complete withdrawal of Israeli troops from the Gaza Strip and the West Bank, and the Palestinians' right to self-rule in those territories. Qurei, who is strongly associated with the Oslo Accords, has reemerged at a time when many Israelis and Palestinians are talking about the total failure of the agreement. Officially called the "Declaration of Principles," the accords were negotiated secretly by Israeli and Palestinian delegations in 1993 in Oslo, Norway, guided by Norwegian Foreign Minister Johan Jorgen Holst.

**Figure 1: Question Answering Scenario with real and target output**

In Phase I, we started as newcomers to the question answering area, having done no prior work directly on analyzing questions or producing answers. In the two years since, we developed

new methods for assigning the appropriate answer type to incoming questions, obtaining an automatic semantic analysis of questions and sentences within relevant documents, identifying opinions, annotating events, and generating descriptive definitions and biographies. We implemented and evaluated these components, and developed a fully functional end-to-end question answering system, capable of interpreting input questions and generating responses for definitions, biographies, and opinions. We also developed components for multimodal input of questions (text and speech) and an initial dialog interface for handling a succession of questions.

Building on these achievements, we intend to continue exploring the research issues we identified in these two years, extending our ability to produce responses to definitional, biographical, opinion and event questions and addressing issues such as change in opinion over time, identification of networks of events, definitions in the context of historical or cause-effect information, and generation of biographical descriptions from different viewpoints. We will extend all components and integrate those that we have not yet incorporated into the full system, including event-based questions, spoken input and the dialog interface. We will also address new issues raised by our Phase II Goals, including generation of responses that highlight contradictions and changes over time, extension to diverse modalities such as speech, foreign language sources and structured data, and further consideration of context. To achieve these goals, we will carry out research in the following four areas:

- **Extracting rich semantic information** for use in the question answering process. We will apply three kinds of sophisticated semantic analysis to annotate documents with rich information, and extract important pieces of information. These include robust semantic role parsing, identification and extraction of events, and extraction of opinions. We are proposing further enhancements in parsing to obtain both more accurate and richer semantic parses. We also will also develop a new framework, analogous to that proposed for answer generation, where the labeled output stream from the semantic parser is used to produce additional features for direct use in response generation (e.g., roles such as OPINION-HOLDER and OPINION from opinion text). We will extend our work on event extraction to identify clusters of related events, building a network of events and analyzing how events change over time. Similarly, for opinion extraction, we will focus on association of opinions with opinion-holders, analyzing how a person's opinion changes over time. Our work in these three areas will be learned from human annotated training data using methods that leverage statistical, linguistic and knowledge-based methods.
- **Unified, multi-strategy response generation for complex, long answers.** Our focus has been and continues to be on long, complex answers where multiple pieces of information from diverse sources are relevant to the answer. We will complement our current set with questions targeting sequences of related events, either across a timeline or sharing common participants, with complex questions that are conditional on the results of previous questions, and with questions requiring responses that highlight different perspectives, contradictions and changes over time. In Phase I, we experimented with different strategies for different question types. In Phase II, we will merge response strategies and apply a unified multi-strategy response generation approach across question types, allowing us to leverage the use of top-down plans, bottom-up summarization and clustering, and integration of semantic facts drawn from data with textual fragments pulled from text. This central answer generation module will resolve redundancies, merge compatible facts, highlight contradictions and

organize the paragraph-length answer to maximize cohesion. We are developing new evaluation metrics to measure the quality of generated responses.

- **Going beyond the single question-answer pair.** To date, we have been focusing primarily on answering single questions (albeit with multiple and complex answers), and there has been little interaction between successive questions. As part of Phase I, we have implemented a first prototype dialog system which maintains key words and focus across a sequence of questions. In Phase II a large part of our effort will be in developing and using context information across a sequence of queries and answers. Analysis of new questions will be influenced by the prior questions in the set, enabling treatment of the entire set as a scenario providing local context. We will also extend our system to generate clarification questions. Our system will implement new techniques for decomposing complex questions into a sequence of simpler ones and tracking information across successive questions the user asks. We will rely on dialog interaction with the user to help guide question decomposition.
- **Integrating information from multiple sources, multiple modalities, and multiple languages.** Our approach relies on extracting and organizing disparate pieces of information from multiple sources; we have already successfully integrated text sources such as the TREC and AQUAINT collections, the Center for Non-Proliferation Studies (CNS) collection, and current search results returned by Google. We will increase our coverage of different sources by incorporating additional collections with specific characteristics (such as country and time period) and automatically stream search results into virtual “sources” on the fly for the purpose of organizing the answers according to perspective. In terms of modalities, we currently handle spoken as well as typed questions; we will expand our coverage of different modalities significantly by processing spoken data as well as spoken questions and exploring structured and semi-structured sources of information such as databases, tables, and web pages. Finally, we will add to our system support for processing documents in two foreign languages, Chinese and Arabic. This processing will initially involve semantic parsing, question classification, term indexing and event extraction, with the intent of extending our system to full coverage of foreign text in these languages during AQUAINT Phase III.

These specific aims address all four challenge areas identified in the BAA for AQUAINT Phase II. Our dialog, question clarification, and question decomposition approaches directly support the goal of *question-answering as part of a larger information gathering process*, while the new question types and expanded treatment of current question types will allow more complex questions. This first challenge is also addressed by our integration of information from multiple sources and new focus on contradictions and changes over time. Not only are we examining different sources of the same type but also of divergent modalities (text, tables, semi-structured web pages, recorded and live speech) and in multiple languages (English, Chinese, Arabic), satisfying the challenge of *accessing, retrieving and integrating multiple, diverse data sources*. Our fusion of information from multiple sources and our proposed unified treatment of answer generation across different question types support both of the above challenges as well as the goal of *evaluating, validating and presenting an answer*. This latter goal is also explicitly supported by our proposed investigation of new evaluation methods for long answers. Finally, a major focus of our work is on the semantic analysis of questions and answers, an approach that integrates statistical learning with linguistic representations. Not only through our semantic parser, but also through the use of learning to acquire knowledge in all our components, we

address the goal of *exploring boundaries / combinations of knowledge-based, statistical and linguistic approaches to question answering*.

As in our earlier Phase I work, we will integrate the proposed components into a **full system** handling question classification, question analysis, information extraction according to different models for each question type, unified answer synthesis, and modeling context via our dialog subsystem. However, individual components will also be available separately with their own published API for use by other researchers within the AQUAINT program, and we will be active participants in the integration efforts within the AQUAINT testbed system. In addition, we have identified a number of collaborations with other AQUAINT Phase II offerors (detailed in Section B.8) and outline mutual plans for sharing our and their technologies and data across sites.

In section B.2, we provide an overview of our achievements in Phase I. The following sections then move to the four main areas of research that we will explore. Section B.3 discusses extraction of rich semantics from documents, Section B.4 discusses generating complex answers, section B.5 discusses diverse data modalities and section B.6 discuss how we will extend our model beyond single question-answer pairs.

## ***B.2. Architecture and Summary of Achievements in AQUAINT Phase I***

**Scientific Contributions.** The primary new ideas developed in Phase I were:

- A multi-strategy approach to response generation. During Phase I we explored techniques for four question types (definitions, biographies, opinions, and events) and developed different response generation strategies for each.
- Domain independent semantic annotation for question answering. During Phase I we developed a parser for automatic semantic annotation of input and began integration of the semantic information into the question classification and response generation processes.

**System Architecture.** We developed an end-to-end question answering system that incorporates many of our modules for advanced question answering. The first step in our integrated question answering process is to classify the type of the question (fact, definition, opinion, biography, event), which determines the response module to generate a response. For question classification, we use a pattern based classifier that is based on words, parts-of-speech and named entities. Focus and keywords are extracted from the question and are passed to our information retrieval component (we use Lucene (Lucene, 2000) for local collections such as the AQUAINT and CNS collections and Google for searches over the web). The retrieved documents are segmented into paragraphs and scored according to estimated relevance to the answer. After pruning the lowest scoring documents, the remaining ones are annotated with semantic and named entity information and are routed to one of the response generators dealing with short answers, biographies, definitions, or opinions. The response generator processes the relevant document segments, producing an appropriate response. We have also developed a spoken question interface and a dialog context manager for the system.

We have adopted a modular design based on client-server and server-server interactions, allowing modules to operate in parallel and on different machines. Two main servers coordinate question analysis and answer generation. The communication between servers and clients is handled via HTTP, allowing distributed operation of the system with some modules running at Columbia and others at Colorado. We have emphasized standardized services with well-specified APIs, facilitating the addition of new modules (e.g., for new question types) without other

modifications to the core system. We have also built a web-based client for entering questions and displaying results remotely on any standard browser.

**Evaluation.** We evaluated the performance of different components in part by local evaluations, which in some cases used data newly annotated for the specific purpose of our evaluation, and in part by participating in larger efforts such as TREC in 2002 and 2003 and the AQUAINT pilot evaluations on definitions, dialog, and opinions. We also investigated new techniques for measuring the success of question answering approaches in synthesizing long answers.

### ***B.3. Extraction of Rich Semantics from Documents***

In order to answer factoid questions, it is often sufficient to rely on simple term-indexing and named-entity processing of documents to find answer snippets. But answering hard questions, and producing long, complex answers to them, requires more sophisticated semantic analysis of documents. The first of the four major components of our proposal addresses this problem. We propose to apply four kinds of sophisticated semantic analysis to annotate documents with rich information, and extract important pieces of information.

The first of these kinds of semantic analysis is robust semantic role parsing. A semantic role parser assigns a shallow meaning structure to a sentence, essentially labeling the different entities in a text, and the meaning-relations between them. For each predicate (e.g. verb) in a sentence, a semantic role parser finds the arguments of the predicate. For example in the sentence “Microsoft acquired Google on Saturday” , “Microsoft” is the AGENT of the acquiring, “Google” is the THEME, and “on Saturday” is the TIME.

The second kind of semantic analysis is event extraction. Events correspond to a larger semantic unit than the single predicate focused on by the semantic parser. Where the semantic parser focuses on a single clause at a time, events correspond to a description of “what somebody did to whom, when, and where”. Events are hierarchical; larger events can be composed of smaller atomic events.

The third kind of semantic analysis is opinion extraction. Identifying opinions and other subjective information is crucial in answering questions about what a given person, organization, or nation might believe.

The fourth kind of analysis is identification of contradictions, timelines, and perspectives.

#### ***B.3.1. Semantic Role Parsing***

One of our core technologies is our semantic role parser, which annotates input sentences with the roles played by constituents relative to target predicates (verbs).. During Phase I, the performance of this basic semantic annotation system has improved dramatically, from precision/recall of 62/62% to 87%/80%. We achieved this large improvement by implementing a completely new classification system based on Support Vector Machines (SVMs), and adding and modifying the features used by our classifier. Our current system first produces a syntactic parse of a sentence (using the Charniak parser (Charniak, 2001)), decides which constituent nodes are semantic arguments, and then uses the SVM classifier to assign a semantic label (AGENT, PATIENT, INSTRUMENT, etc) to each such node. These labels are derived from the PropBank (Kingsbury and Palmer, 2002) corpus, in which the Penn TreeBank was hand-labeled with semantic roles for the arguments of each verb in the corpus. Our 87%/80% performance (reported in (Pradhan et al., 2003)) is the best performance currently reported by any laboratory for the PropBank labeling task.

In Phase I we also developed a prototype of a new semantic labeling algorithm based on SVMs that treats the problem as a chunking task (Hacioglu and Ward, 2003). Rather than rely on a syntactic parser to identify constituents that are then classified, the SVM is used to both segment and label the semantic roles. Because this architecture does not require a full syntactic parser, it should be easier to port to new languages, and should be more robust for processing transcribed speech. We have only just begun our evaluations of the system, but our preliminary results are very encouraging.

Although the improvements in our semantic parser have been dramatic in Phase I (from 62%/62% to 87%/80%), we have still not achieved our goals of building a tool that can construct a robust, domain-independent semantic representation. Our first goal in phase II will be to improve the basic performance of the parser, since even 87% is still far from perfect. We will explore new features for our statistical classifier. These include novel semantic features for words, such as features derived from ontologies. We will also investigate adding features from dependency structures to those from constituent structure parses. Our parser is able to produce a role lattice as well as the highest probability set of roles. We will investigate the use of additional techniques to re-process the lattice, analogous to producing a word lattice in a speech recognizer and using more sophisticated language models to process the word lattice for improved performance. We will also increase the size of our training set. During Phase I we have begun to analyze the differences between the FrameNet (Baker et al., 1998) and PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002) corpora in an attempt to combine them. We discovered that the relationship between the two is complex, and that integration will require careful alignment and potentially an interface representation.

In addition to improving the basic performance of the semantic parser on verbs, we will improve the depth and sophistication of the semantic parser output. Two augmentations are required immediately. The first is an extension of the parser from verbal to nominal predicates. Currently the parser only labels the arguments of verbs. But a large proportion of the semantic predicates in a sentence come from nouns or adjectives. For example, in the sentence “The foreign minister voiced Jordan’s belief that the roadmap represents the only alternative” the noun *belief* has the arguments “*Jordan*” and “*the roadmap represents the only alternative*” in the same way that a sentence with the verb *believe* would. Our current parser cannot handle such sentences, which are quite common. The second key augmentation is to combine the semantic parse of different predicates in the sentence. Currently the parser outputs a separate representation for the arguments of each predicate. In order to be useful for other parts of our question answering research, the parser needs to combine these representations into a single hierarchical semantic representation for the whole sentence.

The result of these augmentations will be a crucial tool for our Phase II work: a semantic parser that can take an entire sentence, find and label the semantic arguments of all predicates (nouns and adjectives as well as verbs), and integrate them into a single hierarchical semantic representation, and do all this with extremely high accuracy.

We will then fully integrate our semantic parser throughout our question-answering system. This includes improvements to areas we have already begun integration in Phase I. For example the semantic parser will be used to further improve our answer type classifier, by including more information about the semantic role of the answer type as a crucial feature, and we will continue our work on porting the semantic parser to Chinese and Arabic, previously supported by the KDD project. The key new use of the semantic parser, however, will be in providing rich, specific labels for improving extraction of information for response generation. The semantic

parser will be used to extract key information that will be useful for each question-type. Each question type (fact, opinion, biography, definition, event) depends on extraction of key information from potential answer text. This information can be viewed as a kind of semantic role. The semantic annotation produced by the parser will be used to extract these roles. For example, the previous section already discussed how the semantic parser output can be used to extract events. Similarly, opinion texts will be labeled with the OPINION-HOLDER and the OPINION, roles that can reliably be derived from the more general AGENT and PROPOSITION roles now produced by the parser. Definition texts will be labeled with roles such as GENUS and SPECIES, biography texts with role such as BIRTH-PLACE and BIRTH-DATE, and so on.

This mapping from more general roles to question-specific ones will be learned from human annotated training data using methods that leverage statistical, linguistic and knowledge-based methods. One method will be to train separate versions of the parser for each of the question types. Another will be to produce type-independent thematic roles, and then use a small set of induced rules to map the thematic roles onto type specific roles. We will use combinations of rule-based, supervised and unsupervised models to segment and label the input for extraction. The supervised and unsupervised approaches are nicely complementary; combining these two approaches via rover (Fiscus, 1997) and other methods will allow us to benefit from statistical, knowledge-based, as well as linguistic methods.

### B.3.2. Identifying and Annotating Opinions

Identifying subjective information and segregating it from factual statements is an important dimension in organizing information relevant to a question. Opinions exhibit a clear difference in their support or opposition for a given position and the response can be organized around that position. Our work in Phase I has focused on identifying opinion sentences and separating them into positive and negative statements. To this end, we implemented a Bayesian document-level classifier, trained on Wall Street Journal articles labeled as “news” or “business” (approximately corresponding to facts) and “editorial” or “letter to the editor” (approximately corresponding to opinions). The classifier achieved very high performance (97% F-measure) when cross-validated on 4,000 articles from the Wall Street Journal. For the more difficult problem of deciding opinion status at the sentence level, we explored three complementary strategies: lexical similarity of a given sentence to all opinion and all factual documents within a given topic; Bayesian classification with approximate training labels inherited from the document level; and calculating the log-likelihood ratio of words in the sentence that have been automatically found to be loaded for or against a position (Yu and Hatzivassiloglou, 2003). The combination of the three methods led to 80% accuracy in detecting opinion sentences in unseen texts. We detect loaded words automatically by bootstrapping from a small set of known such words (e.g., *good*, *awful*) and expand to all word classes with a measure based on word co-occurrence. These results improve upon earlier published results on the similar task of detecting subjective sentences (Wiebe et al., 1999; Hatzivassiloglou and Wiebe, 2000).

To generate a response for opinion questions, we are using the bottom-up techniques and ordering strategies that are used in DEFSCRIBER (Section B.4.1) and adapting them to select and order opinion sentences. We use as input the sentences that were identified by the analysis component to be opinion sentences about the questioned topic (e.g., *welfare reform*). This can be a very large set of sentences and it is the task of the response generation component to filter and order these sentences. It does so by first clustering *pro* and *con* sentences separately, producing themes of similar sentences. It then selects themes to include in the response by choosing one

sentence from each theme that is most closely related, lexically, to the preceding sentence in the response. Our current system is able to generate coherent paragraph-length responses to a request for an opinion, as well as the individual subjective sentences shown in Figure 1.

A second direction of research related to opinions that we started exploring in Phase I involves taking advantage of the semantic structure of sentences, and in particular the fact that many frequent verbs dictate which part of the sentence contains the actual opinion (e.g., the propositional complement of *believe*). We have annotated 3,041 sentences with labeled propositions in FrameNet (Baker et al., 1998), marking each proposition as OPINION or OTHER, and used this corpus to extract verb-specific opinion constraints. Finally, we built a prototype opinion-labeler by training our standard semantic parser on these hand-labeled opinions, achieving precision of 62% and recall of 61%.

In order to effectively answer opinion-related questions, we need to further analyze the opinion sentences and provide additional structure that will be used by our universal answer strategy. While in Phase I we classified entire sentences as opinions or not and as generally positive or negative, in Phase II we will refine the output of this module so that components of opinion sentences are annotated with specific roles relative to their opinion.

One such component is the actual subjective part of the opinion sentence, as opposed to additional factual material; often a sentence that contains subjective clauses expresses an opinion only in the main part or one of the clauses. For example in the propositional opinions discussed above, the opinion is only expressed by the complement of verbs like “*believe*”; the rest of the sentence is non-subjective. We propose to use information from the semantic parse of the sentence to delineate the different clauses and their relative structure and re-score each such part with our statistical and lexical techniques to automatically locate the opinion component. For propositional opinions, these two passes can be combined into a single pass by using “loaded” or opinionated words directly as features in the semantic parser’s support vector machines (SVM) classifier.

Other crucial components include the opinion-holder (which person or organization expresses the opinion), and the target (what the opinion is about). We have already labeled 2,000 opinion-holders in the PropBank database (Kingsbury and Palmer, 2002; Kingsbury et al., 2002). We will use the semantic parser, trained on these hand-labeled opinion-holder labels, together with information about commonly repeated words and phrases in a set of opinion sentences found for the same question to locate the opinion-holder and target. In both this task and in further refining our classification of opinion text, we will use information from WordNet in tracing not only loaded words but also their hyponyms and hypernyms.

Finally, we will attach to each selected opinion sentence information about its source, geographical origin (derived from source information but overridden by explicit information in or nearby the analyzed sentence), and time. We will obtain time information by a combination of existing shallow tools (Mani and Wilson, 2000; Mani and Wilson, 2001) and the role labels assigned to prepositional phrases by the semantic parser. Once the target of the opinion has been identified, we will calibrate the positive/negative score of the opinion part of the sentence with regard to that target, so that instead of generally positive or negative classes we will be classifying by the end of Phase II opinions according to their position relative to the specific subject mentioned in the question. We will still focus on primarily binary (pro and con) viewpoints regarding a given subject during Phase II.

Once opinion text fragments have been detected and separated, we will pass them on to the clustering component for eliminating redundancies. Then our generation component will pack

together similar information from multiple input sentences and reorder the sentences to maximize cohesion. The generation component will utilize the different annotations (target, time, source, geographical location) to provide appropriate organization for the answer according to the focus specified in the question and to trace the source of information (and misinformation). By focusing on different annotation elements in the opinion fragments, we will produce by the end of Phase II output that is one to several paragraphs long and will answer questions such as:

- “What is Ariel Sharon’s opinion of Yasser Arafat?”
- “What different opinions have been expressed on Mahmud Abbas’ resignation?”
- “How do perspectives on Israeli settlements differ between Israel and the United States?”
- “How have opinions about the road map evolved during the summer of 2003?”

### *B.3.3. Event Extraction*

Events in text correspond at a basic level to a description of “what somebody did to whom, when, and where.” As such, they can form the building blocks for answering low-level factual questions about individual actions. More interestingly for us, events are linked, and together they form complex descriptions of other events at a higher level. For example, an event such as Mahmud Abbas’ recent resignation can be decomposed into smaller atomic events such as the actual submission of the resignation, reactions to it by Israel and the United States, attempts to persuade him to reconsider, and its final acceptance by Arafat. Events can also be linked by other relationships even when they are not part of the same larger event: for example, we may be interested in a progression of related, successive events across time, or about all the events where a particular person was involved.

In Phase I we focused on primary operations on events, detecting which sentences in the text describe events, and analyzing them to identify the participants, locations, time periods, and main actions involved. We first conducted a study by annotating text with event/non-event labels and automatically estimating the ability of features such as pronouns, proper names, and numbers to signal an event. Using information from this study, we built a prototype system (Filatova and Hatzivassiloglou, 2003) that starts from pairs of named entities (found to be most likely to signal an event) and uses frequency and repetition information to highlight those pairs that are most likely to contain the important constituents of events (e.g., actors). Our system is able to classify a pair like <“China Airlines Flight 676”, “Bali”> as event-related (flight 676 crashed on Bali’s airstrip) while rejecting the pair <“Bill Hazzard”, “New York Times”> (Bill Hazzard is a journalist for the New York Times). Pairs of named entities that are linked in an article but also appear in multiple other topics (e.g., <“Arafat”, “Sharon”>) are penalized, so that the most specific pairs of named entities are selected for each topic.

Once we have obtained pairs of named entities, we test the different words between them and keep those that are most prominent for that pair and topic, either directly or via synonyms and hypernyms in WordNet (Miller, 1995). This filters out many of the associations between named entities that occur by chance. We combine the remaining pairs and labels (words linking them) in a graph, merge entities in that graph (for example, “Taipei” and “Taiwan”) on the basis of shared links, and construct links between entities involving multiple atomic links. We have built two interfaces for displaying the output: one graphically presents the labeled links between the entities, and the other traces these entities back to the text and displays the sentences most closely matching the two named entities and the label(s) for the atomic event.

Our Phase I research has led to a robust domain-independent event detection module which suggests the most salient pairs of event-related named entities and proposes single-word labels

that describe the event. The quality of this module's three outputs has been rated at precision values of 75% for valid relationships, 63-68% for importance ranking, and 40% for event labeling. In Phase II, we will initially extend this module by incorporating multiple-word labels for events and generalizing the input used by the module, so it can be applied to arbitrary collections of text. We have found that the quality of the extracted events increases substantially when multiple texts on the same topic are used; this enables us to rank higher the most salient events. Presently, topic labels have to be supplied in our texts for the event module to take advantage of them. In order to automatically obtain topic information, we will link our event detector with document-clustering technology similar to that produced by Topic Detection and Tracking systems, but augmented with linguistic information about noun phrases and named entities (Hatzivassiloglou et al., 2000).

Beyond these two extensions, our main goal in Phase II will be to move from *event detection* to *event linking*. Linking related events will allow us to provide long answers on questions about a given protagonist or the evolution of an event across time. By detecting links between events, we will also construct representations of composite events combining information from simpler events, a process that will be recursively repeated on multiple levels. Our Phase II work will involve attaching to events new labels, including source, geographical origin, and time, and subsequently exploring links based on these attributes and shared participants (already detected in Phase I). For extracting time phrases, we will utilize existing shallow syntactic detectors for time phrases and the labels the semantic parser produces. We will extend techniques developed for Columbia's NewsBlaster (McKeown et al., 2002; McKeown et al., 2003) (funded by the TIDES and KDD projects at Columbia University) for annotating text pieces with geographical locality information and assessing the importance of geographical information for a given question (Gravano et al., 2003). We will also use a combination of syntactic and statistical techniques, together with information such as the AGENT role obtained by the semantic parser, to determine the direction of actions between two or more protagonists (who is the initiator and who is the recipient).

This fully unsupervised approach to extracting and linking events will be complemented with a second, supervised approach that makes extensive use of the role labels provided by the semantic parser. In this supervised approach, we will map the predicate-argument relations and semantic roles (AGENT, LOCATION, etc.) produced by the semantic parser into domain-independent atomic events and their participants. The two approaches are nicely complementary; one requires no training data but produces weakly labeled output (with "action" words), while the second approach will require training data but will produce atomic events which have event labels as well as labels for the roles the participants play. Combining these two approaches will allow us to benefit from statistical, knowledge-based, as well as linguistic methods.

A similar dual approach will be adopted for linking events. In addition to our statistical method based on observed sharing of event components, we will explore an event merging algorithm that maps atomic events to macro events using training examples labeled with linking event relations such as CAUSE, RESULT, PRECEDE, FOLLOW, ENABLE, and SUBEVENT. Once again, we will combine the supervised and unsupervised approach to create a merging algorithm which has the advantages of both statistical and knowledge-based methods.

#### B.3.4. *New Analysis Directions: Contradictions, Perspectives, and Timelines*

Our approach to question answering is predicated upon the fusion of information from multiple sources. For questions that require several different pieces of information to be presented in the answer (such as a question asking about world reaction to Mahmud Abbas' resignation), information fusion can identify text segments that are similar and aggregate them. The next challenge is to meaningfully organize the presentation of information nuggets that stand in relationship to each other. For example, an official Israeli source may express its disappointment in Abbas' resignation, while an Egyptian source may accuse Israel that it secretly undermined Abbas in order to sabotage the roadmap process. To enable our generation component to highlight this contradiction between the two sources, we need not only semantic annotation at the individual statement level, as we argued in Sections B.3.1-B.3.3, but also the detection and annotation of *relationships* between pieces of information.

We propose to explore three such relationships in Phase II: Contradictions, as in the example above; differences in perspective and geographical source; and timelines, progressions of events that share one or more common participants and can be both linked to each other and ordered on a time line. For all three approaches we will utilize the rich semantic information that we obtain in the form of semantic roles, event labels and roles, and opinion classes. We will also adapt our generation component to prioritize detected contradictions, differences in perspective, and event timelines in its reordering of the information and its choice of linking words between different text segments.

Identifying **contradictions** requires identifying direct negation of facts and antonyms used in wording, and finding semantic representations that are similar with the exception of one or two roles. Contradictions range from the simplest cases where a single number changes between two related sentences, to cases where the meaning of the words and differences in syntactic structure express contradictory information. We will adopt a two-step approach, where we will first locate parts of two sentences that are almost identical in meaning (starting with the easiest case, repetitions), and parts that are in direct opposition. We will use our capability to detect similar pieces of answers, as developed in Phase I (Blair-Goldensohn et al., 2003) for the first task, and shallow syntactic analysis and knowledge of antonyms from WordNet (Miller, 1995) for the latter task. The semantic parser will segment the text into fragments such as propositions that can be compared with this approach; it will also provide us with additional constraints on which of the many combinations of fragments we should compare. For example, we will use the semantic parser to match short phrases that are associated with the same concept, even though the wording may be different (e.g., recognize that two numbers should be compared if they both refer to the count of human victims in a terrorist attack). We will also leverage our opinion classification system by matching positive and negative sentences that are about the same topic and otherwise similar.

We will extend this approach to deal with **multiple perspectives** in the second year of the contract. Unlike contradictions where we only need to detect that two pieces of information are in direct opposition, for multiple perspectives we need not only to detect oppositions but also link them to specific positions. We will use a statistical technique based on extracted geographical locality information (Gravano et al., 2003), examining the similarity of sentences within a geographical region to sentences outside that region that are otherwise similar in content. Subsequently, we will automatically explore if this variation in similarity can be detected without powerful source labels such as those provided by geography. We anticipate that this could well lead to the automatic formulation of representations for perspectives on the basis

of identified signal words, just as we currently automatically find words loaded for or against a particular subject (Hatzivassiloglou and McKeown 1997; Yu and Hatzivassiloglou, 2003). Again, we will use the semantic parser's breakdown of sentences into propositions and semantic roles to constrain our search. We will also explore a combination of supervised and unsupervised techniques; if an analyst knows that pro-Israeli sources are unlikely to use the term "al nakba" (literally *catastrophe*, and used by Palestinians to refer to Israel's founding in 1948), this word (or its absence) can be used to automatically identify and link sources according to this perspective.

For **timelines**, we will expand upon our work in event detection and linking to identify new time information and will use this input to extract and contrast changes. An example of the information we want to link is an unfolding event, where a number (e.g., the number of casualties) or the status of a person is regularly updated. We will incorporate the detection of time phrases in our event system, and will explore ways to formulate links across time as well as a measure of similarity between atomic events. For the former, we will automatically determine the time scale of an event (which will vary depending on the level of granularity we are interested in) by aggregating information across multiple events that are linked together in retrieved documents. We will base our similarity of events not only on the lexical similarity of the predicates or labels but also on the compatibility of shared participants and locations. Linking events through time will be a step in our vision of obtaining a set of general operators for events such as decomposition and aggregation.

#### ***B.4. Generating Complex Answers***

When generating responses to open-ended questions, a single sentence is unlikely to provide enough information. Instead, the system must be capable of pulling information from different sources, fusing it to remove redundancy and generating a sequence of coherent sentences that convey this information. A key contribution of our work is the combined use of language generation and summarization techniques to identify appropriate content and generate fluent sentences for the response. We use generation strategies, such as content plans, to represent the kind of information needed to respond to a particular question type, and summarization strategies to identify and select repeated information on the web.

Given a query and a set of document segments relevant to that query, our research investigates techniques to fuse information from the different segments to produce an answer. Critical to this task is the problem of selecting fragments of text from different documents that should be included in an answer and determining how to combine them, removing redundancy and integrating complementary information fluently. To date, we have experimented with the use of two strategies to select and merge information across the segments: fusion of similar information and top-down organizational strategies which may specify both additional information to extract as well as how to order and frame the response. We are exploring the use of semantic information derived from semantic parses of retrieved text in each of these processes.

In Phase I, we experimented with different strategies for different question types. For the next phase of the proposal, we will explore additional strategies for content selection that exploit new sources of potential content, including both text and data. A key problem will be fusing different types of information, from data elements, to semantically typed phrases such as named entities, to untyped strings such as clauses and sentences. We will need to extend each of the strategies individually, test their use across question types, extend them for use with new question types

and determine how they can best be integrated. This central answer generation module will resolve redundancies, merge compatible facts, highlight contradictions and organize the paragraph-length answer to maximize cohesion.

In this section, we first describe our achievements and ongoing work on response generation for definitions and biographies. We then describe our proposed work on new strategies and a unified generation strategy that can be used across question types.

#### B.4.1. Definitions

Definitional question answering is concerned with questions of the form “What is X?” Over the course of Phase I, we have implemented DEFSCRIBER, a system that generates multi-sentence definitions to answer such questions from Internet documents, using an innovative combination of top-down and bottom-up strategies (Figure 1 shows an example of DefScriber output). The bottom-up techniques in DEFSCRIBER use similarities and themes from the data to determine content. These are statistical methods often used in text summarization, including similarity metrics coupled with clustering (Hatzivassiloglou et al., 1999; Lin and Hovy, 2002a; Radev et al., 2000). This portion of DEFSCRIBER produces *themes* as output, each of which consists of a set of sentences about the same thing. Each theme will potentially contribute one sentence to the response, depending on answer length. The top-down part of DEFSCRIBER uses a set of *definitional predicates* to identify types of information which should ideally be included in a definition. These predicates model core properties of definitions discussed in the literature (Sager and L'Homme, 1994) and identified in our own research. Of the predicates we have identified, *genus* and *species* are probably most central to definitions, modeling superordinate information and discriminating attributes (as in *The Hajj is a type of ritual beginning in the twelfth month of the Islamic year*, where the genus and species are shown by the two underlined parts). We use two methods to automatically identify the predicate types in text: feature-based classification from machine-learned decision trees, and pattern-recognition (Grishman, 1997) using patterns manually extracted from a hand-marked corpus we created.

Bottom-up techniques are fully implemented in DEFSCRIBER and we have currently implemented three of the predicates: *genus*, *species* and *non-specific definitional*. DEFSCRIBER generates a paragraph-length response to any question for which information exists on the Internet.<sup>2</sup> Once the top-down and bottom-up techniques have determined potential content, we use ordering techniques based on coherence to prioritize and select the sentences in order for the response. We evaluated DEFSCRIBER in Phase I using an online questionnaire asking subjects to rate responses for relevance, redundancy, structure, breadth of coverage and term understanding. 24 terms were selected for the experiment and 38 judges participated. Our evaluation indicated that the combination of techniques used in Phase I in DEFSCRIBER achieves significant improvement over any single technique used alone, obtaining the best performance in four of the five categories.

In Phase II, we will be continuing our work on top-down strategies. We have identified five additional predicates that would be useful for definition; these include *target partition* (which divides an entity into or more concepts), *history*, *cause*, *effect*, and *etymology*. Furthermore, our implementation at this stage involved a fair amount of manual effort in annotating corpora from which we constructed patterns for recognizing sentences that match a predicate. In the next phase, we will investigate the use of statistical learning techniques that will help us to automate

---

<sup>2</sup> We have also implemented the ability to look for definitional information inside a specific collection, such as the Center for Non-Proliferation Studies (CNS) collection.

the development of additional predicates. Our implementation of these predicates will require experimentation with response generation given the large number of predicates. We will study how predicates interact with each other; when text for multiple predicates can be found in the input, which ones have priority?

In addition to adding new predicates, we also will improve the recall of existing (and new) predicates. Our current system retrieves only one or two sentences for a *genus/species* in the context of a specific question, although there are many *genus/species* sentences in the relevant documents that are provided as input to DefScriber. Once more sentences are returned, we will need a more sophisticated mechanism for fusing information across the sentences to produce a single *genus/species* sentence that adequately represents input for the response. This approach will be used for all predicates. As we investigate methods for fusion and extend the predicate implementation, we expect to move closer to language generation approaches, where instead of selecting a full sentence from the input, we will select matching phrases that can be combined to produce novel sentences for the output. Finally, we will explore how context of an information gathering session affects generation of definitional descriptions. When the user engages in dialog with the system, we can use previous questions to narrow the scope of a suitable answer. Content of a definition will be constrained in these cases by context. For example, if a user has been querying about oil fields and then asks for a definition of “SPUD”, DefScriber can rule out any sentences having to do with potatoes.

#### *B.4.2. Biographies*

Our approach to biographical descriptions also uses top-down strategies, but within a language generation framework. The generation framework should allow more control over exactly what information gets included in a response since it selects potential context from a semantic representation and not from text. It also has the potential for producing a more fluent response since wording and ordering are determined entirely by the system and not from the input text. However, language generation systems to this point have been quite domain dependent; it has not been possible to use them in unrestricted, open domains such as question answering. Our work from Phase I has us poised for a breakthrough on this front.

In Phase I, we began to develop a system that can use information extracted from web pages to automatically generate descriptions of people. For this type of question, we are experimenting with language generation techniques that produce text from data, as opposed to the approach we use in definitional question answering, where we generate responses from textual input. Our goal is to allow the possibility of obtaining different *views* of an individual (Acker and Porter, 1994) according to the interests of personnel accessing the biographies or according to the situation at hand.

The language generation process includes three main components: a content planner which determines what information should be selected and its order in the response, a sentence planner which aggregates information into sentences, and a realizer which produces the actual English. Our focus is on the content planning component. We are using examples of biographies found on the web to automatically construct content plans that determine biography content. Such plans will guide the generation of biographies on unseen people so that, for example, different biographies are produced for a military rather than a civilian leader. The ability of our system to automatically adapt its output also extends to cover different classes of system users: for instance, different biographies of a terrorist can be generated for use by intelligence analysts versus law enforcement personnel. Our work to date includes building learning systems for

inferring order constraints on content plans (Duboue and McKeown, 2001), for automatically constructing a content planner (Duboue and McKeown, 2002), and for learning content selection rules (Duboue and McKeown, 2003).

Our work in Phase I has resulted in an initial prototype of a biographical response generation system, PROGENIE. PROGENIE uses the content plans learned by our machine learning approaches and, given data culled from web sites through information extraction techniques, generates descriptions in rough form. The sentence planner and realizer must be extended so that PROGENIE can combine simple sentences and make better word choices in order to produce more fluent output. Nonetheless, the prototype is capable of varying the description to produce different views of the same person through the use of different plans.

In Phase II, we will extend our work to date on information extraction from web pages, which provides the input for the biography generator. We will use semantic parser output to robustly identify slots in a data frame to provide input data for the biography generator. In addition to output from the semantic parser, we will also experiment with input derived from web page scraping software developed at Univ. of Massachusetts at Amherst by Bruce Croft's team. Finally, we will diversify our sources of input, using online databases (e.g., the CIA Factbook) in addition to data extracted from web pages.

In addition to robustly harvesting input data from the web, we also need to represent the words which can be used to refer to this data. During the process of harvesting data, we will develop learning and statistical acquisition techniques that can help us to build a lexicon that can be used for biography descriptions. This lexicon will associate semantics, which we expect to find in the input, with words that can be used to express this semantics.

Finally, we also will work on extending our learning techniques for acquiring content selection rules. This will allow us to incorporate a variety of content plans that can be automatically selected and tuned to create different views of the same person for end users with different needs. We will integrate the content planner with other language generation components (sentence planner and realizer) to create fluent text.

#### *B.4.3. New response generation strategies*

We propose to create new generation strategies that determine content for a response by invoking a short-answer question answering system and by using data that is available in online databases on the web. This new form of input for response content creates challenges for a generation strategy; granularity and type of the input is different from the input we have been using to date. This will influence fusion of input to create a fluent answer. The granularity of our current input is at the proposition level; for definitions it is a full sentence and for biographies, an implicit proposition is filled with data created through information extraction (e.g., <Person> was born in <X>). A short answer question answering system will either give a sentence containing the response or just a fragment of text containing the response. A database will provide simply a value. We will develop a top-down strategy consisting of a plan that makes a sequence of calls to a short-answer QA system. We will need to determine:

- How can we merge short-answer QA responses to create a fluent and concise answer?
- When and how do we need to remove information from a short-answer QA response given the larger context of the long-answer question?

#### *B.4.4. Integration of strategies and use across question types*

During Phase I, we developed alternative strategies for long, complex answers that we applied to distinct question types. In Phase II, we plan to leverage the results of these explorations by integrating the different strategies into a single, flexible approach. For example, currently we are using a hybrid strategy for definitional questions applied to text (i.e., top-down plan consisting of predicates interacting with a bottom-up strategy looking for data similarities) and a top-down, learned, content plan in the traditional language generation style (i.e., specifying semantic constraints on content) for biographical questions. This has allowed us to experiment with different approaches in separate testbeds for different question types. But there is no reason why our definitional approach could not have been used for biographical descriptions nor why our biographical approach could not have been used for definitional questions. Taken one step further, using all strategies together, including proposed new strategies, we might be able to exploit the advantages of each strategy for all question types.

A key challenge will be the creation of a single strategy that integrates rhetorical and semantic constraints on selection of content. Our content plan for biographies selects content using semantic specifications on information and ordering, while the definitional predicates we use are more rhetorical in nature (e.g., genus-species, exemplification). As an example, consider the issues in augmenting our strategy for definitions with calls to a short-answer QA system; the strategy would have to be augmented to indicate where such material could be used. The definitional strategy retrieves sentences from text. We might use the responses from short-answer QA to augment the sentences with additional, related material, adding in the returned phrases or values as phrasal modifiers in the existing sentences. We suspect that this approach might work well for specific types of additions such as the addition of modifiers to referential noun phrases in order to clarify the referent for the reader (Nenkova and McKeown, 2003). An alternative approach would be to use language generation techniques that can handle both strings and semantically typed values or phrases as input to generate a totally new sentence in output.

#### *B.4.5. New question types*

We will apply our response generation strategies to new question types involving perspective, comparison, contradiction and change over time. For example, we will address the generation of responses to questions such as “How does Iraq’s position on religion and government differ from Egypt’s?” or “How has the situation of women in Afghanistan evolved since 1999?”. Part of our approach to responding to these questions will involve new research on analysis, described in Section B.3.4. In the response generation components, we will need to consider how the contradictory, contrastive and dynamic facts identified through analysis can be integrated in a response. When there are many facts that are identified, how do we select which ones to highlight in the response, how do we merge them, and how do we order them?

In addition, we will apply the strategies to questions about events. We have made progress over the first phases of the Aquaint program in analysis that identifies persons, locations, organizations and places and the relationships between them. In the next phase, we need to take the information produced by analysis of events and form coherent responses to questions such as “What was the sequence of events leading to the resignation of Abbas?”.

#### *B.4.6. Long Answer Evaluation*

Several of our approaches for answering questions with long answers involve issues similar to those explored in automatic text summarization, in particular eliminating redundancy and

improving readability of the output. Consequently, we have explored new evaluation techniques for summarized text, which we intend to specialize on generated answers. We have developed a scoring procedure we refer to as *the pyramid method* for evaluating content in human or machine-generated summaries (Passonneau and Nenkova, 2003), which allows prioritization of information (Radev et al., 2003). Our first step involved manual analysis of summaries from DUC 2003 (Nenkova et al., 2003); we developed a consistent procedure for manual identification of the same Summary Content Units (SCUs, (Jing et al., 1998)) across different summaries.

A pyramid score of the content in a summary depends on first constructing an SCU pyramid, a partition of the set of relevant SCUs by their weights, which are determined based on their frequency. The sets which occur in many summaries tend to have fewer members. The sets are thus vertically stacked with the largest set of lowest relevance on the bottom. Our pyramid formula expresses quantitatively the notion that a higher scoring summary should draw SCUs from the top tier of a pyramid and exhaust that tier before taking SCUs from the next tier.

During our phase I funding, we also collected a dataset relevant to our evaluation. To constrain the problem and maximize the ease of data collection, we began with definitional questions, where the rhetorical and semantic constraints on answers have been well studied (e.g., (McKeown, 1985; Blair-Goldensohn et al., 2003; Klavans and Muresan, 2000)). We chose medical consumer texts because it is a high interest area to the lay person, people who seek medical information on the web are often highly motivated in their information seeking tasks, and the large amounts of information about the same thing distributed across multiple pages makes it a good fit for an abstracting approach to QA.

In Phase II we will apply the pyramid scoring method to evaluation of open-ended answers in QA, by working on automated methods for two tasks:

- scoring a summary, given a pyramid consisting of a partition over a relevant set of SCUs, according to their weighted importance
- constructing such a pyramid for a set of source texts, given a set of human-generated summaries, source texts, and other data

We will use our SCU annotations of the DUC summaries, along with other data, as training material for developing automated methods for both tasks, along with other data sets.

The second task will make use of our current medical paraphrase dataset, consisting of 400 propositions and 6,000 paraphrases, in conjunction with the SCU annotations of summaries of DUC 2003 newswire to investigate information alignment based on a gradient notion of alignment, and where alignment is not of one textual paraphrase with another, but of textual expressions with a more abstract notion of content. The propositions in this dataset were selected based on their relevance to definitional questions about medical terminology. For example, we assume that an answer to the question "What is hemophilia?" leads to a different relevance weighting of the information in texts about hemophilia than would be appropriate for a general purpose summary. Thus we will also use this dataset to investigate how the question affects the prioritization of information in the text. Our goal is to arrive at a more general pyramid scoring method that takes into account the information seeking goal.

### ***B.5. Diverse Data Modalities***

Our approach to question answering is unique in its use of sophisticated algorithms for response generation: generating complex answers by fusion information. But even the best response generation system is useless without a wide range of information to fuse. A

sophisticated question answerer must be able to combine information from many kinds of data sources, including structured and unstructured text, web pages, or audio documents (speech recordings), as well as text in foreign languages. *Accessing, retrieving, and integrating diverse data sources* is a key goal of the AQUAINT 2 project.

We propose to develop the capability to query each of these diverse data modalities. We will combine information from multiple text sources such as the TREC, AQUAINT and CNS collections; Google; structured and semi-structured data sources such as databases, tables, and web pages; spoken-language documents via speech recognition, and we will partially support text documents in Chinese and Arabic. All possible sources will be queried to return answer candidates. Returned information from all of the sources will be integrated into a single response. Our approach is thus the *Diverse-Data* strategy, as defined in section 5.2.3.2 of the BAA, including structured web data, recorded speech, three languages (English, Chinese and Arabic), and at least two genres (Newswire and News Broadcast).

Incorporating these multiple diverse data sources requires a way to select among and integrate the results from various sources. The system must be capable of using our domain independent event-oriented representations to integrate results across the various materials being consulted. More specifically, the system will attempt to identify and merge representations that refer to the same events, detect and report events that are similar and likely to be of interest to the user, produce timelines of events reported in different sources, and note contradictions across sources. The next two sections describe in detail our plans for dealing with foreign languages (Chinese and Arabic) and for dealing with spoken data.

#### *B.5.1. Multilingual Support for Chinese and Arabic*

A key part of our project is to extend our system to other languages, both to test the robustness of our algorithms to language-independence issues, and to provide more diverse data sets. We have chosen two key languages to focus on: Arabic and Chinese.. It will not be possible in the time-course of Phase II to port our entire system to work in these two new languages. It is our goal therefore to focus specifically on processing of questions, and relatively shallow processing of answer documents. This will require that we have Chinese and Arabic versions of the following components:

- the semantic parser
- the question classification module
- term indexing of all documents
- simple event extraction

The first of these components, the semantic parser, will leverage the results of our prior one-year project in the KDD program. That project focused on producing prototype semantic parsers for Chinese and Arabic. For Chinese, we produced a small training set by hand-annotating semantic roles for 1000 sentences. We then ported the Collins (Collins, 1997) syntactic parser to Chinese, including writing new Chinese head rules, and trained the parser on the Penn Chinese TreeBank (Xue et al., 2002). The resulting Chinese syntactic parser has quite high accuracy; precision/recall of 83%/86% on the standard Chinese TreeBank test set. We then ported the remainder of our English semantic parser to Chinese, including Chinese named entity tagging. Our current prototype Chinese semantic parser achieves a precision/recall of 72.5%/60.2% given perfect parses. This result of our KDD project is a good start, but in order to be useful in question-answering, higher accuracy will be required. One improvement will be to add more

data by using the forthcoming Chinese PropBank, which should be available in the spring of 2004. Our Arabic semantic parser is still in design. Like the Chinese system, this will require a labeled training set, a parser, a named entity tagger, etc. We are currently working on porting the Collins parser to Arabic to train on the Penn Arabic TreeBank (Maamouri et al., 2003). Since Arabic has relatively complex morphology, we expect that we will need to modify the code to use lemmatization and morphological stemming algorithms. We currently plan to use the morphological analyzers available from the LDC. Since there are no plans for an Arabic PropBank, we are currently labeling a small (1000-sentence) training set, modeled on our Chinese training set, to train the Arabic semantic parser.

The term indexing of all documents is relatively simple to port from our English system. We will build a rule-based question-type classification module, based on standard Named Entity tagging software for Chinese. For event extraction, we will investigate both the CSLR event-extraction approach based on the output of the semantic parser and the Columbia approach based on predicate-argument induction from frequently occurring predicates in documents.

### *B.5.2. Spoken Language (Audio) Documents*

Spoken language data is now widely available, including radio broadcasts, transcriptions of speeches, lectures, and so on. We propose to augment our question-answering system to allow it to draw answers from these kinds of spoken language data. We will use the state-of-the-art real-time CSLR SONIC speech decoder (Pellom, 2001) for this task. Like all modern speech recognizers, SONIC will not produce error-free output. Transcriptions of most spoken language genres exact have approximate error rates of 20–30%. While high, such error rates have been shown to give acceptable performance for segmenting and keyword indexing, and so the resulting transcript should be accurate enough to extract enough information for basic question answering purposes, including named entities, keywords and the basic relationships between them. The transcript should also be accurate enough for segmenting documents into smaller paragraph-sized units and for topic identification. Documents will be decoded ahead of time, as they are acquired, and indexed for topics and keywords.

We will need to extend SONIC to optimize its performance for this type of task. The challenges and proposed solutions are:

1. **Acoustic Channel** – The documents to be decoded are likely to be recorded with different microphones and under varying acoustic conditions. CSLR has developed a number of algorithms for adapting to and dealing with such input in our Digital Voice Libraries project. These algorithms currently are at a research stage and must be implemented in an efficient way into SONIC.
2. **Speaker Adaptation** – If an entire document is spoken by the same set of speakers, there is an opportunity to use speaker adaptation to improve recognition performance. The system would have to detect speaker changes in the input and segment the input into sections labeled by speaker. For each speaker, new models will be created by adapting the initial speaker-independent models. The input would then be reprocessed using the appropriate speaker-adapted models.
3. **Dynamic Lexicon Expansion** – New entity names will occur as different types of documents or more recent documents are processed. In order to recognize words not already in the system lexicon the system must generate a pronunciation model for the word and must also add it to the language model used by the recognizer. Since the system also processes text documents from the same domains as voice documents, we can mine the text for new entity

names. For recent documents we can also mine web sources, such as news wires, GlobalSecurity.com, etc. We will extract new named entities (organized by topic) from existing databases and web resources to develop a named entity lexicon and will automatically generate pronunciations for these. We will extend the techniques we developed under Phase I to generate pronunciations for new words. The focus will be especially on recognizing proper nouns.

4. **Dynamic Language Model Expansion** – New words must also be added to the language model used by the recognizer. This is accomplished by use of a class-based language model. Class-based language models estimate the probability of a word sequence by multiplying the probability of a class sequence with the probability of the words given the class sequence. New names will be tagged with the type of entity class they belong to and then added to the list of words represented by the class.

Since a key aspect of our proposal is rich semantic annotation of our data sources, we will need to modify our semantic parser and our event and opinion extraction routines to work on errorful transcribed text. Recognition errors tend to disrupt standard syntactic parsers, even if all of the key words are correct, since the syntax structure is not correct. Statistical chunking parsers should be more robust to recognition errors. We have already begun experimentation with generating shallow semantic parses from statistically chunked input. Similar extensions for increasing robustness (e.g., by lowering statistically estimated thresholds for declaring relationships between entities) will be pursued in the event extractor.

### ***B.6. Going Beyond Single Question-Answer Pairs***

Our focus to date has been primarily on answering single questions and there has been little interaction between successive questions. As part of Phase I, we implemented a first prototype dialog system which maintains key words and focus across a sequence of questions. In Phase II a large part of our effort will be in developing and using context information across a sequence of queries and answers via use of context and clarification dialogs.

We will extend the notion of context to model the sequences of information type transfers that constitute a scenario. We will also include in the context the information extracted thus far in the dialog. This context will be used by the response generation processes involved. In realistic scenarios, fact, definition, biography, event and opinion questions will be interleaved in context. The processing of any one of them should be influenced by the results from the prior processing of the other types. For example, if the user first asks a question such as “What is the status of the roadmap?” and follows this with a biography question about Mahmud Abbas, the answer to the latter question should focus on his role in implementing the roadmap and the impact of resignation to the roadmap. We will incorporate a context mechanism to uniformly represent annotated information from the scenario history, including events at various levels of granularity, relations between events, time information, perspectives and opinion holders. This ability to integrate the results of prior processing is enabled by the system’s use of common rich semantic representations across modules.

The next major issue to be addressed for dialog interaction is the generation of clarification questions. Clarifications will involve generating questions to the user to resolve ambiguities in the system. The ambiguities could arise because of multiple semantic interpretations of the user’s question or because of contradictory returned answers. When the system is unsure how to interpret something, or needs additional information to provide an answer, it will compose a question to the user to elicit the needed information.

## ***B.7. Leveraging efforts in other programs***

### ***B.7.1. TIDES and KDD at Columbia University***

For our work under the DARPA TIDES program, we have developed components for finding similar information across documents for summarization and we have used these components within AQUAINT. As part of our work on summarization, we are continuing work under TIDES to revise extracted sentences to increase the fluency of the summary. For example, we have developed techniques to replace first-mention references in extracted sentences with modified references so that it is clear who is being referred to. We also modify subsequent references, removing modifiers. We are working on more sophisticated approaches for common noun references. We are also working on techniques for ordering of summary sentences. All of these techniques can be exploited for response generation. We will use them to improve fluency of our generated responses.

We have integrated our work on NewsBlaster as part of an Integrated Feasibility Experiment (IFE) carried out by DARPA. Columbia, UMass Amherst, and IBM, under the leadership of BBN, have integrated modules across the Internet. The resulting TAP-XL system features tracking of large numbers of documents on the same event across days, translation from Arabic, named entity detection, and summarization of translated documents on the same event. It is evaluated every 4 months. We will use our experience in integrating our work in a large testbed for testbed development under the AQUAINT program.

As part of Columbia's KDD program, we are funded for research on email summarization. The language used in email is much less formal, and often less grammatical, than the language used in news and other web sites. We treat email as errorful input and are working on techniques that can help us to reduce the errors and improve fluency in summaries that are generated. We will use techniques that we develop under KDD to help in response generation from speech, another errorful source of input.

### ***B.7.2. Communicator and KDD at the University of Colorado***

Under the DARPA Communicator program, CSLR developed the Sonic recognition system, language models, dialog models and a full spoken dialog system. While these capabilities must be extended to go from the domain specific Communicator scenarios to the domain independent AQUAINT task, we are starting with a state-of-the-art dialog system, in all respects. We can also use data from the communicator task to include in our corpora for training speech system for AQUAINT.

Under the KDD project, we improved our semantic parser and ported it to Chinese. As described above in section B.4.2, our resulting prototype semantic parser for Chinese achieves Precision and Recall of 72.5% and 60.2% when using TreeBank parses. We are also in the process of porting to Arabic. Since multilingual semantic parsing will be one of the cornerstones of our Phase II work, the progress we have made under KDD will apply directly to our continued progress in AQUAINT Phase II.

## ***B.8. Collaborations with Other Sites***

We have planned interact with several other sites, using research developed elsewhere when possible to improve our end-to-end system. These sites are also interested in making use of our results.

There are three key possibilities for collaboration with the Pittsburgh/Cornell/Utah group of Wiebe, Cardie, and Riloff who also analyze textual sources to detect the opinions, attitudes and biases of the author or of quoted individuals (Cardie et al., 2003). At the data level, we will incorporate the richly annotated opinion texts collected by the Pittsburgh team in order to supplement our data, and thus improve our opinion classifiers. We will experiment with using their template output to augment our own work and do comparative evaluation.

We plan to draw on work done at the University of Massachusetts on extracting information from tables, forms, and format-rich Web pages. This will be used to more robustly use information from structured data sources for response generation.

BBN has baseline tools for multi-lingual NLP, including name extraction and parsing for Chinese and Arabic. Under a Phase 2 proposal, “Breaking the Cross-lingual Barrier to Question Answering”, BBN proposes improving performance of those linguistic capabilities and distributing them to the AQUAINT community. These tools would be of great benefit to our project.

### ***B.9. Conclusion***

The unique contributions of our research include the following innovations, which are directly responsive to the AQUAINT Phase II goals:

**1. Question Answering as Part of a Larger Information-Gathering Process:** We will answer complex questions that require cooperative interaction with the user, de-composition into component questions, and understanding context and background knowledge. Specific innovations include the use of interactive, spoken dialog to allow the user to become an active participant in finding information through clarification dialogues; speakers can ask follow-up questions, pursue a line of thought, and clarify their intentions when the system misinterpreted a question. Our linked event representation will allow us to detect potentially related information and propose alternative directions to the user.

**2. Evaluating, Validating and Presenting an Answer:** We will answer questions whose long, complex answers require integrating multiple pieces of information from diverse sources. Specific innovations include: new evaluation methodology, significant expansions of our earlier work on long-answer questions such as opinion, definition, biography, and event questions; new types of long-answer questions, including contradictions and timelines, as well as new unified methods to extract key facts and merge them to produce coherent answers.

**3. Combining Knowledge-based, Statistical and Linguistic Approaches to QA:** We will answer questions by integrating deep linguistic knowledge-sources and annotations with sophisticated state-of-the art statistical and machine learning methods. Specific innovations include new tools for extraction of rich information from documents, including significantly extended robust semantic role parsing which assigns a shallow meaning structure to a sentence, extraction of a network of events composed of smaller atomic events, and extraction of opinions differentiating between subjective and factual, positive and negative and holder of the opinion and opinion held, all based on combining large linguistically-annotated databases with sophisticated statistical classifiers.

**4. Accessing, Retrieving and Integrating Diverse Data Sources:** We will answer questions by extracting and organizing knowledge from a wide variety of diverse data sources. Specific innovations include the combination of information from multiple text sources such as the

TREC, AQUAINT and CNS collections; Google; structured and semi-structured data sources such as databases, tables, and web pages; spoken-language documents via speech recognition; partial support of text documents in Chinese and Arabic. All possible sources will be queried to return answer candidates, and the resulting information will be fused into a single response. Our approach implements the *Diverse-Data* strategy, as defined in section 5.2.3.2 of the BAA.

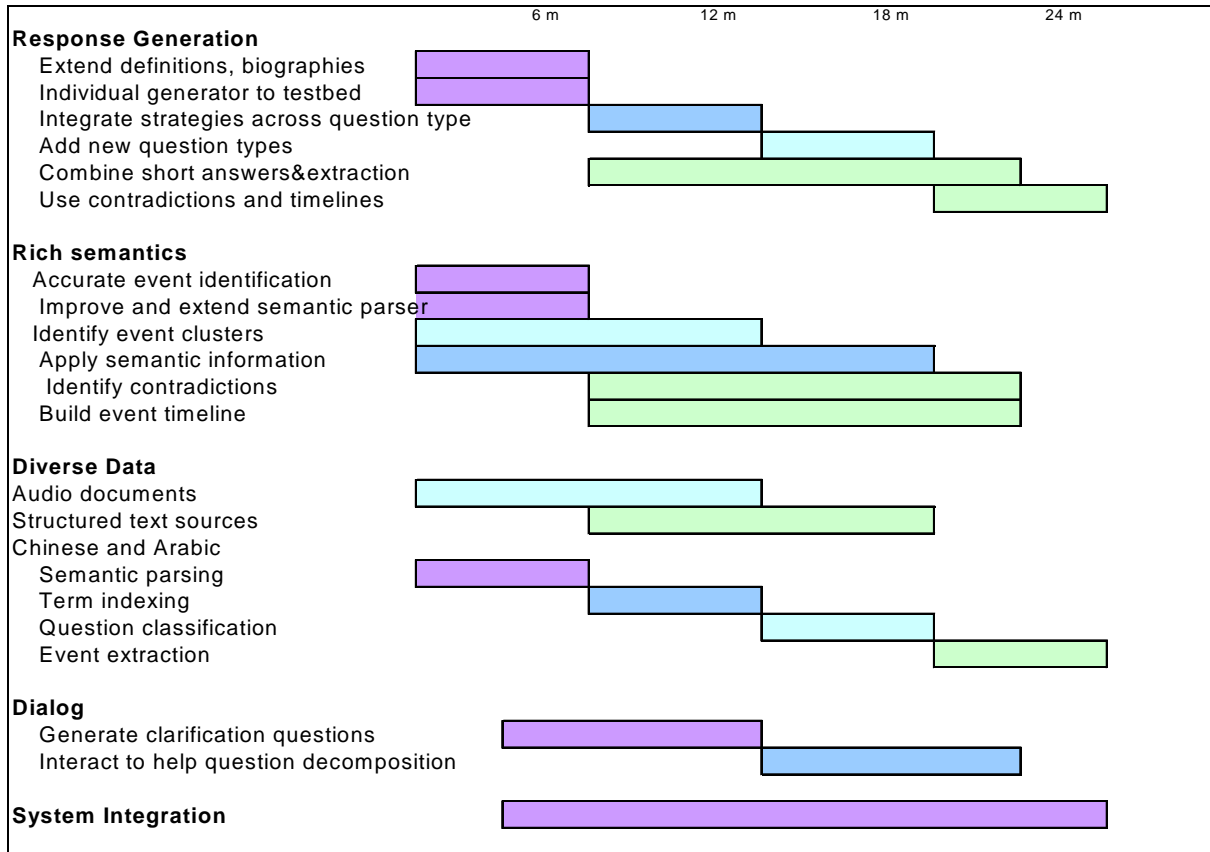
In summary, our unique integration of technologies such as information fusion, language generation, speech recognition, dialogue modeling, and Arabic and Chinese processing, together with sophisticated statistical machine learning algorithms applied to rich linguistic knowledge about events, opinions, contradictions, semantic structure, and question-types, will allow us to build a system which significantly extends the range of possible question types and responses available to analysts, and seamlessly fuses these to generate a response.

## C. Statement of Work

### C.1. Scope

The objective of the proposed work is to extend and expand the work we did under Phase I to improve the technology for answering questions that require complex answers. We will also extend our coverage of sources and languages. Our specific tasks under Phase II are:

- |  |                    |
|--|--------------------|
| 1. Extraction of rich semantics                | <b>\$689,474</b>   |
| a. Semantic role parsing                       |                    |
| b. Identifying and annotating opinions         |                    |
| c. Event extraction                            |                    |
| d. Contradictions, perspectives and timelines  |                    |
| 2. Generating complex answers                  | <b>\$372,746</b>   |
| a. Definitions                                 |                    |
| b. Biographies                                 |                    |
| c. Unified answer fusion                       |                    |
| d. Evaluation of long answers                  |                    |
| 3. Diverse data modalities                     | <b>\$415,194</b>   |
| a. Multilingual support for Chinese and Arabic |                    |
| b. Spoken language (audio) documents           |                    |
| c. Semi-structured data                        |                    |
| 4. Integrated system                           | <b>\$501,923</b>   |
| a. Interactive dialog over multiple questions  |                    |
| b. Integration of independent modules          |                    |
| c. Overall project coordination                |                    |
| <b>TOTAL</b>                                   | <b>\$1,979,337</b> |



## D. Expected Results and Technology Transfer

**Deliverables:** The primary deliverable of this effort is a multimodal question-answering system that engages in either spoken or written dialog with a user, consults a variety of spoken or written sources to identify information relevant to the user, and presents that information to the user in a natural and effective manner. The following items represent the core components to be delivered:

- Biography, definition, and opinion processing systems – these systems will classify questions falling into these categories, analyze source materials with respect to these categories, and generate responses appropriate to each type.
- Event detection software – this component will detect events in source materials and generate semantic representations of these events suitable for use in question-answering.
- Semantic parsers – including a functional semantic parser for English, along with training and testing materials developed as a part of this proposal. Prototype systems for Arabic and Chinese, along with training and testing materials developed as part of this effort.
- Interactive dialog management system – A system capable of handling both spoken and written inputs, tracking the dialog focus, as well as generating clarification and follow-up questions.

**Technology Transfer:** We will deliver all the code for all components of the system as well as all required sub-systems. We will document all of these required components including all

scripts for data pre-processing, post-processing, training and evaluation. In this documentation, we will specify all relevant interfaces (APIs) needed to allow other ARDA contractors to incorporate our code into their systems. ARDA will have non-exclusive rights to all the code and data generated as a part of this project.

We will also make our software freely available to other AQUAINT/ARDA contractors as well as to other universities and research laboratories for research purposes. We will also facilitate the licensing of this software to commercial entities in a licensed form through active programs in place at both universities. Both the Columbia NLP team and Colorado's Center for Spoken Language Research have provided a wide-variety of systems to the broad research community. Columbia's FUF/Surge generation system and SEGMENTER tool have been transferred to sites around the world. CSLR also pursues an aggressive policy of promoting and distributing its systems. Our flagship Communicator system, as well as SONIC, our state-of-the-art speech recognition system, are freely available for non-commercial use. CSLR also offers periodic short courses on these systems. In addition, we have a thriving industrial affiliates program to facilitate the transfer of our technologies to the commercial sector.

## **E. Related Work**

Our hybrid approach to QA builds on research in summarization and generation. Previous work in multi-document summarization has developed solutions that identify similarities across documents as the basis for summary content (Mani and Bloedorn, 1997; Carbonell and Goldstein, 1998; Hatzivassiloglou et al., 1999; Barzilay et al., 1999; Radev et al., 2000; Lin and Hovy, 2002a). Whether similarities are included through sentence extraction or information fusion (Barzilay et al., 1999), all of these approaches are data-driven because similarities in the data determine content.

Goal-driven, or top-down, strategies are more often found in generation. Schemas (McKeown, 1985), rhetorical structure theory (Mann and Thompson, 1988; Moore and Paris, 1992; Hovy, 1993; Marcu, 1997) and plan-based approaches (Reiter and Dale, 2000) are examples of goal-driven approaches, where the schema or content plan specifies the kind of information to include in a generated text. In early work, schemas were used to generate definitions (McKeown, 1985), but the information for the definitional text was found in a knowledge base. In contrast to our work on biographical descriptions, previous work used hand-encoded plans (Teich and Bateman, 1994; Kim et al., 2002); we are working on automatic learning of plans. In more recent work, information extraction is used to create a top-down approach to summarization (Radev and McKeown, 1998) by searching for specific types of information which can be extracted from the input texts (e.g., perpetrator in a news article on terrorism). Here, the summary briefs the user on domain-specific information assumed *a priori* to be of interest.

Other long-answer QA systems are currently under development as part of the AQUAINT program (Voorhees, 2003a). Some of these share attributes with DefScriber and ProGenIE. Weischedel et al. explore biographical questions, using a combination of methods that are largely complementary to those used in DefScriber and ProGenIE namely identification of key linguistic constructions and information extraction (IE) to identify specific types of semantic data. Another important contrast between our work and most of the long-answer systems developed under the AQUAINT program has to do with answer format. While these systems mostly produce answers as a ranked list of descriptive phrases or sentences, Columbia's system uses summarization methods to produce a coherent, multi-sentence, paragraph-length response.

Research on responding to opinion and event questions requires analysis of input documents to identify the opinions or events. There are a variety of definitions of events. In the topic detection and tracking (TDT) community (Allan, 2002; Yang et al., 1999), an event is an instantiation of a topic. For example, an *earthquake* is a topic and the *Kobe 1995 earthquake* and the *Afghanistan 1998 earthquake* are two specific events. In TDT, the problem studied is finding texts on a particular event. We are interested, in contrast, on identified sentences and phrases about events in a text. In the information extraction community (Marsh and Perzanowski, 1997), a template, which is filled by extracting information from a text, represents an event. In linguistics (Chung and Timberlake, 1985; Miller, 1995), linguists study verbs or nouns which can refer to events and specify constraints on how they can be used. Our work attempts to build on these three approaches in order to identify participants of events and the text that relates to the same event.

As in our work, several studies have found that lexical features were useful for subjectivity identification, needed for opinions (Hatzivassiloglou and Wiebe, 2000; Riloff et al., 2003). (Turney, 2002) showed that semantically oriented words could be used to label the semantic orientation of phrases. (Pang et al., 2002) adopted a more direct approach, using supervised machine learning with words and n-grams as features to predict orientation at the document level. Earlier work for automated opinion detection discriminated between subjective and objective text at the phrase, sentence and document levels using manually annotated text for training (Hatzivassiloglou and Wiebe, 2000; Wiebe et al., 1999).

QA systems that accept questions beyond the factual type, and where the document collection is extensive or changing, pose enormous challenges for evaluation. Where a target behavior can be well-specified, system evaluation can be achieved by assembling an "answer set." For example, with factual questions, open-domain QA systems have been evaluated against a standard set of answers, using Mean Reciprocal Rank, as in the QA track of TREC (Voorhees, 2003a). Here, the difficulty of defining a standard answer set is addressed by assembling sets of answer patterns. The Bleu approach to evaluating machine translation (Papineni et al., 2002) is conceptually similar. This method depends on quantifying ngram overlap of a translation with a repository of model translations. In both cases, while there are multiple ways of "presenting" the same answer or translation, there is nevertheless explicit human consensus on what constitutes "the answer" (Kupiec, 1993), or "translational equivalence," in the case of MT. Bleu scores, for example, are calibrated to human scores. In the case of summarization, however, it is difficult to achieve consensus either indirectly, by comparing different human summaries to each other, or directly, by eliciting human judgements. In our view, the current work on summarization evaluation either takes the lack of explicit consensus as an appropriate standard to calibrate an evaluation with (Lin and Hovy, 2002b), or attempts to discover an implicit, more fluid consensus based on prioritizing information (Radev et al., 2003). We believe to take the former tack is to give up on teasing apart what causes the instability of human judgements, whereas evidence suggests it is possible to design elicitation methods that achieve more consistency (Radev et al., 2003) (Passonneau and Nenkova, 2003).

Our work on long-answer QA relies on work in robust semantic parsing. Much of the recent work on shallow semantic parsing, including our own, follows the general architecture outlined by Gildea & Jurafsky (Gildea and Jurafsky, 2002). The goal of these systems is to annotate the predicate-argument structures in a sentence. Arguments are given labels that indicate their roles relative to the predicate. An input sentence is first parsed by a syntactic parser. Then, for each predicate (verb), features are extracted for each constituent and used to classify the role label for

the constituent (which can be NULL). The G&J classifier used a backed-off combination of predictor features to estimate posterior role probabilities directly. More recent work has involved adding new features or using a different classification mechanism (or both). (Gildea and Palmer, 2002) report results on the PropBank corpus using essentially the same system as G&J. (Surdeanu et al., 2003) used a decision tree classifier and a set of additional features (such as part-of-speech of headword). (Gildea and Hockenmaier, 2003) used features extracted from a Combinatory Categorical Grammar in an attempt to better model dependencies. (Chen and Rambow, 2003) attempt to capture deeper syntactic and semantic representations using extractions of a Tree Adjoining Grammar (TAG) from the Penn TreeBank. (Fleischman et al., 2003) used a maximum entropy framework for classification.

All of these systems used either the PropBank or FrameNet for training and test data. Currently the configuration that reports the highest performance for both corpora is our AQUAINT Phase I system, which we call ASSERT. This system uses the general G&J architecture with a set of additional features and a Support Vector Machine based classifier.

## **F. Prior Accomplishments**

### ***F.1. TIDES: Multilingual Multidocument Information Tracking and Summarization***

Contract N66001-00-1-8919, \$1,799,851, March 1, 2000 – February 28, 2004, DARPA/ITO

We are developing a practical, multilingual and multidocument summarization system. Our design features the integration of robust, statistical techniques, shallow linguistic approaches and machine learning to achieve scalability within languages and portability across languages. To realize these goals, we have developed methods for summarization across documents using information fusion and identification of key differences, summarization across languages relying on identification and translation of terms, and new methods for identification, expansion and translation of terms. To date, we have developed an English prototype multi-document summarization system, MULTIGEN, which results in a dramatic decrease in amount of text to read. We have developed a second summarization strategy, DEMS, that looks for important new information in a document based on statistical metrics. Both of these approaches have been integrated in Newsblaster (<http://newsblaster.cs.columbia.edu>) to provide daily updates of multiple news sites on the web, including clustering of news articles on the same event.

### ***F.2. DLI2: A Patient Care Digital Library: Personalized Retrieval and Summarization of Image, Video and Language Resources (PERSIVAL)***

Contract IIS-9817434, \$5,002,375, September 1, 1999 – August 31, 2004, NSF

The goal of this project is to provide personalized access to a distributed patient care digital library through the development of a system, PERSIVAL (PERsonalized Retrieval and Summarization of Image, Video And Language resources). Our initial prototype of PERSIVAL tailors search, presentation, and summarization of online medical literature and consumer health information to the end user, whether patient or healthcare provider using the secure online patient records available at Columbia Presbyterian Medical Center (CPMC).

### ***F.3. Monitoring of Online Information Sources***

Contract: IIS-9817434, \$697,000, NSF and NSA, June 1, 2002 – July 31, 2003

We are augmenting our work on Newsblaster towards real-time interaction, to include tracking of events across days and search over summarized news. Our research also focuses on

development of new techniques for summarization of email. This task is difficult because of the informal nature of the language used in email; it contains aspects of dialogue in the frequent use of quoted replies and it often contains ungrammatical and incomplete fragments of sentences.

#### ***F.4. Next Generation Conversational Interfaces***

Contract: N66001-00-2-890602, \$1,968,000, 10/17/00 – 4/3/02, DARPA

This project was funded under the DARPA Communicator program for spoken language systems. The goal of the project was to advance the capabilities of spoken dialog systems for interfacing to computer applications. The project resulted in the development of the CSLR Conversational Agent Toolkit and of two applications, CU Communicator and CU Move. CAT is the set of modules (audio front-end feature extraction, speech decoder, parser, dialogue manager, TTS) used at CSLR to develop conversational (spoken dialogue) applications. CU Communicator is a system implementing the DARPA Communicator Travel Task; obtaining information over the telephone about air travel, hotels and rental cars. We participated in all of the system evaluations sponsored by the program and had one of the top ranked systems. CU Move is an in-vehicle conversational interface for route navigation information.

#### ***F.5. Modeling Pronunciation Variation for Universal Access to Speech Understanding***

Contract: NSF IIS-9978025, \$503,956, 9/15/99 – 8/31/02, NSF

Another research goal of our lab has been to improve the robustness of speech recognizers to pronunciation variation. Pronunciation variation is known to be one of the major sources of errors in speech recognition. Through our NSF project “Modeling Pronunciation Variation for Universal Access to Speech Understanding”, we have focused on understanding the causes and implications of this variation. We first developed analytic techniques for studying how discourse, lexical and syntactic context, and dialect effect word pronunciation variability (Bell et al., 2003; Jurafsky et al., 2002; Jurafsky et al., 2001a; Bell et al., 1999; Jurafsky et al., 1998). We then showed that only certain of these variations (especially syllable deletions) cause error in current ASR technology (Jurafsky et al., 2001b). Finally, we showed how pronunciation variation due to Spanish accent affected the ASR lexicon (Ward et al., 2002) and the difference in performance due to re-training acoustic versus language models on Spanish-accent English training sets (Ikeno et al., 2003).

## **G. Facilities**

### ***G.1. Columbia University Natural Language Processing Laboratory***

The NLP Group has numerous computers purchased and supported by research funds: 1 Sun Ultra 80 server, 4 Sun Ultra 30 servers, several Sun Ultra 20 servers, 1 Sun Blade server, a Terabyte PC Linux-based fileserver, 4 high-end PC Linux servers, and a number of Unix-based (Sun Ultra 10 and PC Linux) and Microsoft Windows lab workstations. All of these machines are connected to the departmental ethernet. In addition, all the group's 17 Ph.D. students and 5 research staff persons have a workstation on his or her desk.

Another important asset of the group is its sophisticated set of software tools. Many tools have been obtained from external sources: Church's Part-of-Speech tagger from AT&T; Collins' parser, a robust statistical parser from AT&T; the Alembic Workbench from MITRE; CLASSIC (an implementation of KL-ONE); LFG Grammar Writer's Workbench from Xerox; PC-KIMMO from the CLR; WordNet from Princeton University; FrameNet from ICSI; and Identifinder from

BBN. The group's locally developed tools include: FUF, the Functional Unification Formalism; CFUF, a graph-based implementation of the FUF language implemented in C and embedded within a Scheme interpreter; Surge, a syntactic realization grammar for text generation; Crep, a regular expression matcher for corpus retrieval; Segmenter, a text segmentation utility; Verber, a utility design to conflate semantically related verbs together; Xtract, an automatic collocation compiler; LinkIT, a tool for identifying and relating noun phrases within a document; Centrifuser, a domain- and genre-specific multidocument summarization system; SimFinder, identifies spans of texts that convey similar meaning; MultiGen, a multi-document text summarizer; DEMS, the Dissimilarity Engine for Multidocument Summarization; DEFINDER, a text-mining tool for extracting definitions from medical text; and DefScriber, a definitional question-answering system.

### ***G.2. University of Colorado Center for Spoken Language Research***

The University of Colorado provides exceptional facilities to CSLR for both research and education. CSLR is housed in 6,000 sq. feet of laboratory space, including 23 single occupancy rooms, 10 double occupancy rooms, and equipment room, a room for demonstrations and meetings, and two classrooms for computer laboratory classes. In addition, CSLR faculty hold appointments in academic departments (Computer Science, Psychology, Linguistics, Electrical Engineering, and Speech Hearing and Language Sciences), and have access to the research and educational facilities of these departments, as well as the campus computing network, libraries, and all other available campus resources. CSLR supports both Windows and Unix operating systems, and provides powerful desktop computers for individual researchers, and powerful data and computer servers for developing and evaluating systems for collecting data.

The principal investigators also have access to the talented and dedicated faculty, staff and students associated with CSLR. The Center currently has six faculty, five post-doctoral researchers, ten full time staff, three part-time staff, and numerous graduate students. The CSLR staff administers grants and budgets, assist with report preparation, and provide general operational support to CSLR faculty, staff, and graduate students.

## **H. Use of Government Property**

No use of government property is required for execution of the proposal. We will use data provided by the government (e.g., CNS data). While it is desirable for us to have access to such data, it is not required for completion of the research. We will only use such data after signing appropriate release forms.

## **I. Support and Teams**

The leaders of the Columbia team include Drs. Hatzivassiloglou, McKeown, and Passonneau; the team from Colorado consists of Drs. Jurafsky, Ward, and Martin. The Columbia team members are world leaders in the area of natural language processing, with particular expertise in the areas of statistical natural language processing, natural language generation, summarization, and digital libraries. Ongoing support to the Columbia group comes from NSF, KDD, ARDA, DARPA, and NIH. The Colorado team members are world leaders in the area of natural language processing with particular expertise in the areas of spoken dialog systems, semantic analysis, and statistical language processing. The primary recent sources of support from all three have been from the NSF, KDD and ARDA. Drs. McKeown, Jurafsky, and Martin hold regular academic appointments and are committing 10% of their AY time along with one-month

summer commitments. Dr. Martin is committing additional time during the 2004-05 academic year as part of a sabbatical. Drs. Hatzivassiloglou, Ward, and Passonneau are research faculty and are committing 40%, 25%, and 20% of their time respectively.

The PIs from each institution will have the primary responsibility for supervising the work of the postdoctoral researchers, graduate students and programmers at their institutions. Extensive collaborations between the groups have been established and will continue. To facilitate these collaborations, group members have traveled between the institutions and one large joint retreat was held in Colorado during the summer of 2003. Current and continuing joint work by Drs. Jurafsky, Hatzivassiloglou, McKeown and Yu includes the integration of the semantic parser into the processing of opinion texts and biographies. Future joint work includes the use of the semantic parser for event processing. The primary personnel leading this effort will be Drs. Hatzivassiloglou, Martin, and Ward.

## J. Intellectual Property Rights

The Offeror (Columbia University) will retain title to any algorithms, software, data, and documentation developed by the Offeror under the proposed contract. However, the Offeror hereby grants (a) licenses for the use of such algorithms, software, data, and documentation for non-commercial purposes to all other participants/contractors in the AQUAINT program, including a system integrator or independent evaluator that may be selected by ARDA according to the provisions of Section 5.4 of BAA 03-06-FH; and (b) an irrevocable license to full rights on such intellectual property to the United States Government.

## K. Evaluation

**Program-Level Evaluations:** We will participate in at least two Program-Level evaluations each year. Further, we believe progress depends on a flexible approach to evaluation metrics that attempts to quantify multiple dimensions of system performance. For example, evaluation of **(1) presentation of divergent opinions on the same topic**, should address the elements involved in constructing opinion answers, such as scoring of ability to differentiate documents and sentences expressing (a high degree of) opinions from those that do not, and identification of the opinion, whether the opinion is pro or con, and the opinion holder. For each element, recall and precision could be computed against a test set. For **(2) generation of a time line for a complex event**, depending on the length and constituency of the time line, a mixed set of metrics seems appropriate in order to separate identification of an event from ordering. For the former, human assessors may be needed to determine whether the system output corresponds to a specific event, e.g., the August 14, 2003 power blackout, which can be referred to in multiple ways. For evaluation of ordering of recognized events on a time line, a metric such as string edit distance or an alignment function might be appropriate. For **(3) fusion of answers from multiple documents**, if answers are scored for content, the metric should attempt to address the differential importance of different facts, as is being done with certain approaches to summarization evaluation (Passonneau and Nenkova, 2003; Radev et al., 2000). In addition, a user-centered evaluation design is also appropriate that would allow users to rate relevance, fluency, and so on, as in (Blair-Goldensohn et al., 2003). For **(4) assembly of a biography**, again, evaluation should address informational and user-centered dimensions separately, such as identification of key achievements, roles, and dates on the one hand, and on the other, fluency or organization of the response.

**Project-Level Evaluations:** We will perform a minimum of four project level evaluations to quantify progress in system components. Distinct components require distinct evaluation methodologies and metrics as noted below. These evaluations will include evaluations of *intermediate data*, such as biographical content plans, definitional predicates, event labels, opinion polarity; evaluations of *system components*, e.g., of algorithms for filtering text, extracting text strings, clustering textual or semantic data, information fusion; and *user-centered evaluations* of answers.

**Semantic Parsing:** In order to evaluate semantic parsing performance, we have been using the standard training and test sets from the PropBank and FrameNet corpora. These test sets have been annotated by hand. For evaluation, the test set is processed by the system and precision and recall statistics are generated for annotating predicate argument structure. We are also hand annotating test sets from the TREC corpus to allow us to measure performance on new genre of text.

**Audio Document Transcription:** For evaluation, we will obtain audio documents from the web containing short clips of current new items, transcribe them by hand and annotate them with semantic labels. We will then automatically transcribe the documents with Sonic and use word error rate for recognizer evaluation. Using the hand annotated transcription as the standard, we will compute precision and recall performance for semantic parsing of the speech recognizer output. We will also generate a set of questions to compare the end-to-end performance on the manual versus automatic transcriptions.

**Definitional Answers:** We apply machine learning with cross-validation to the problem of learning rules for definitional predicates that allow us to extract distinct types of information so as to control how we structure a definitional answer. We have achieved accuracy rates of 80% and higher. We use human judges to evaluate automatically generated definitions we produced under distinct system parameter settings to test the relative contribution of statistical and top-down methods. In our human questionnaires, we elicit five dimensions of structure, redundancy, term understanding, relevance and coverage.

**Biographical Answers:** We have developed a novel approach to generating and evaluating content plans using genetic programming techniques. For evaluating order constraints learned from annotated corpora, we generate text and align it with human-generated texts. The final score is an average of the alignment scores produced for a set of semantic inputs against a set of human texts. In the past we have applied this evaluation method to plans for medical briefings (Duboue and McKeown, 2002) and will adapt the method for evaluation of biographical content plans.

**Event Tracking:** In our approach to atomic event detection in which we extract binary links among potential event participants from newswire, we assemble constellations of inter-participant links into atomic events ranked by importance, and assign event labels. Our evaluations include qualitative comparison with IE systems on MUC Scenario templates (DARPA, 1997), and precision scores of the link, importance and label output of our algorithm with human scores (Filatova and Hatzivassiloglou, 2003). As we apply the event tracking algorithm to generation of system answers, we will evaluate using a mix of intrinsic and user-centered evaluations.

**Opinions:** For our opinion classifiers at the document and sentence level, we measure performance by standard recall and precision, whereas we evaluate our automatic identification of opinion words and labels against an external established gold standard and quantify results in terms of accuracy. For future work on identifying opinion labels, opinion holders and targets, we

will evaluate against corpora that have been manually and automatically tagged with semantic features and roles from FrameNet, WordNet, PropBank and the semantic parser.

## L. Data

Here we list the Data Needs (**DN**), and Data Collection/Creation Efforts (**DCCE**) supporting the investigation of mixed knowledge intensive and statistical methods for answer generation.

**Definitional answers. DN:** 1. Manual annotation of rhetorical predicates for definitional answers to provide training data. 2. Questionnaire data from human subjects on answer quality. **DCCE:** Based on previous results with annotation of small set of documents, creation of coding manual, and document markup language, we need annotators to apply existing annotation method on new texts; we need annotators for analysis of new predicates.

**Biographical answers. DN:** Match biographical facts about known individuals (currently 600 facts for 1K individuals) against biographical texts to evaluate content selection. **DCCE:** 1. Automatic methods to link texts and semantic facts into a Text and Knowledge Resource (TKR); currently done for 300 individuals. 2. Human annotations of a subset of TKR to locate and tag verbalizations of semantic facts; pilot study demonstrated feasibility of annotation task.

**Event detection. DN:** Texts labeled with event information for training and testing statistical methods for extraction of events, their importance, and their participants. **DCCE:** Tagging Topic Detection and Tracking Corpus (TDT2) with hand-labels for semantic roles, atomic events, and macro events, as well as automatic labels for Named Entity and POS. Human annotation and questionnaire needs include evaluation of dynamic event labeling.

**Opinions. DN:** 1. Texts labeled with lexical, sentential, semantic and text-level information as training data for automatically identifying opinion sentences, propositional opinions, semantic roles, and answer structure. 2. Human annotations and/or questionnaire data to extend training data and perform evaluations. **DCCE:** Texts annotated manually with multiple levels of opinion classification: document level, opinion sentence semantics (including propositional opinions and the identity of the opinion-holder); opinion answer structure, as well as automatic annotation with semantic role values. Questionnaire data; human generated answers.

**Answer Evaluation Methodology. DN:** Texts with human-generated summaries and answers and linguistically annotated for training automated evaluation methods; human subject responses to source texts and answer texts. **DCCE:** Human-generated answers to definitional (or other) questions given a set of source texts; human-reader responses to source texts and text extracts and to answers; manual annotation of content units and relevance relations in answers and document sources; manual, automatic annotations of semantics and pragmatics of content units.

**Multilingual (Arabic and Chinese) Question Answering. DN:** Training data for various components of our Arabic and Chinese systems. **DCCE:** Manual annotation of Penn Arabic Treebank with semantic role values. Collection of Chinese and Arabic database of questions, and manual annotation of questions with question-types.

## M. Testbed Participation

We will provide a published interface to our full system for use in the AQUAINT testbed. We will also provide an interface to components, including each question type (definitions, opinions, biographies, and events), the semantic parser, and the speech interface. We have adopted a modular design based on client-server and server-server interactions, allowing modules to operate in parallel and on different machines. Two main servers coordinate question analysis and

answer generation. The communication between servers and clients is handled via HTTP, allowing distributed operation of the system with some modules running at Columbia and others at Colorado. We have emphasized standardized services with well-specified APIs, facilitating the addition of new modules (e.g., for new question types) without other modifications to the core system. We have also built a web-based client for entering questions and displaying results remotely on any standard browser. In addition to access to components, we will also provide any data that we collect for use in the testbed.

We will provide each answer component to other AQUAINT participants when evaluation indicates that it is of high enough quality to be useful. We expect to be able to provide versions of the definition and opinion components in the first six months of the project, versions of the biography component and the semantic parser at the end of the first year, and versions of the event component and speech interface within 18 months.