Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)

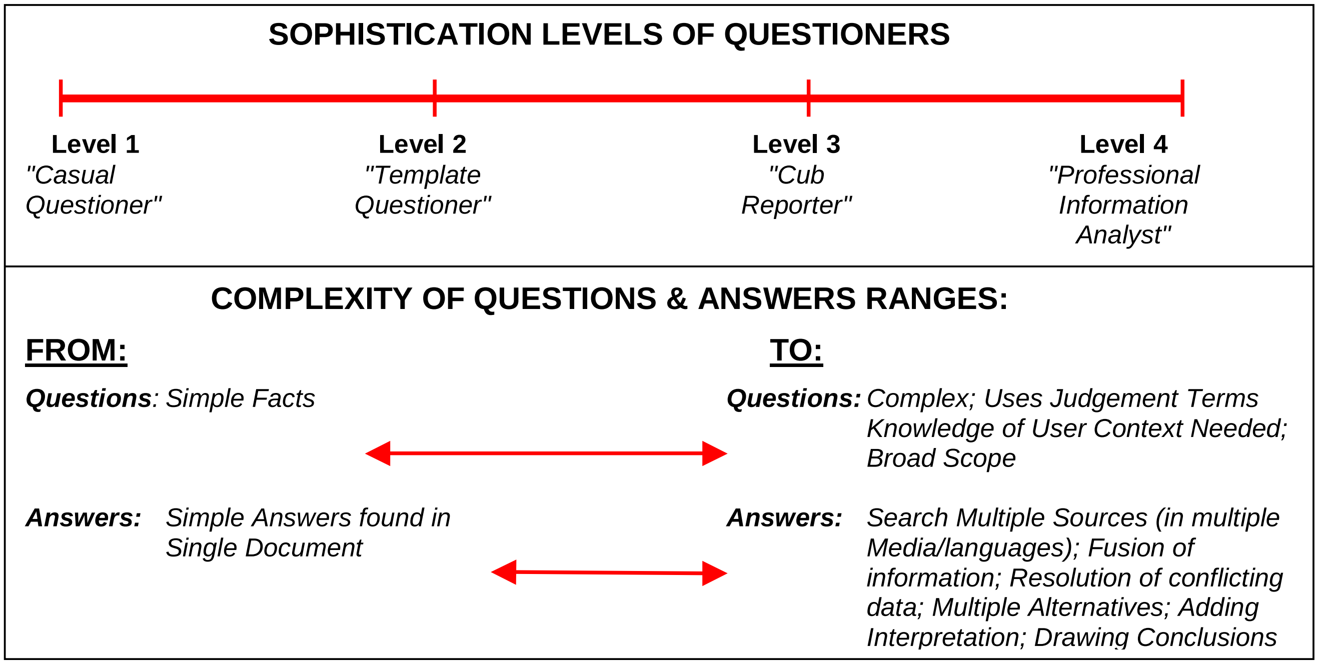
John Burger[1], Claire Cardie[2], Vinay Chaudhri[3], Robert Gaizauskas[4], Sanda Harabagiu[5], David Israel[6], Christian Jacquemin[7], Chin-Yew Lin[8], Steve Maiorano[9], George Miller[10], Dan Moldovan[11], Bill Ogden[12], John Prager[13], Ellen Riloff[14], Amit Singhal[15], Rohini Shrihari[16], Tomek Strzalkowski[16], Ellen Voorhees[18], Ralph Weishedel[19]

1. **INTRODUCTION**

Recently the Vision Statement to Guide Research in Question Answering (Q&A) and Text Summarization outlined a deliberately ambitious vision for research in Q&A. This vision is a challenge to the Roadmap Committee to define the program structures capable of addressing the question processing and answer extraction subtasks and combine them in increasingly sophisticated ways, such that the vision for research is made possible.

[1] MITRE, john@mitre.org
[2] Cornell University, cardie@cs.cornell.edu
[3] SRI International, vinay@ai.sri.com
[4] University of Sheffield, robertg@dcs.shef.ac.uk
[5] Southern Methodist University, sanda@seas.smu.edu
[6] SRI International, israel@ai.sri.com
[7] LIMSI-CNRS, jacquemin@limsi.fr
[8] ISI/USC, cyl@isi.edu
[9] Advanced Analytic Tools, stevejm@ucia.gov
[10] Princeton University, geo@clarity.princeton.edu
[11] Southern Methodist University, moldovan@seas.smu.edu
[12] New Mexico State University, ogden@crl.nmsu.edu
[13] IMB T.J. Watson Research Center, john@us.imb.com
[14] University of Utah, riloff@cs.utah.edu
[15] AT&T, singhal@research.att.com
[16] State University of New York at Buffalo, rohini@cedar.Buffalo.EDU
[16] University of Albany, SUNY, tomek@cs.albany.edu
[18] NIST, ellen.voorhees@nist.gov
[19] BBN, weischedel@bbn.com

The Vision Statement indicates a broad spectrum of questioners and a range of answers, as illustrated in the following chart:



SOPHISTICATION LEVELS OF QUESTIONERS

Level 1 *"Casual Questioner"*

Level 2 *"Template Questioner"*

Level 3 *"Cub Reporter"*

Level 4 *"Professional Information Analyst"*

COMPLEXITY OF QUESTIONS & ANSWERS RANGES:

FROM:

***Questions**: Simple Facts*

***Answers:** Simple Answers found in Single Document*

TO:

***Questions:** Complex; Uses Judgement Terms Knowledge of User Context Needed; Broad Scope*

***Answers:** Search Multiple Sources (in multiple Media/languages); Fusion of information; Resolution of conflicting data; Multiple Alternatives; Adding Interpretation; Drawing Conclusions*

The Q&A Roadmap Committee has the ultimate goal to provide a research roadmap that enables meaningful and useful capabilities to the high-end questioner. Thus the research roadmap must have several milestones set, to check intermediary goals in the capabilities offered to the full spectrum of questioners and in the answer range that is offered. Real-world users of Q&A systems find such tools useful if the following standards are provided:

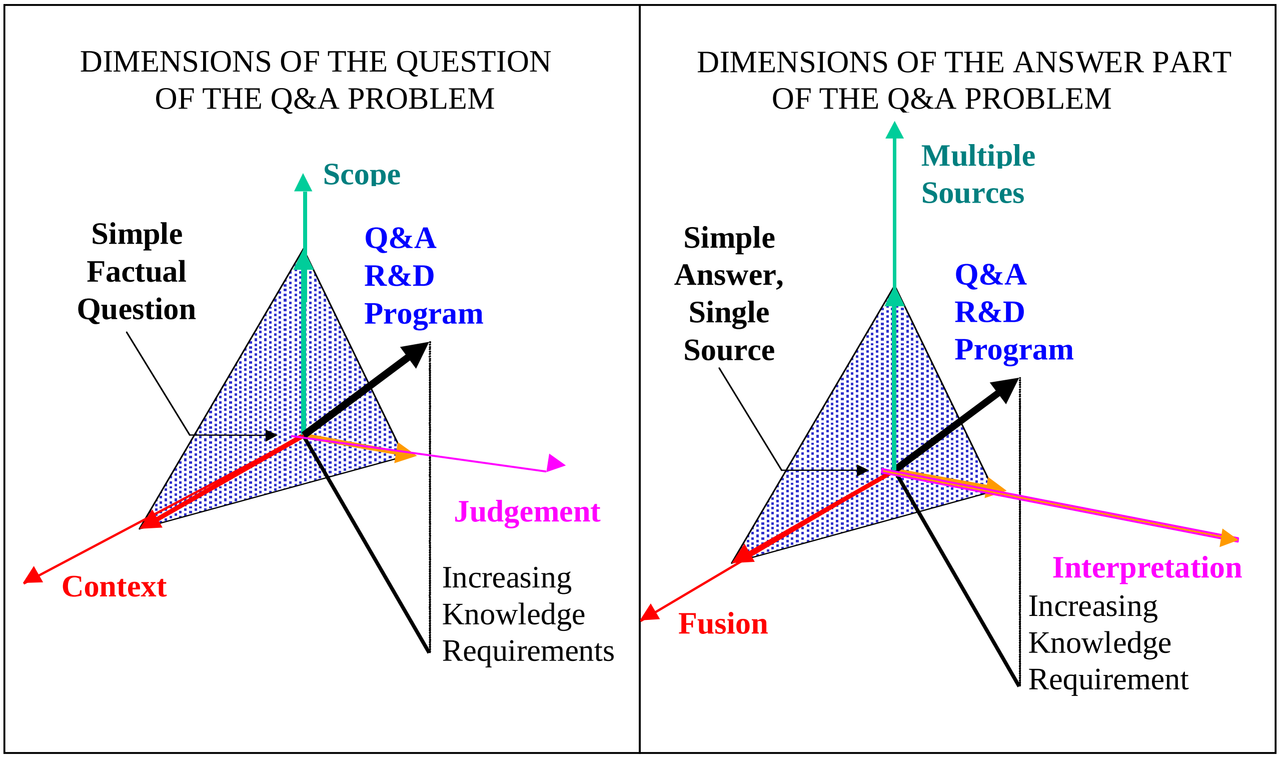
- Timeliness. The answer to a question must be provided in real-time, even when the Q&A system is accessed by thousands of users. New data sources must be incorporated in the Q&A systems as soon as they become available, offering the user an answer even when the question refers to the most recent events or facts.
- Accuracy. The precision of Q&A systems is extremely important – as incorrect answers are worse than no answers. Research in Q&A should focus on ways of evaluating the correctness of provided answers, that comprise also methods of precisely detecting cases when the available data does not contain the answer. Contradictions in the data sources must be discovered and conflicting information must be dealt with in a consistent way. To be accurate, a Q&A system must incorporate world knowledge and mechanisms that mimic common sense inference.

- Usability. Often, knowledge in a Q&A system must be tailored to the specific needs of a user. Special domain ontologies and domain-specific procedural knowledge must be incorporated. Rapid prototyping of domain-specific knowledge and its incorporation in the open-domain ontologies is very important. Often, heterogeneous data sources are used – information may be available in texts, in databases, in video clips or other media. A Q&A system must be able to mine answers regardless of the data source format, and must deliver the answer in any format desired by the user. Moreover it must allow the user to describe the context of the question, and must provide with explanatory knowledge and ways of visualizing and navigating it.

- Completeness. Complete answers to a user's question is desirable. Sometimes answers are distributed across one document or even along multiple documents in the data sources. Answer fusion in a coherent information is required. The generation of the complete answer must rely on implicatures, due to the economic way in which people express themselves and due to the data sparseness. Moreover, world knowledge together with domain-specific knowledge must be combined and reasoned with, sometimes in complicated ways. A Q&A system must incorporate capabilities of reasoning and using high performing knowledge bases. Sometimes analogies to other questions are necessary, and their judgement must be done either in the context defined by the user or in the context of the user's profile. The automatic acquisition of user profiles is a method of enabling collaborative Q&A and of acquiring feedback information regarding Q&A.

- Relevance. The answer to a user's question must be relevant within a specific context. Often the case, interactive Q/A, in which a sequence of questions helps clarify an information need, may be necessary. Question complexity and the related taxonomy of questions cannot be studied without taking into account the representation of context, the common ground between the user and the Q&A system and without allowing for follow-up questions. The evaluation of Q&A system must be user-centered: humans are the ultimate judges of the usefulness and relevance of Q&A systems and of the ease with which they can be used.

To achieve these desiderata, the Vision Statement proposes to move research in Q/A along the six directions indicated in the following two diagrams:


DIMENSIONS OF THE QUESTION OF THE Q&A PROBLEM
Scope
Simple Factual Question
Q&A R&D Program
Judgement
Context
Increasing Knowledge Requirements
DIMENSIONS OF THE ANSWER PART OF THE Q&A PROBLEM
Multiple Sources
Simple Answer, Single Source
Q&A R&D Program
Interpretation
Fusion
Increasing Knowledge Requirement

The Roadmap Committee's role is to consider how far away from the origin along the six axes we should move the R&D plane and how rapidly we believe technological solutions can be discovered along such a path.

2. ISSUES IN Q&A RESEARCH

Research in the area of Open-Domain Question Answering generates a lot of interest both from the NLP community and from the end-users of this technology, either lay users or professional information analysts. Open-Domain Question Answering is a complex task, that needs a formal theory and well-defined evaluation methods. The theory of Q&A does not appear in a vacuum – several theories have been developed earlier in the context of NLP or cognitive sciences. First, we have the conceptual theory of question answering, proposed by Wendy Lehnert, with an associated question taxonomy and then we have the mechanisms for generating questions developed by Graesser & al. However, these theories are not open-ended. They did not assume large scale real-world resources, and were not using high-performance parsers, named entity recognizers or information extractors, tools mainly developed in the last decade, under the impetus of the TIPSTER program. Nevertheless, these former theories of Q&A relied on complex semantic information, that needs to be reconsidered and remodeled for the new broader task of Q&A. If in the 90s semantics was put on the back burner, as the Vision Statement acknowledges, it is in the interest of Q&A Research to revitalize research in NLP semantics, such that we can better understand questions, the context in which they are posed, and deliver and justify answers in contexts.

In moving along the six degrees of freedom on the Q&A Research space, the Roadmap Committee has identified a number of research issues, that are further decomposed into a series of tasks and subtasks, useful to the definition of the Roadmap Research program structures. The issues are:

1. *Question Classes*: Need for question taxonomies All previous theories of Q&A contain special taxonomies of questions. QUALM, the system developed by Wendy Lehnert is based on thirteen conceptual categories in which questions can be mapped by an inferential analysis procedure. The taxonomy proposed by Lehnert is primarily based on a theory of memory representation called *Conceptual Dependency*. In contrast, the taxonomy proposed by Graesser has foundations both in theory and in empirical research. Two theories provided most of the categories in Graesser's taxonomy. The speech act theory based on quantitative research on interpersonal behavior developed by D'Andrade and Wish identifies eight major speech act theories that can be used to categorize virtually all speech acts in conversations: questions (equivalent to interrogative), assertion, request/directive, reaction, expressive evaluation, commitment and declaration. These eight categories were abstracted from act theories in philosophy, linguistics and sociology (Austin ,1962; Labov & Fanshel, 1977; Searle, 1969). The question taxonomy proposed by Graesser & al. includes questions, assertions and request/directives because they were the only categories that provide genuine inquiries. Research needs to be done to expand and consolidate these categories for the larger scope of open-domain question answering.

   The taxonomy of questions proposed by Graesser & al. comprises eighteen different question categories that provide genuine inquiries. For the major category of "questions", Graesser and his collaborators have used QUALM's thirteen categories to which they have added several new categories: a "comparison" category (which was investigated by Lauer and Peacock, 1990), a "definition" category, an "example" category, and an "interpretation" category. For all the categories in the taxonomy, Graesser conducted a study of empirical completeness, showing that the taxonomy is able to accommodate virtually all inquiries that occur in a discourse. The study focused on three different contexts: 1) college students reading passages, 2) individuals using a computer, and 3) citizens asking questions in newspaper media. The study sessions spanned a variety of topics, including basic mathematics, statistics, research methods, a computer network, climate, agricultural products, and population density. However, Graesser's taxonomy was not implemented in a Q&A system, and was used only by humans to score the reliability of the taxonomy itself. It is clear that it is not open-ended and it has severe limitations, based on the ad-literam incorporation of the QUALM question categories – requiring processing that cannot scale up to large collections of texts. It is time to go back to the future!

The question taxonomy proposed in the TREC-8 Lasso paper describes a classification of questions that combines information from the question stems, questions focus and phrasal heads. This classification is simplistic, and needs to be extended along several dimensions. Nevertheless, it is a good start, since it does not require question processing based on a specific knowledge representation and is clearly open-ended in nature. Moreover, it unifies the question class with the answer type via the question focus.

As simple as the question classification used in Lasso was, it needs extensions and clarifications. The notion of *question focus* was first introduced by Wendy Lehnert in her book "*The Process of Question Answering*". In this book, at page 6, section 1.1-7 the focus of a question is defined as the question concept that embodies the information expectations expressed by the question. Because of that, Lehnert claims that some questions are not fully understood until their focus is determined. This intuition was clearly supported by the question classification employed in Lasso. However, many more nuances of the interpretation of a question focus need to be taken into account. For example, Lehnert exemplifies the role of the question focus with the inquiry:

> *Q: Why did John roller-skate to McDonald's last night?*

Interestingly, Lehnert points out that if someone would have produced the answer:

> *A: Because he was hungry.*

the questioner might not have been satisfied, as chances are that (s)he really wanted to know:

> *Q: Why did John roller-skate instead of walk or drive or use some other reasonable means of transportation?*

In this case it is clear that the question asked about the act of roller-skating, not the destination. Therefore, the classification of questions based on their focus cannot be performed unless world knowledge and commonsense reasoning capabilities are added to Q&A systems.

In addition, world knowledge interacts with profiling information. As Wendy Lehnert states, for most adults, going roller-skating is more unusual than going to McDonald's; and any unusual fact or situation requires explanation, thus may become the focus of a question. However, if everyone knew that John was an eccentric health-food nut who roller-skates everywhere he goes, the question

> *Q: Why did John roller-skate to McDonald's?*

would reasonably be interpreted as asking about McDonald's or activities taking place at McDonald's and involving John, rather than roller-skating. Clearly, there is a shift in the interpretation of the focus based on the available world knowledge, the information about the question concepts and their interaction. Furthermore, the recognition of the question focus has different degrees of difficulty, depending on the four levels of sophistication of the questioners. The following table illustrates questions and their focus for all the four levels:

Level 1 *“Casual Questioner”*	**Q:** Why did Elian Gonzales leave the U.S.?	**Focus:** the departure of Elian Gonzales.
Level 2 *“Template Questioner”*	**Q:** What was the position of the U.S. Government regarding the immigration of Elian Gonzales in the U.S.?	**Focus:** set of templates that are generated to extract information about (1) INS statements and actions regarding the immigration of Elian Gonzales; (2) the actions and statements of the Attorney General with respect to the immigration of Elian Gonzales; (3) actions and statements of other members of the administration regarding the immigration of Elian Gonzales; etc
Level 3 *“Cub reporter”*	**Q:** How did Elian Gonzales come to be considered for immigration in the U.S.? --*translated into a set of simpler questions*: Q1: How did Elian Gonzales enter the U.S.? Q2: What is the nationality of Elian Gonzales? Q3: How old is Elian Gonzales? Q4: What are the provisions in the Immigration Law for Cuban refugees? Q5: Does Elian Gonzales have any immediate relatives?	**Focus:** composed of the question foci of all the simpler questions in which the original question is translated. **Focus Q1:** the arrival of Elian Gonzales in the U.S. **Focus Q2:** the nationality of Elian Gonzales. **Focus Q3:** the age of Elian Gonzales. **Focus Q4:** immigration law. **Focus Q5:** immediate relatives of Elian Gonzales.
Level 4 *“Professional Information Analyst”*	**Q:** What was the reaction of the Cuban community in the U.S. to the decision regarding Elian Gonzales?	**Focus:** every action and statement, present or future, taken by any American-Cuban, and especially by Cuban anti-Castro leaders, related to the presence and departure of Elian Gonzales from the U.S. Any action, statements or plans involving Elian’s Miami relatives or their lawyers.

As Wendy Lehnert states in her book “The difficulties involved in natural language question answering are not obvious. People are largely unconscious of the cognitive processes involved in answering a question, and are consequently insensitive to the complexities of these processes”. What is difficult about answering questions is the fact that before a question can be answered, it must be first understood. One level of the interpretation process is the classification of questions. This classification should be determined by well defined principles.

Related subtasks:

1/ Identify criteria along which question taxonomies should be formed.

2/ Correlate question classes with question complexity. Study the complexity of each class of question. For example, start with the study of the complexity of all trivia-like, factual-based questions and the question processing involved as well as the answer extraction mechanisms. Moreover, for each level of sophistication, determine all the question classes and their classification criteria.

3/ Identify criteria marking the complexity of a question.

4/ Study models of question processing based on ontologies and knowledge bases. These models should cover the gap between current question processing of factual data (with emphasis on Named Entity Recognition) and question processing imposed by complex domains (e.g. similar to those developed in the Crisis Management Task in the HPKB program).

2. *Question Processing*: Understanding, Ambiguities, Implicatures and Reformulations. The same information request can be expressed in various ways – some interrogative, some assertive. A semantic model of question understanding and processing is needed, one that would recognize equivalent questions, regardless of the speech act or of the words, syntactic inter-relations or idiomatic forms. This model would enable the translation of a complex question into a series of simpler questions, would identify ambiguities and treat them in context or by interactive clarification.

Question processing must allow for follow-up questions and furthermore, for dialogues, in which the user and the system interact in a sequence of questions and answers, forming a common ground of beliefs, intentions and understanding. New models of dialogue need to be developed, with well formulated semantics, that allow for open-domain NLP processing.

A special case of questions are the inquiries that require implicatures. A class of such questions is represented by questions involving comparatives (e.g. "*Which is the largest city in Europe? "* or superlatives *"What French cities are larger than Bordeaux ?"*) The resolution of such questions requires the ability to generate scalar implicatures. For example, the answer to the first question can be detected even when in the text the information is not explicitly stated. If in a document we find "with its 3 million 437 thousand inhabitants, London is the second-largest capital in Europe" whereas in another document we find "Paris, larger in both size and population than London, continues its booming trend", scalar implicatures, based on the pragmatics of ordinals, infer the answer that Paris is the largest city in Europe. Similarly, the answer to the second question is a list of French cities that surpass Bordeaux in size and/or population. Moreover, both questions display "comparative ambiguity", in the sense that a city may be larger than another for multiple reasons: the number of inhabitants, the territorial size or the number of businesses.

The implicatures needed by question processing are not limited only to scalar implicatures – sometimes more complicated inferences- that are based on pragmatics are needed. A classical example was introduced in Wendy Lehnert's paper:

Q: Do you have a light ?

This is a request deserving a performative action – giving light to someone – to light up her cigarette! Only a full understanding of this question would enable the performative action. An incomplete understanding of the question could generate an answer of the type:

A: Yes, I just got a new lighter yesterday.

resulting from the incorrect translation of question Q into question Q':

Q': Do you have in your immediate possession an object capable of producing a flame?

Question processing must incorporate also the process of translating a question into a set of equivalent questions, which is a classification process based on very sophisticated NLP techniques and a well defined semantics of questions and full-fledges question taxonomies.

The same information request may be formulated through different speech acts and different natural language expressions. A high-performance Q&A system should be able to extract the same answer for any equivalent question posed. Research in question reformulation and generation will move current state-of-the art question processing beyond shallow processing of questions and keyword matching.

There are many forms of question ambiguities. Related to comparative questions mentioned previously is the question "How large is Paris ?". Answers informing about the population size of Paris, about its area size are relevant. However, answers informing about its number of museums or public gardens are not. Therefore a study of question ambiguities and their possible translation into more precise questions is needed. Sometimes these ambiguities can be resolved by world knowledge – e.g. the size of a city is generally measured by the number of inhabitants and by its territorial size. Other times ambiguities are resolved by context or by domain-specific knowledge. Knowledge of History of the Antiquity would enable the answer to the *question "Which is the greatest city in Magna Grecia ?*" (answer: Syracuse) or its follow-up question*: "In what country is it located ?*" (answer: Italy (Sicily))

Ambiguities are present at different levels of question processing. Sometimes it is difficult to establish the question focus since it is ambiguous, other times the question type is not unique and furthermore, the answer type is not always clearly determined, again because of the many forms of ambiguity. For example, for the question:

Q: How rich is Bill Gates?

although the focus can be recognized as the wealth of Bill Gates, the answer is difficult to identify, as wealth can be measured in different ways (e.g. Bill Gates owns 43% of Microsoft, which is a very profitable large company, thus he is very rich with respect to Joe Doe, the owner of 93% of an obscure software company).

Linguistic ambiguity interacts with linguistic diversity. Various expressions of the same concept encountered in a question need to be conflated correctly, thus resolving most of the ambiguities. Ambiguities range from various spellings to the presence of vocabulary terms that have different meanings in different domains (words sense ambiguity), and thus classify differently the question. Furthermore, not all ambiguities can be solved by specifying the question context. The following table illustrates several forms of ambiguities associated with the four levels of questioner sophistication:

Level 1 *"Casual Questioner"*	**Q:** Where is Pfizer Pharmaceutical doing business?	**Ambiguity**: Location Granularity: Country vs. State vs. City
Level 2 *"Template Questioner"*	**Q:** What recent drugs has Pfizer introduced on the market?	**Ambiguity**: Drugs developed by Pfizer or marketed by Pfizer.
Level 3 *"Cub reporter"*	**Q:** When was Viagra introduced on the market? **Q:** Was Viagra a successful drug?	**Ambiguity**: market location: Drug introduction in the U.S., in Europe, in Mexico.
Level 4 *"Professional Information Analyst"*	**Q:** Why are pharmaceutical companies using Oracle Clinical Application?	**Ambiguity**: -intend in the usage of software package – impact on the customer satisfaction vs. impact on productivity. - What pharmaceutical companies? Are any other besides Pfizer using Oracle Clinical Application?

Related subtasks:

1/ Develop theoretical models for question processing.

2/ Study semantic models for question similarity, question implications and question subsumption.

3/ Study models of question implicatures.

4/ Learn classification schemes of questions and their translation into more precise inquiries.

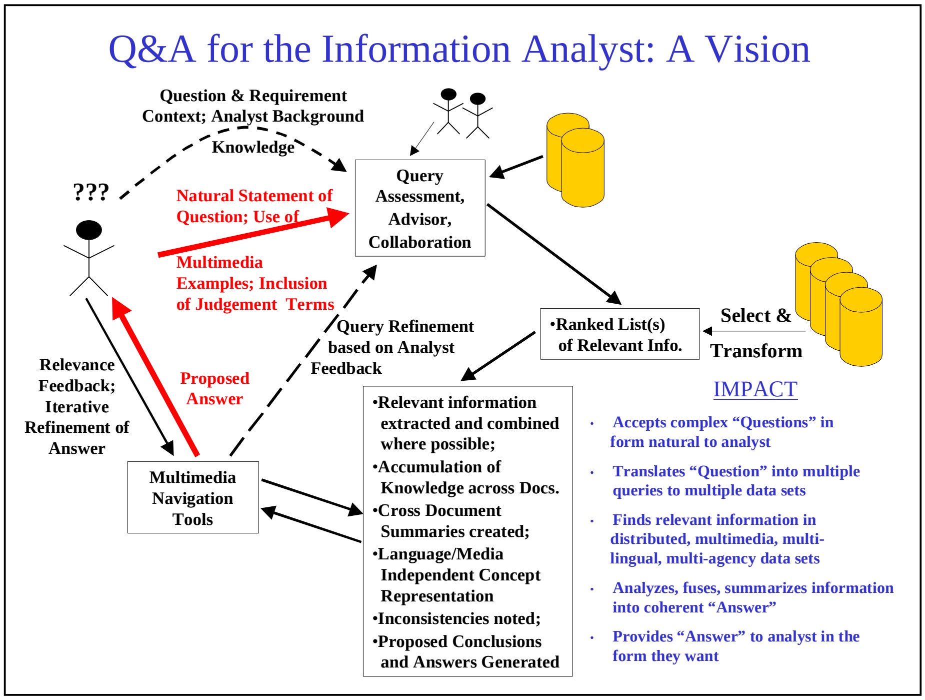
5/ Study models of question ambiguities for each class of question and for various degrees of complexity. Base the study on various ontologies and knowledge bases.

3. *Context and Q&A*. Questions are usually asked within a context and answers are provided within that specific context. Moreover, ways of defining, visualizing and navigating contexts must be studied and implemented. The context can be used to clarify a question, resolve ambiguities or keep track of an investigation performed through a series of questions. Context processing is also very important for collaborative Q&A – in which different software agents processing Q&A tasks collaborate to resolve different questions. A software agent trying to resolve the *question "What happened to Bill Gates ?*" needs to collaborate with another software agent resolving the question "*How much did the Microsoft stock dropped last week ?*"

The notion of context is very complex – and modeling context is not simple. A formal theory of the logic of contextual objects has been proposed by John McCarthy. Revisiting this theory and creating models for Q&A is necessary. Probably the most challenging definition of the *context structures for Q&A* is provided by the scenario when an intelligent analyst, while reading reports produced by two different analysts from different agencies has the "ah hah" experience and uncovers evidence of strong connections between two previously unconnected groups of people or events that are tracked separately. To pose questions like "*Is there any other evidence of connections, communication or contact between these two suspected terrorist groups and its known members?*" the context of the activity of the terrorist groups and evidence of their connection needs to be provided.

The context model should support not only information about the terrorist activities but also pragmatic knowledge about terrorism in general as well as mechanisms for reasoning by analogy with other terrorist groups for which there is evidence that they had joint operations. Since information about the two terrorist groups as well as information about collaborating terrorists may be needed to be fused from different sources, there is strong likelihood that there will be conflicting facts, duplicate information and even incorrect data. Therefore, before being integrated in the context of a question, this data needs to be evaluated for reliability and confidence.

However, the context data interacts with the background knowledge of the analyst. The contextual information needs to be (1) visualized through multimedia navigation tools; (2) modified by the analyst to better reflect her/his background knowledge; and (3) evaluated by the human judgement of the analyst, as indicated in the following diagram from the Vision Paper:



Q&A for the Information Analyst: A Vision

Question & Requirement Context; Analyst Background Knowledge

???

Natural Statement of Question; Use of Multimedia Examples; Inclusion of Judgement Terms

Query Assessment, Advisor, Collaboration

•Ranked List(s) of Relevant Info.

Select & Transform

Query Refinement based on Analyst Feedback

Relevance Feedback; Iterative Refinement of Answer

Proposed Answer

Multimedia Navigation Tools

•Relevant information extracted and combined where possible;
•Accumulation of Knowledge across Docs.
•Cross Document Summaries created;
•Language/Media Independent Concept Representation
•Inconsistencies noted;
•Proposed Conclusions and Answers Generated

IMPACT

- Accepts complex "Questions" in form natural to analyst
- Translates "Question" into multiple queries to multiple data sets
- Finds relevant information in distributed, multimedia, multi-lingual, multi-agency data sets
- Analyzes, fuses, summarizes information into coherent "Answer"
- Provides "Answer" to analyst in the form they want

Although there may be a substantial human intervention in the definition and evaluation of the question context, a research challenge is presented by the detection of the unknown facts or attributes in the context, such that the questioner is not presented with known facts unless (s)he wants so. For example, if the questioner is interested in the terrorist activities of a certain group, (s)he should be presented with the latest actions of that group only, unless she specified that (s)he wants to know all about that group.

Moreover, to include in the answer predictions about possible future courses of action, the context must interact with analogous data and rely on knowledge bases similar to those developed under DARPA's HPKB (High Performance Knowledge Bases) program for the Crisis Management task. In the HPKB competition, the Q&A task was oversimplified. First of all, the domain was clearly defined and supportive data was provided. Second, the questions were parameterized and no NLP was performed. Humans encoded knowledge base axioms with information they read from texts, encoded the questions in the same format and performed theorem proving to find the answers. The reasoning mechanisms were complex, and should be built for a large number of domains. This will enable complex open-domain Q&A. As the Q&A TREC-8 Track was an unqualified success in terms of achieving the goal of answering open-domain questions, the availability of high performance knowledge bases, equipped with complex reasoning mechanisms will enable answering complex, multi-faced questions across several domains. Contextual structures for Q&A need to be defined to control the interaction with knowledge bases.

Techniques for rapid development of large knowledge bases and complex reasoning mechanisms are currently under study in the DARPA RKF (Rapid Knowledge Formation) project and may be available in the near future.

However, not all four levels of questioner sophistication require complex contextual structures. The following table depicts examples of contexts and ambiguities they resolve:

Level 1 *“Casual Questioner”*	**Q:** Where is the Taj Mahal?	**Context:** Gazeteers: Taj Mahal – casino in Atlantic City : context 1 Taj Mahal– restaurant in Dallas TX : context 2 Taj Mahal – palace in Agra India: context 3
Level 2 *“Template Questioner”*	**Q:** Where is Microsoft recruiting from?	**Context:** Linguistic Patters for information extraction <Organization – hire – employee – as – position> <Employee joins – Organization – after leaving Organization 2>
Level 3 *“Cub reporter”*	**Q:** How many software products does Microsoft sell?	**Context:** Ontologies Software product = {operating system, word processor, spreadsheet, …} Operating system = {drivers, file managers, window managers, ….}
Level 4 *“Professional Information Analyst”*	**Q:** How is Microsoft struggling to maintain its leading position in the Software Industry?	**Context:** Knowledge Bases and Reasoning Mechanisms Leading position in one Industry => dominate market & develop good products, etc

Related subtasks:

1/ Develop theoretical models of context useful for Q&A.

2/ Study semantic models of dialogue and their representation of context.

3/ Contexts defined by users – in various formats, ranging from keywords to textual data or database entries to very complex situations modeled in a natural way.

4/ Study the impact of context representation on both question processing and answer processing. Analyze various retrieval models that find relevant information when context is provided.

5/ Model context from user profiles or from history of previously asked questions. Compare contexts of questions resulting in the same answer or in related answers. Study context and its impact on the explanatory knowledge derived from an answer.

6/ Integration of contextual knowledge into and from world knowledge and special purpose ontologies as well as axiomatic knowledge.

4. *Data sources for Q&A*:

Before a question can be answered, it must be known what knowledge sources are available. To see how correct the answer is, we need to have evaluation methods and test-beds in place. However, the processing of questions as well as the evaluation of the answers is dependent in a large measure on the data sources. If the answer to a question is not present in the data sources, not matter how well we perform question processing, retrieval and extraction of the answer, we shall not obtain a correct result.

TREC-8 allowed participants to use any knowledge source in their search for answers to 200 questions. In addition, it was guaranteed that the answers were present in some of the approx. 500,000 documents provided by NIST for the competition. NIST also provided the questions and performed a human-centered evaluation. The paper "*TREC-8 Q&A Track Evaluation*" by Ellen Voorhees and Dawn Tice details the question selection process as well as the evaluation used at TREC-8.

Not only that the complexity and the range of questions should scale up in future TRECs, but it is also necessary that the data sources become more heterogeneous and of larger size. It is possible to include a significant number of digital libraries (e.g. the Medline library for questions like TREC-8's Q8*: "What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearings and incoherent vocalizations (grunts, shouts, etc)?*")

At present, a significant amount of information is structured in databases of different formats and different semantics. Of extreme value would be the access to such databases, to support the Q&A process and to provide with more realistic scenarios, in which information is not available only in textual form, but in other formats as well. Moreover, even for the textual data, different file formats should be considered. Instead of only ASCII files with SGML tags, textual information should also be retrieved from PDF files, postscript files, Word documents, Excel and Powerpoint documents. The problem is to gather and organize large-scale collections of files in such varied formats. Similarly, there is need to access several large databases (e.g. the EDGAR database at the U.S. Securities and Exchange Commission). In addition several knowledge bases, in different formats should be made available. An example is the upper ontology of CYK or some other ontologies developed in LOOM. They will complement WordNet, the lexical database widely used in Q&A systems, as well as FrameNet, a database of several thousands of frame elements.

In addition, large sets of questions need to be generated. Not all questions should have an answer in the available data sources. Moreover, some questions should be formulated with a specification of their context. TREC-9 will introduce a new feature of the test questions - the reformulation of a large part of questions – i.e. several questions will ask for the same information, thus should produce the same answer. This feature should be maintained in the future TRECs, as it contributes to the measurement of the accuracy and robustness of Q&A techniques implemented in different systems.

As the ultimate goal of the Q&A paradigm is the creation of an integrated suite of analytical tools, the Q&A systems should be extended to dialogue processing systems that are queried and respond in real time to problems along the four levels of questioner sophistication. In the scenario of dialogues and follow-up questions, the dialogue-provoking questions have to be selected with special care, such that they determine the unfolding of multiple possible dialogues.

Related subtasks:

1/ Collect heterogeneous data formats for Q&A that grow naturally over time.

2/ Provide with databases in different formats and with heterogeneous data.

3/ Provide access to several digital libraries.

4/ Extract answers from multimedia data.

5. *Answer Extraction*: Extraction of simple and distributed answers; Answer Justification and Evaluation of Answer Correctness The ultimate goal of a Q&A system is to extract the answer from the underlying data sources. Answer extraction depends on the complexity of the question, on the answer type provided by question processing, on the actual data where the answer is searched, on the search method and on the question focus and context. Given that answer processing depends on such a large number of factors, research for answer processing should be tackled with a lot of care and given special importance.

   The requested information might be present in a variety of heterogeneous data sources, that must be searched by different retrieval methodologies. Several collections of texts might be organized into separate structures, with different index operators and retrieval techniques. Moreover, other information may be organized in databases, catalogues or may simply reside on the WWWeb, in HTML or XML format. The context could be represented in different formats, according to the application. However – the search for the answer must be uniform in all data sources and data entries, either in textual or other formats must be identified uniformly. To enable this process – different indexing and retrieval techniques for Q&A must be studied – methods of detecting overlapping information as well as contradictory information must be developed. A framework in which data retrieved from various data sources are passed to the answer processing module in a coherent and complete way must be defined and developed.

   Answers should not be searched only throughout text collections (the TREC case), but also in databases, catalogues and Web pages from Intranets. Moreover, answers may be found in existing knowledge bases or ontologies, or FAQ (Frequently Asked Questions) repositories. New retrieval models, that integrate efficiently all these data formats must be studied and developed. An early experiment was developed at Oracle by creating an in-house knowledge base comprising 200,000 concepts classified into 2000 major themes, to be used for knowledge retrieval from texts.

Answer extraction in itself implies the process of recognizing the answer of a question. In doing so at least three problems arise. The first one is related to the assessment of answer correctness, and implicitly to ways of justifying its correctness. The second problem relates to the method of knowing whether all the possible correct answers have been found throughout the data sources, and in a related way – whether no correct answer can be found to a given question. Thirdly, there is the problem of generating a coherent answer when its information is distributed across a document or throughout different documents. This problem can be extended to the case when an answer part is found in a given data source, whereas the other parts, with which it must be fused, are retrieved from different data sources, in different formats. Another extension is given by the case when special implicatures must be generated prior to the extraction of the answer – as the simple juxtaposition of the answer parts is not sufficient.

Answer justification: to assess the correctness of an answer, justifications, in the most natural format, should be made available. Justifications may be based on a wide range of information, varying from evidence enabled by natural language appositions to commonsense and pragmatic knowledge, available from on-line knowledge bases. A back-chaining mechanism implemented into a theorem prover may generate the justification to an answer in the most complex cases, and translate it into a format easily and naturally understood. In the case of textual answers, textual evidence is necessary for the final justification. Answers extracted from databases will be justified more easily, due to the inference implemented in such data sources.

Answer evaluations: Develop methods for answer evaluation with different quantitative measures that unify scoring across different data sources and throughout different text collections. A very low score may indicate the absence of data where the answer may be found. Evaluations must be done in conjunction with the treatment of contradictory answers.

Distributed answers: a particular interest is presented by questions that have answers distributed across different text collections or across different documents. Such questions and their underlying data sources pose research problems that add new complexity dimensions to the Q&A process. Combining information from different data sources is not a trivial problem. The answer may be either scattered throughout the collections, but its generation may not require any inference; or the answer may be provided only by inferring new information or relations between its subparts. The fusion of information from different documents and from different sources often relies on complex inferential processes. The interpretation of tables in text document adds to the complexity of the process. The following table presents some instances of challenging answer extraction at different sophistication levels:

Level 1 *“Casual Questioner”*	**Q:** When was Queen Victoria born?	**Text 1:** Queen Victoria (1854, 1889) ruled Britain with an iron fist …. **Text 2:** British monarchs: Victoria 1832-1889 Edward 1874-1946 Elizabeth 1923- **Answer: 1832**
Level 2 *“Template Questioner”*	**Q:** How many casualties were reported last week in Fredonia?	**Text 1:** Last Monday two people were killed on the streets of Beautiville, Fredonia, after a bomb exploded **Text 2:** The terrorists murdered a family with a small child in Fredonia last Friday, near its border with Evilonia. The father just returned home the day before. **Answer: five people**
Level 3 *“Cub reporter”*	**Q:** How many U.S. households have a computer?	**Text 1:** Two families in three are connected to the Internet in the U.S. **Text 2:** Last year, IRS has received 150 million individual return forms. **Answer: 90 million**
Level 4 *“Professional Information Analyst”*	**Q:** Why there were hacker attacks on the computers at University of California, Santa Barbara?	**Text 1:** U.S. colleges have powerful computing facilities. **Text 2:** Computer hackers need speedy processors to break security passwords. **Answer: To use their computers for password cracking**

Related subtasks:

1/ Develop theoretical models of answer extraction.

2/ Study quantitative evaluation metrics for answer correctness.

3/ Study qualitative models of answer completeness – part of these studies will develop models for the detection of the absence of an answer in the underlying data sources.

4/ Study analytical methods of answer justification and the acquisition of its necessary knowledge.

5/ Study NLP techniques that enhance answer extraction procession: e.g. coreference resolution, incorporation of world knowledge

6. *Answer formulation*

The result of a Q&A system should be presented in a way as natural as possible. In some cases, simple extraction is sufficient For example, the question Q8:” *: “What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearings and incoherent vocalizations (grunts, shouts, etc)?”* should be answered by returning only “*Tourette’s Syndrome*”. When the question classification indicates that the answer type is a name (of a person, organization, shop or disease, etc), a quantity (monetary value, length, size, distance, etc) or a date (e.g. the answer to the question “On what day did Christmas fall in 1989?”) the extraction of a single datum is sufficient. For all the other cases, instead of returning 50 or 250 bytes of text, we should return the answer in a format natural to the questioner. The format depends on the sophistication of the questioner. An immediate solution is to generate Web pages on the fly, containing the answer either as a generated text or as a table. The page should have hyperlinks to explanatory information and to the data sources from where the answer was extracted. In addition, the user should be allowed to input his/her own judgements or data.

The fusion of information from different sources requires "cut and paste" generation capabilities. Since the sentences containing the information to be fused are disconnected (as they come from different documents or from different parts of the same document) when they are strung together the resulting answer may be incoherent and sometimes misleading. Moreover, it might contain unnecessary information. To produce a coherent, concise textual answer from different sources several operations are needed. First, sentences must be reduced to contain only information relevant to the answer. Then they must be combined. Machine learning techniques (e.g. decision trees) can be used to learn combination operators. The implementation of combination actions involves joining several parse trees, substituting a parse subtree with another or adding additional subtrees. Often more than one sentence needs to be generated as an answer, thus RST-based generation needs to be implemented to produce a coherent textual answer.

To avoid redundant statements in the answer, we need to identify *theme intersections* between different sources. The identification of theme intersections requires knowledge about paraphrases, i.e. alternative ways in which a human can chose to "say the same thing". Moreover, multi-document fusion of information should be performed with a time perspective, i.e. within the correct temporal sequence of events.

Similarly, when the answer formulation takes the format of a database, information fusion involves the unification of the semantics of the different sources with the semantic of the desired database.

When the answer is part of an ongoing dialogue, the answer must be generated as a coherent part of that dialogue –and its natural format might involve generating referential expressions. Needless to say, that regardless of the format, most of the cases of information fusion require advance inference capabilities and additional knowledge.

For example, in the last table illustrating examples of answers, the answer to the level 1 (Casual questioner) question "When was Queen Victoria born?" is 1832 because the semantic of the table was understood. The table represents the life spans of the British monarchs. The other time period identified for Queen Victoria represents the duration of her reign. In this case the fusion was not necessary, as one of the sources was discarded. In contrast, the answers for all the other levels required fusion. In the case of the question for the Template questioner, the Q&A system had to have the ability to add the number of victims from last Monday to the number of victims from last Friday. The addition was not trivial, since only the first text explicitly states the number of victims. In the second text, a family is said to have been killed- and the family had one child. Then the system had to infer that this is a three-member family. The inferences for the other levels of questioners are harder.

Related subtasks:

1/ Models for answer fusion from different sources. Full range of question complexity. Process questions *like "What are the three most polluted cities in the U.S. ?*" up to questions like "*How likely is it that the Fed is going to rise interests at their next meeting?*"

2/ Study unification of retrieval methods and detection of overlapping or contradictory information

3/ Time stamping: When presenting a Q&A system with a question, often the present date, time or year is assumed by default. However, questions like "Who is the President of the U.S.?" have different correct answers, depending on the time stamp when the question was asked. Modeling distinctions among current facts or events and past facts or events is necessary.

4/ Q&A for the case of event tracking, time stamping of events and subevent identification or event correlations (text and data mining). These models comprise the recognition of new features, new attributes or development of events in a story that can be tracked across multiple documents.

5/ Treatment of sets of correct answers and of contradictory answers. Development of models of probable answers.

6/ Develop inference mechanisms for fusion of answers in different formats

7. *Real Time Question Answering* There is need for developing Q&A systems that are capable of extracting answers from large data sets in several seconds, regardless of the complexity of the question, the size and multitude of the data sources or the ambiguity of the question.

Related subtasks:

1/ Detect time bottlenecks: retrieval and answer extraction.

2/ Study fast models of retrieval.

3/ Study answer extraction techniques that have small execution times – enhance reasoning mechanisms with fast execution capabilities. In many applications execution time is essential: either the customer looses patience and interest or people are in critical missions – pilots in action, military in mission.

4/ Study scalability issues – Solution: distributed Q&A.

8. *Multi-Lingual Question Answering* The ability of developing Q&A systems for other languages than English is very important. Moreover, the ability of finding answers in

texts written in languages other than English, when an English question is asked is very important.

Related subtasks:

1/ Develop part-of-speech taggers, parsers and Named Entity recognizers for other languages than English.

2/ Translate English questions in other languages. Translate the answer in English.

3/ Tools for answer retrieval in other languages

4/ Develop knowledge bases and ontologies that contain concepts that are language-independent (interlingua type of hierarchies, linked to concepts or words in other languages than English).

5/ Develop multi-lingual retrieval engines. Generate parallel sets of answer paragraphs.

9. *Interactive Q&A*

It is often the case that the information need is not well captured by a Q&A system, as the question processing part may fail to classify properly the question or the information needed for extracting and generating the answer is not easily retrieved. In such cases, the questioner might want not only to reformulate the question, but (s)he might want to have a dialogue with the system. It is generally acknowledged that developing a successful computational model of interactive natural language, a dialogue component based on extensive analysis of sample dialogues needs to be implemented. Although the problem is complex, much analysis of human-human interactions has been done, such as Walker and Whittaker's work, Oviatt and Cohen's research. Most of the data employed was either human-human dialogues in relevant task domains or the Wizard-of-Oz dialogues in which a human (the Wizard) simulates the role of the computer as a way of testing the dialogue model. Theories of mixed initiative dialogues and results of their implementations have been recently published.

Most prior work on natural language dialogue has either focused on individual subproblems or focused on database query applications. Wizard-of-Oz experiments for Q&A dialogues need to be performed and a dialogue processing model implemented in various systems needs to be developed. Some of the characteristics of such a model are (1) it has coherent subdialogue movement; (2) user profile usage; (3) expectation usage; (4) variable initiative behavior.

Related subtasks:
1/ Dialogue models for Q&A: follow-up questions, reference resolution, detection of intentions, common goals and plans.
2/ Develop models that detect new developments – distinguish what is new from previous answers
3/ Develop conversational models in which the system makes suggestions: e.g. "You mean …"

10. *Advanced Reasoning for Q&A*

More sophisticated questioners expect answers which are outside the scope of written texts or structured databases. To upgrade a Q&A system with such capabilities, we need to integrate reasoning components operating on a variety of knowledge bases, encoding world knowledge and common-sense reasoning mechanisms as well as knowledge specific to a variety of domains. We also need to allow for the representation of scenarios of interest, capabilities of inferring new facts if required by the answer and a way of assembling all these facts and presenting them to the answer generator component.

A knowledge-based component of a Q&A system should be viewed as an augmentation or, rather than a rival to the retrieval-based approaches. The potential benefits are:

Customization: As answers are synthesized from a knowledge base, answers can be customized to the user's particular situation.

Controllable level of detail: The level of detail can be dynamically controlled to suit the user's level of expertise, by controlling how much information from the knowledge base is included in the answer.

Robustness: By inferring answers rather than extracting them, the Q&A system can respond to unanticipated questions and can resolve situations in which no answer could have been found in the sources of data.

Related subtasks:
1/ Incorporate knowledge representation and reasoning mechanisms that allow complex inferences (e.g. reasoning by analogy).
2/ Incorporate models of common sense reasoning.

11. User Profiling for Q&A

The user profile captures data about the questioner, comprising context data, domain of interest, reasoning schemes frequently used by the questioner, common ground established within different dialogues between the system and the user etc. The profile may be represented as a predefined template, where each template slot represents a different profile feature. Profile templates may be nested one within another. The Q&A system fills the template slots for each questioner that uses it. Two research issues arise. First, we need to identify the profile template slots characteristic for each class of questioners. The following table lists some of the features useful for the four levels of sophistication identified in the Vision Paper.

Level 1 *“Casual Questioner”*	**Profile features:** user satisfaction, frequency of questions, number of follow-up questions; domain of interest; number of questions asked repeatedly over time; number of questions with related topics
Level 2 *“Template Questioner”*	**Profile features:** number of related templates, number of retrieved answers and their relevance, frequency of reuse of templates; frequency with which new templates are introduced
Level 3 *“Cub reporter”*	**Profile features:** length of dialogues; dialogue topics; complexity of context, number of unanswered questions; number of questions reformulated; number of similar or related questions within a Q&A session and throughout sessions.
Level 4 *“Professional Information Analyst”*	**Profile features:** number of newly discovered facts, relations, feedback usage; time spent navigating new mined evidence; number of new axioms in the knowledge base that were added; complexity of the reasoning schemes required

Related subtasks:
1/ Develop profiling models: acquire knowledge about the user’s interests, habits, intentions, area of interest.
2/ Develop models that detect related questions or answers among multiple users.

12. <u>Collaborative Q&A</u>

Related subtasks:
1/ Models for Q&A that detect users operating in the same context, having the same or similar agendas
2/ Develop models that mine for unrelated questions that have related answers and vice versa.

3. **MILESTONES IN THE PROGRAM**

The following table summarizes the tasks and their respective subtasks and sets milestones that correspond to the TREC competitions spanning the following years. We have associated with each subtask three different levels of difficulty:

- 1 – for an “easy task”
- 2 – for a “more difficult task”
- 3 – for an “advanced task”

As it can be seen in the table, when an advanced task is introduced in a given year, all the following years will continue with more requirements at the same level of difficulty.

	2000 TREC 9	2001 TREC 10	2002 TREC 11	2003 TREC 12	2004 TREC 13	2005 TREC 14
1.Question Classes						
1. Question taxonomy		1	2	3	3	3
2.Question complexity		1	2	2	3	3
3.Question complexity criteria		1-2	2-3	3	3	
4. KB, ontologies			1	2	3	3
2.Question Processing						
1. Models of Question Processing		1	2	3	3	3
2. Q similarity, subsumption		1	2			
3. Q implications				1	2	3
4. Map Q into simple inquiries		1	2	3	3	3
5. Q ambiguities			1	2	3	3
3.Context and QA						
1. Context models		1	2	3	3	3
2. Context from dialogue				1	2	3
3. User defined context		1	2	3	3	3
4. Impact of context on QA		1	2	3	3	3
5. Context from user profile				1	2	3
6. Contextual K + KB				1	2	3
4. Data Sources						
1. Heterogenous data formats		1	2	2-3	3	3
2. Databases		1	2	3	3	3
3. Digital Libraries			1	2	3	3
4. Multimedia				1	2	3
5.Answer Extraction						
1. Models of Answer extraction		1	2	3	3	3
2. Evaluation metrics of Answer Extraction		1	2	3	3	3
3. Answer Completeness			1	2	3	3
4. Answer justification			1	2	3	3
5. Coreference, KB			1	2	3	3
6.Answer Formulation						
1. Answer fusion		1	2	3	3	3
2. Overlapping & Contradictory			1	2	3	3
3. Time stamping		1	2	3	3	3
4. Event tracking				1	2	3
10. Sets of correct A, probable A, contradictory A				1	2	3

11. Inference for fusion			1	2	2	3
7.Real-time QA						
1. Time complexity analysis		1-2	3	3	3	3
2. Fast IR models			1	2	3	3
3. Fast AE models tradeoffs			1	2	3	3
4. Scalable QA				1	2	3
8.Multilinqual QA						
1. POS, parsers, NE for other languages than English			1	2	3	3
2. Q: E -> L; A : L -> E			1	2	3	3
3. IR in other languages			1	2	3	3
4. KB in other languages				1	2	3
5. Multilingual IR engines				1	2	3
9.Interactive QA						
1. Dialogue models				1	2	3
2. QA for new developments				1	2	3
3. Conversational models				1	2	3
10.Advanced QA						
1. Complex KR & Reasoning				1	2	3
2. Common sense reasoning				1	2	3
11.User profiling						
1. Profiling models			1	2	3	3
2. Related Q			1	2	3	3
12.Collaborative QA						
1. Model collaborative QA same context, same agenda				1	2	3
2. Model different agendas				1	2	3

4. **GENERAL FRAMEWORK FOR EVALUATING THE Q&A TECHNOLOGY**

4.1 Overview

The main assumption in our evaluation proposal is that each year we shall address several new problems that would enable the capability of testing at the end of the 5th year a number of sophisticated Q/A systems, capable of handling questions posed by users whose expectations are somehow similar to those of professional information analysts. In the five-year program that we outline here we propose to continue with the user-centric evaluation already used in the TREC-8 and TREC-9 Q/A tracks, as we believe that Q/A systems should primarily satisfy humans rather than get aligned to some functional evaluation scheme. However, due to some inconsistencies noticed initially in the TREC-8, which became more significant in the TREC-9, we propose to have at least three human evaluators judging independently the answers to each question, such that inter-judgement agreements can be analyzed.

It is to be noted that we envision this very ambitious and simple evaluation program not only as a succession of five steps of accelerated difficulty. We believe that each of the five sets of problems will generate specific issues brought forward by the actual tests. It is thus natural that each year will springboard a new Q&A sub-track, with its own followers, that will consider variations, re-definitions and perhaps several test cycles, in which progress could be measured and lessons be learned. Setting up a very ambitious evaluation plan is not only driven by an enthusiastic vision, but it also offers a good starting point for scaling back instead of only moving forward the complexity of the Q&A task. It is a different perspective than the one that has initiated the Q&A evaluations. At the beginning, in TREC-8, simplicity was sought because it was the only way of having Q&A take off. The result was an unqualified success, that prompted immediately the question “But how successful we *really* are?”. By providing several degrees of complexity, each year a new set of them, the proposed evaluation enables scenarios that generate data that will tell how successful we really are. Moreover, each year the tasks impose new constraints on the user-centric aspect of the evaluations, that we hope will be studied more in depth in following cycles of tests.

We also propose that in the fourth year of the roadmap, the Q/A effort should merge with the Summarization effort, thus incorporating the evaluation methods already tested in the summarization competitions. For the first four years we propose to use the TREC data as the text collections mined for Q/A. The fifth year will employ both structured and unstructured data relevant to a few specific domains. For training purposes, prior to the fifth year competition, several such data collection will be made available to the participants (e.g. on-line technical manuals, catalogs etc). Throughout the roadmap we assume that the evaluations will take place within the context of the TREC competitions.

Year 1: TREC-10

- New requirements

(a) *The answer is scattered across two or more documents*. A step towards more realistic situations is to consider questions for which single documents contain only partial answers, with a complete answer found across several documents. In these cases the global answer has to be assembled from partial answers via a fusion process. The format of the answer will no longer be a text snippet of 50/250 bytes, but rather would change to “the answer”, representing a name, a date for currency value expression, a list or even a one or several sentences if the question requires more complex answers.

(b) The answer is *no longer guaranteed* to be present in the text collection and the systems are expected to notify the absence of an answer.

- Examples

(Q1) *What countries from the South America did the Pope visit and when?*
Answer:
Argentina – 1987 [Document Source 1]
Columbia – 1986 [Document Source 2]
Brazil – 1982, 1991 [Document Source 3]
(Q2) *When did President Kennedy visit Cuba ?*
Answer:
There is no answer in the collection.

(Q3) *Who advised Nixon to open relations with China and why?*
Answer:
Henry Kissinger. [Document Source 1, 2, 3] He believed that it is in the U.S. strategic and business interests to have trade relations with China. [Document Source 1, 4]

(Q4) What are the major causes for aircraft crashes?
Answer:
- *bombs, evidenced by the Lockerbee tragedy [Document Source 1, 2 , 3]*
- *mechanical failure, as in the case of the Paris Concorde crash [Document Source 4, 5, 6, 7, 8]*
- *explosive cargo, as in the case of the 1996 Everglades accident [Document Source 9 10 11 12]*

- Discussion

The main challenges of the first year evaluations are determined by:
- The answer extraction from different documents and its fusion into a single answer.

- Detection of overlapping and contradictory information.
- Time stamping, as exemplified in question Q1.
- Event tracking, comprising first story detection and well as story continuation.
- Recognition of questions with no answers in the collection.

Year 2: TREC-11

- New requirements

The Q/A process will take place *"within a context*". We propose to represent the contexts through TREC queries, and the data collections will be the relevant documents for those queries. For any given topic, represented within a TREC query, series of questions shall be formulated, that take into account the fact that the system has access to the title, the description and the narrative associated with the topic. The capability of processing follow-up questions is a first step towards a dialogue Q/A environment.

- Examples

Topic 168:

- Title: *Financing AMTRAK*
- Description: *A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).*
  - Narrative: *A relevant document must provide information on the government's responsibility to make AMTRAK an economically viable entity. It could also discuss the privatization of AMTRAK as an alternative to continuing government subsidies. Documents comparing government subsidies given to air and bus transportation with those provided to AMTRAK would also be relevant.*

(Q1) *Why AMTRAK cannot be considered economically viable ?*
(Q2) *Should it be privatized ?*
(Q3) *How much larger are the government subsidies to AMTRAK as compared to those given to air transportation ?*

- Discussion

The main challenges of the second year evaluations are determined by:

- *The comprehension of the context.* The interactions between the context and the processing of the questions, on one hand, and the extraction of the answer on the other hand. For example the answer to question Q1 needs to fuse in different sources of information that present why AMTRAK is not currently a viable economic entity. It is not expected that the answers would comprise world knowledge facts, such as the assumption only non-economically viable entities receive subsidies from the Government. However, the answer is expected to provide with evidence about the economical viability of AMTRAK and especially the causes that are explicit in the underlying texts. The search for such evidence and for its causes is expected to employ information from the context, such as the privatization of AMTRAK or its government subsidies.

- *Update of the context.* As a series of questions shall be related to a given context, the answers bring forward new information, thus updating the contextual knowledge. For example, the answer of question Q1 will provide information about the reasons that explain the economic situation at Amtrak. The processing of Q2 should take into account this novel information as well. Moreover, this new information might not be relevant to question Q3, whose processing falls back on the initial context. Such decisions will be determined by the content of the answer extracted and formulated for each previous question.

- *The common ground.* The processing “within context” entails also a common ground, thus a dialogue-like environment. The evaluations will consider only the resolution of reference, in three particular situations:

  (a) Reference between two questions from the same context. For example:
  (Q’) Why does *AMTRAK* receive subsidies from the government?
  (Q”) Should *it* be privatized ?

  (b) Reference between two entities in the same answer document. For example:

  AMTRAK employs 50000 people and connects over 1000 cities in the U.S. However *the company* was in red for the past couple of years.

  (c) Reference between the answer to a previous question and a following question. For example:

  (Q’) How many people are employed by AMTRAK ?

  (A’) AMTRAK employs *50000 people*,

  (Q”) Is *this* more than in air transportation?

Year 3: TREC-12

- New requirements

The answer will no longer be a text snippet or a table of items, but a product *of text generation*. Moreover, if desired, the generation will comprise an *explanation* of possible ambiguities and a *justification* of why the answer is right.

- Examples

(Q1) Where is the Taj Mahal?

Answer: *If you are interested in the Indian landmark, it is in Agra, India. If instead you want to find the location of the Casino, it is in Atlantic City, NJ, U.S.A. There are also several restaurants named Taj Mahal. A full list is rendered by the following hypertable. If you click on the location, you may find the address.*

The Taj Mahal Indian Cuisine	Mountain View, CA
The Taj Mahal Restaurant	Dallas, TX
Taj Mahal	Las Vegas, NV
Taj Mahal	Springfield, VA

(Q2) How did Socrates die?

Answer: *He drunk poisoned wine. Anyone drinking or eating something that is poisoned is likely to die.*

- Discussion

The main challenges of the third year evaluations are determined by:

- *The quantification of the "naturalness" of the generated answer.* It is a fine line dividing the situations requiring explanations or justifications and those that do not impose any explanations or justifications, as the answer "says it all". For example, the answer to question Q1 does not require explanations of why the Taj Mahal Palace is in India, whereas the Casino is in Atlantic City. However, since there are lots of Taj Mahals, it organizes the generation of this information into a hypertable. What are all the situations that will require the generation of a hypertable? How many elements are allowed to be listed in a table at a time ? These are issues that must be investigated before starting the evaluations.

- *The recognition and explanation of ambiguities.* Natural language is extremely ambiguous, thus for an open-domain application such as Q&A, an open list of ambiguities should be allowed. An important part of these evaluations is going to be dedicated to the recognition of ambiguities, similar to those described in pages 7-9 of this document.

- *The generation of answer justifications.* The ultimate test of the correctness of an answer is provided by a justification. In the COLING-2000 paper "Experiments with Open-Domain Textual Question Answering", Harabagiu et. al report a method of generating abductive justifications by combining three forms of knowledge: (a) information derived from the facts stated in the text; (b) world knowledge, accessed from WordNet and other sources and (c) information derived by reference resolution. In the experiments reported, justifications were traces of a theorem prover. The challenge is to generate justifications expressed in natural language.

Year 4: TREC-13

- New requirements

More complex questions will be asked, requiring the answers to be *summaries* of the textual information comprised in one or several documents. The summarization is going to be driven by the question, an approach already tested and implemented in the context of Frequently Asked Questions (FAQ) documents (cf. Berger and Mittal's ACL-2000 Paper "Query Relevant Summarization using FAQs"). Moreover, the summary offered as an answer will present in a coherent manner information from one or multiple documents, using text generation capabilities. "Cut and Paste" summaries, delivered by approaches similar to the one reported in Jing and McKeown's NA-ACL-2000 paper "Cut and Paste Based Text summarization" or summaries obtained by fusion of information from different articles, similarly to the work reported in Radev and McKeown's Computational Linguistics 24(3) paper "Generating Natural Language Summaries from Multiple On-Line Sources". The parallel evaluations of summarization systems will also highlight some other possible summaries that are driven by questions. We propose three kinds of questions:

(a) *Context-based summary-generating questions*. Such questions are more difficult than the questions used in TREC-11, as only a summary will provide a satisfactory answer.

- Example

Topic 168 (detailed above)
Question: *What is the financial situation of AMTRAK?*

(b) *Stand-alone summary-generating questions*. Such questions are open-topic, thus information relating to several topics may be relevant and needed to be added in the summary. Furthermore, these questions might not relate to any of the topics offered as possible contexts. Once again, the requirement that only a summary provides a satisfactory answer is maintained.

- Example

Question: *How safe are commercial flights?*

(c) *Example-based summary-generating questions*. Such questions are a special case of the context-based summary-generating questions, as they try to find similar situations, facts or entities are those described by the current topic. They also are answered only by summaries.

- Example

Topic 168 (detailed above)
Question: *What other companies are operated with Government aid?*

- Discussion

The main challenges of the fourth year evaluations are determined by:

- *The interaction between the context, the question and the quality of the context-based summary.* Given a specific topic, e,g, Amtrak financing, and a question asking about the financial situation at Amtrak, should the summary address all the points from the context that relate to the question or just some of them. How is the relationship between the question and the context being established? How large should the summaries be?

- *The interaction between the example's context, the question and the quality of the example-based summary.* Given a specific situation, event or series of events and a question asking about other examples of similar situations or examples, should the summary address only the similar points or also the eventual differences.

- *Measuring the "informativeness" of stand-alone summaries.* Measuring the precision and the recall of open-domain summaries is not trivial, especially when the underlying text collection is large, as in the case of the TREC collection.

Year 5: TREC-14

- New requirements

The questions asked will be at *expert-level*, as sufficient structured and unstructured information for different domains will be made available. Delivering expert-like answers is based on the capability of mining domain knowledge and mastering the relationships between all activities, situations and facts within a specific domain. Reasoning by analogy, comparing and discovering new relations are expected.

- Examples

(Q1) What are the opinions of the Danes on the Euro?

Provided with data from the European Parliament regarding the performance of the Euro, the legislature related to it and with opinions from politicians of different European countries as well as data from Danish polls, the answer presents the statistics from the polls and motivates the opinions by showing the pros (e.g. the economy in France) and the cons (e.g. the British Pound untouched by recent fluctuations of the Euro).

(Q2) Why so many people buy four-wheel-drive cars lately?

Consumer reports, technical data from manufacturers and statistics from several car dealers support the market tendency, and explain the answer. The expertise is shown through the analysis of the advantages of driving a four-wheel-drive.

(Q3) How likely is it that the Fed will raise the interest rates at their next meeting?

Data regarding the past decisions of the Fed, given certain values of inflation, stock market performance, employment data and other major political factors are used to make a prediction of the Fed's expected actions. The answer should also formulate comparisons to previous situations and the Fed's action on the interest rates.

- Discussion

The main challenges of the fifth year evaluations are determined by:

-*The heterogeneity of domain data.* Domain data will be provided in several formats: (a) collections of news articles; (b) database entries in different database formats; (c) political essays; (d) digital library entries – in the forms of technical manuals or journal publications. Mining domain knowledge from all these different formats will be a non-trivial challenge. Several months prior to the competition, the participants will be offered some training data, to get accustomed with all the various formats of the data input.

-*Techniques for mining information on the fly and at expert-level will need to be developed.* Given a specific domain, the identification of all its relevant knowledge as well as all the necessary analogies and comparisons between different situations or events/entities important in that domain need to be performed. Along with the training domain knowledge, the participants will be provided with training expert questions.

-*The evaluation of the expertise displayed in the answer.* The evaluation of the correctness of the answers will be more difficult, as good command of the domain knowledge is required.

5. **Appendixes**

APPENDIX A

The 13 conceptual question categories used in Wendy Lehnert's QUALM

Question Categories	Examples
1. Causal Antecedent	Why did John go to New York? What resulted in John's leaving? How did the glass break?
2. Goal Orientation	For what purposes did John take the book? Why did Mary drop the book? Mary left for what reason?
3. Enablement	How was John able to eat? What did John need to do in order to leave?
4. Causal Consequent	What happened when John left? What if I don't leave? What did John do after Mary left?
5. Verification	Did John leave? Did John anything to keep Mary from leaving? Does John think that Mary left?
6. Disjunctive	Was John or Mary here? Is John coming or going?
7. Instrumental/Procedural	How did John go to New York? What did John use to eat? How do I get to your house?
8. Concept Completion	What did John eat? Who gave Mary the book? When did John leave Paris?
9. Expectational	Why didn't John go to New York? Why isn't John eating?
10. Judgmental	What should John do to keep Mary from leaving? What should John do now?
11. Quantification	How many people are there? How ill was John? How many dogs does John have?
12. Feature Specification	What color are John's eyes? What breed of dog is Rover? How much does that rug cost?
13. Request	Would you pass the salt? Can you get me my coat? Will you take out the garbage?

APPENDIX B

Arthur Graesser's Taxonomy of Inquiries

Question	Abstract Specification	Examples
1. Verification	Is a fact true? Did an event occur?	Is an F-test a type of statistic? Did it rain yesterday?
2. Comparison	How is X similar to Y? How is X different from Y?	In what way is Florida similar to China? How is an F-test different from a t-test?
3. Disjunctive	Is X or Y the case? Is X, Y, or Z the case?	Do the mountains increase or decrease the rain in Oregon? Did he order chicken, beef, lamb of fish?
4. Concept completion	Who? What? When? Where? What is the referent of a noun argument slot?	Where are the large population densities in North America? Who wrote the song? What did the child steal?
5. Definition	What does X mean? What is the superordinate category and some properties of X?	What is a factorial design? What does interaction mean?
6. Example	What is an example of X? What is a particular instance of the category?	What is an example of an ordinal scale? What experiment supports this claim?
7. Interpretation	How is a particular event interpreted or summarized?	Does the graph show a main effect for "A"? What happened yesterday?
8. Feature specification	What qualitative attributes does entity X have? What is the value of a qualitative variable?	What is George like? What color is the dog?
9. Quantification	What is the value of a quantitative variable? How much? How many?	How many rooms are in the house? How much profit was made last year?
10. Causal antecedent	What caused some event to occur? What state or event causally led to an event or state?	How does warm air get to Ireland? Why is the kite going backwards?
11. Causal consequence	What are the consequences of an event or state? What causally unfolds from an event or state?	What happens to the warm winds when they reach the mountains? What are the consequences of double-digit inflation?
12. Goal orientation	What are the motives behind an agent's actions? What goals inspired an agent to perform an action?	Why did Roger move to Chicago? What was the purpose of the city's cutting taxes?
13. Enablement	What object or resource anables an agent to perform an action?	What device allows you to measure an earthquake? What do I need to bake this fish?
14. Instrumental/Procedural	How does an agent accomplish a goal? What instrument or body part is used when an agent performs an action? What plan of action accomplishes an agent's goal?	How does a person perform long division? How do you move a mouse on a computer?
15. Expectational	Why did some expected event not occur?	Why wasn't there a war in Iraq? Why doesn't this doll have a mouth?
16. Judgmental	The questioner wants the answerer to judge an idea or to give advice on what to do.	What do you think about the new taxes? What should I do to stop the fight?
17. Assertion	The speaker expresses that he or she is missing some information.	I don't understand what this message on the computer means. I need to know how to get to the Newark airport.
18. Request/Directive	The speaker directly requests that the listener supply some information.	Please tell me how to get a printout of this file.