

# IBM Research Report

## Towards the Open Advancement of Question Answering Systems

**David Ferrucci<sup>1</sup>, Eric Nyberg<sup>2</sup>, James Allan<sup>3</sup>, Ken Barker<sup>4</sup>, Eric Brown<sup>1</sup>,  
Jennifer Chu-Carroll<sup>1</sup>, Arthur Ciccolo<sup>1</sup>, Pablo Duboue<sup>1</sup>, James Fan<sup>1</sup>,  
David Gondek<sup>1</sup>, Eduard Hovy<sup>5</sup>, Boris Katz<sup>6</sup>, Adam Lally<sup>1</sup>, Michael McCord<sup>1</sup>,  
Paul Morarescu<sup>1</sup>, Bill Murdock<sup>1</sup>, Bruce Porter<sup>4</sup>, John Prager<sup>1</sup>,  
Tomek Strzalkowski<sup>7</sup>, Chris Welty<sup>1</sup>, Wlodek Zadrozny<sup>1</sup>**

<sup>1</sup>IBM Research Division  
Thomas J. Watson Research Center  
P.O. Box 704  
Yorktown Heights, NY 10598

<sup>2</sup>Carnegie Mellon University

<sup>3</sup>University of Massachusetts

<sup>4</sup>University of Texas

<sup>5</sup>USC/ISI

<sup>6</sup>MIT

<sup>7</sup>SUNY Albany



# Towards the Open Advancement of Question Answering Systems

December 2008

David Ferrucci (IBM), Eric Nyberg (Carnegie Mellon University), James Allan (University of Massachusetts), Ken Barker (University of Texas), Eric Brown (IBM), Jennifer Chu-Carroll (IBM), Arthur Ciccolo (IBM), Pablo Duboue (IBM), James Fan (IBM), David Gondek (IBM), Eduard Hovy (USC/ISI), Boris Katz (MIT), Adam Lally (IBM), Michael McCord (IBM), Paul Morarescu (IBM), Bill Murdock (IBM), Bruce Porter (University of Texas), John Prager (IBM), Tomek Strzalkowski (SUNY Albany), Chris Welty (IBM), Wlodek Zadrozny (IBM)

## 1 Summary

On February 27-28, 2008, a group of researchers from industry and academia met to discuss the state of the Question Answering (QA) field. The discussion focused on recent experiences from funded research programs (e.g. AQUAINT, HALO) and open evaluations (e.g. TREC, NTCIR). The group acknowledged that funded research programs and evaluations have been instrumental in establishing fundamental QA research. However, major advances in the field of QA are yet to be realized. Advances that can openly accelerate progress and greatly generalize QA technologies to produce scalable, more adaptable methodologies and business applications are within our reach. Although there are deep technical challenges, we believe these challenges can be effectively addressed by open collaboration in the open-source development of integrated QA technologies aimed at well-chosen sets of QA applications.

It is currently difficult to discern the adaptability or generality of algorithms that are published as part of a solution to a particular QA problem. There are no means to leverage individual contributions from distributed groups, so development tends to take place in a single organization, with lots of redundant (as opposed to shared) effort. As a result, sponsors have difficulty in determining which component technologies are really working, which require more research attention, and which are already working well enough for the problem at hand. Attempts to adapt systems that were built and evaluated using TREC datasets for use in other applications indicate that there is much fundamental research left to be done, particularly in improving the run-time speed, answer confidence, and domain adaptability of QA technologies.

The goal of the workshop was not to write another general, all-encompassing road map for QA research problems, but to address the problems mentioned above, so that ongoing research can be streamlined – reducing overall cost and time to innovate, while providing a more supportive environment for individual research contributions from academic organizations. The concept that emerged from the workshop is one of *open advancement*: the use of shared system models, open-source components, collections of challenge problems and common evaluation metrics, so that the contribution of each technology to end to end performance can be accurately measured and the community as a whole can uniformly advance system performance on an ever broadening range of QA problems.

Our specific objective is to combine formal metrics and rigorous module and end-to-end system evaluation with a collaborative research process that allows our field, as a research community, to achieve monotonically increasing performance levels across multiple instances of QA problems, while managing overall research and development cost effectively.

This document captures the thoughts and proposals from the workshop regarding open advancement, and is intended as a working document to aid in the creation of new sponsored research programs in industry and government.

## 2 Introduction

Question Answering (QA) is an application area of Computer Science which attempts to build software systems that can provide accurate, useful answers to questions posed by human users in natural language (e.g., English)<sup>1</sup>.

A QA system is a software system that provides *exact* answers to natural language questions for some range of topics. The notion of *exact* in this context is ultimately a subjective measure intended to indicate that a QA system is distinguished by providing responses that contain just the information necessary to precisely answer the question intended by user. The QA system's exact answer may be supplemented with additional information, including a justification or dialog explaining why the provided answer is correct.

Another distinguishing and important characteristic of QA systems is that their accuracy and their ability to justify an answer should increase with the amount of relevant information provided in the question. They should, therefore respond more accurately and more completely to longer, denser questions, that is, to questions which provide more information about what is being asked. This behavior would suggest that QA systems rely on a deeper semantic “understanding” of the intent of the question.

The resources used to answer questions can vary from unstructured data (e.g., typical web pages, blog posts) and semi-structured data (e.g., Wikipedia) to completely structured data (facts mined from the Web or pre-existing databases). Research and development of QA systems has been evaluated since 1999 in the yearly TREC QA track evaluations conducted by NIST, and has been supported in the U.S. by the AQUAINT program (2001-2008). Question answering systems are also evaluated in the context of two other academic workshops, CLEF (Europe) and NTCIR (Asia).

The AQUAINT program has already demonstrated the practical potential of question answering software. Several teams in the program have delivered working end-to-end systems to the program sponsor, and in some cases systems have been deployed and

---

<sup>1</sup> In this paper we limit our discussion to text-based QA, but acknowledge that QA systems can be developed to directly address other modalities including image, speech, music and video for example.

made to interoperate via web services. Mature systems tuned to TREC QA factoid questions have delivered batch answers with 70% first-answer accuracy.

Tempering these successes, recent TREC experience has also shown that system performance tends to drop dramatically when a new domain or problem (question type) is introduced, which might imply that systems are over-tuned to a particular problem instance for each yearly evaluation, and that research progress has been incremental rather than general. If one considers the TREC scores for a particular system across yearly tasks, one notes that even with constant research and development, teams cannot always sustain the same level of performance on a new task. The superficial level of algorithmic detail presented in a typical academic paper on a complex QA system does not support direct checking and replication of results, so it is difficult to assess the contribution of individually reported algorithms to overall system performance.

In a 2003 document entitled “Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)”, Burger et al. presented a diverse set of research topics in QA along with a corresponding set of planned evaluations (e.g. TREC QA tracks or tasks), but without details on how metrics would be developed and applied for more detailed forms of QA evaluation. In a workshop on QA evaluation held at LREC in 2002 (Maybury, 2002), the participants spent time creating considerable road map detail regarding resources for development and evaluation, methods and algorithms, and systems performance, but neglected to outline the process of collaborative systems integration, measurement and analysis.

Although prior workshops and road mapping exercises in QA have helped to define the agenda for advanced research, less progress has been made on the definition of a transparent, repeatable process that can be used to evaluate component technologies and end-to-end systems working on a variety of problems. Our observation is that precisely such a transparent, repeatable process, and a variety of shared evaluation metrics is what is required to help QA technology reach more general levels of applicability and higher levels of sustained and relevant performance. The process should be general enough to support not only question answering from English text resources, but also video and audio retrieval, and cross-lingual retrieval from resources in other languages.

At the conclusion of the workshop, the participants agreed to collaborate on a white paper to document the vision and requirements for open advancement of QA systems. This document is the result of that collaboration, and contains these main sections:

- **Vision.** What is open advancement? Why is it important? What are the benefits for each stakeholder (industry, academia, government, etc.)?
- **Challenge Problems.** We outline a set of challenge problems (QA applications) that will drive improvements in the state of the art across measurable dimensions (such as answer accuracy, answer confidence, and response time); the problems are diverse by design, so that performance gains can be tested across multiple problem instances to demonstrate general advancement.

- **Approach to Open Collaboration.** A discussion of how a consortium of researchers from industry, academia and government could work together towards open advancement of QA technology.
- **Open Collaboration Model.** An outline of the intended software development process, a first version of which has already been established by IBM and is being tested at Carnegie Mellon University.

### 3 Vision

A typical question answering system is a complex software configuration which combines a variety of text processing modules into a particular dataflow configuration. Two broad objectives of QA research are: a) to identify the most useful text processing modules for acquiring knowledge, analyzing questions, documents, passages, answers, etc.; and b) to identify the most effective dataflow configuration for a run-time QA system. A particular QA technology is comprised of a set of modules and dataflows representing a specific approach or solution, usually tailored to a particular type of question or resource. When the application domain requires a very fast system response to an input question, an additional objective is to design and deploy an effective run-time hardware configuration which takes advantage of parallel distributed computing.

The goal of open advancement is to rapidly accelerate the pace of progress, both in developing effective core algorithms and in deploying practical solutions. We plan to establish a broader collaboration on the open advancement of QA that will help to share innovation across the field as a whole, and provide deeper insight into the generality and domain adaptability of emerging QA technologies. The open advancement of QA is expected to promote teaming of stakeholders from industry, academia and government. Overall, we expect open advancement to provide the following synergies between stakeholder groups:

- *Industrial partners* gain better insight into levels of component and system performance required to solve practical problems, and are better able to leverage commercial applications of QA technology;
- *Government partners* gain better insight into performance bottlenecks and measures that can be used to associate new research efforts with detailed performance evaluation results; as a result, research program management becomes more cost-effective;
- *Academic partners* gain the opportunity to evaluate new component algorithms in the context of end-to-end systems working on a variety of problem instances, improving the breadth and quality of their research output without having to build a complete end-to-end framework on their own.

The broader collaboration will be supported by a group consensus on best practices regarding key elements of QA system development: a) a shared logical architecture for

integration and evaluation; b) a shared deployment architecture for run-time systems; c) a set of common challenge problems to drive innovation along important dimensions; d) a collaborative, distributed process for integration, testing and evaluation; and e) an open collaboration approach to software development, to support transparent, controlled evaluation and straightforward replication of results. These aspects of open advancement are introduced in the subsections that follow.

### **3.1 Shared Logical Architecture**

If one goal of QA research is to identify the most effective algorithms and dataflows for a given problem, then the notion of “effective” must be defined via some set of formal metrics on QA system modules and dataflows. Metrics are applied to a particular dataflow, corpus and question set in order to precisely measure module and system performance. As the system is tested across a broad variety of applications, more data are gathered, and over time greater insight is gained into how to select the best components and configure optimal dataflows for practical applications. If these fundamental tenets of system engineering can be applied to QA software development, it will become possible to relate investment in particular text processing technology with actual performance gains or losses in practical application.

To support this vision of shared modules, dataflows and evaluation measures, an open collaboration will include a *shared logical architecture* – a formal API definition for the processing modules in the QA system, and the data objects passed between them. For any given configuration of components, standardized metrics can be applied to the outputs of each module and the end-to-end system to automatically capture system performance at the micro and macro level for each test or evaluation.

### **3.2 Shared Deployment Infrastructure**

The logical architecture will be complemented by a shared run-time evaluation framework, which allows particular configurations of components to be formally evaluated on particular problem instances. Another important advantage of a shared logical architecture is that it enables the development of parallel, distributed computing infrastructure for large-scale deployment of system components. By sharing the same logical architecture, all teams will be able to take advantage of distributed computing frameworks and resources provided by other partners.

IBM and CMU have begun to collaborate on building an open-source framework for OAQA. Elements of the framework that are contemplated or already in development include:

- A Java implementation of a logical architecture for OAQA (data structures, components, configurations, resources, etc.);
- Tools to configure QA components into a particular end-to-end configuration;
- Tools to manage the execution of end-to-end configurations on distributed commodity hardware;
- Tools to capture execution results, automatically evaluate them against known answer keys, and compare current performance to past performance at the module and system level.

IBM is also considering providing access to specialized high-performance computing resources for the open advancement of QA. Of particular interest are map-reduce-style algorithms for retrieval, answer extraction and answer validation on massively parallel systems like BlueGene.

### **3.3 Challenge Problems**

Each QA application domain has certain characteristics related to the difficulty of the language(s), resource structure, questions, expected answers, etc. More general algorithmic approaches are less sensitive to variations in these characteristics, and can be adapted to new problem instances more readily; more language- and domain-specific approaches are much more sensitive to variations in language and resource structure, and cannot be easily adapted to completely new problem instances.

An effective *challenge problem* is one which cannot be solved without a significant improvement in QA system performance along some dimension of measurement (such as question difficulty, answer confidence, answer accuracy, system response time, etc.). By formulating a series of challenge problems across different information domains, it will become possible to drive core research and development and enhance domain adaptability in the longer term. Section 4 contains a detailed discussion of the different dimensions of a challenge problem, with examples drawn from particular QA problem domains.

### **3.4 Collaborative Process**

During the workshop, we identified three main areas for collaboration across industry, government and academia:

- **Challenge Problems & Metrics.** Challenge problems that drive the innovation to provide better business solutions, and the metrics that measure progress against those challenges, should be defined by representatives from all three groups.
- **Core Technologies.** Identifying and standardizing the use of different core technologies (for example, document retrieval, NLP tools) will help to advance the entire field of research, as well as drive forward our ability to deliver higher quality, lower-cost software solutions.
- **Infrastructure and Evaluation.** By designing and building a shared infrastructure for system integration and evaluation, we can reduce the cost of interoperation and accelerate the pace of innovation. A shared logical architecture also reduces the overall cost to deploy distributed parallel computing models to reduce research cycle time and improve run-time response.

All three areas will involve representatives from industry, academia and government, but it is expected that academic partners will wish to focus on technology research and evaluation, industrial partners will support frameworks for integration and evaluation,

and external sponsors will provide guidance regarding appropriate challenge problems and how they map to application systems that will meet current operational goals.

The open collaboration will include a variety of shared activities; each activity will be associated with a shared process that is agreed upon by the collaborators. The shared activities will include: a) defining a shared logical architecture; b) defining a set of common metrics for modules and end to end systems; c) establishing a framework to support integration and evaluation of end-to-end systems and modules; d) defining a set of challenge problems to be undertaken; and e) defining a shared process for ongoing integration, testing and refinement. These activities are discussed in more detail in Section 5.

### **3.5 Open Collaboration Model**

Open advancement is based on *transparency of evaluation* and *reproducibility of results*. The collaborative process will best be served by an open-source development model, where newly-created code is placed into open source for direct examination by all partners. This will help to ensure that the work is understandable, repeatable and reusable. Examples of successful open-source projects that have helped accelerate the development and application of key technologies include Lemur (document indexing and retrieval with language models), Nutch (web search engine), Hadoop (distributed file system with support for map-reduce), Lucene (general-purpose Java search engine) and UIMA (a framework for content analysis and automatic annotation), just to name a few.

Pre-negotiation of open-source teaming arrangements also helps to expedite research contracting with academic institutions. The role of open-source development is discussed in more detail in Section 6. A compelling example of the success of this model is the annual RoboCup competition, which evaluates the soccer-playing skills of robot players in an open international forum<sup>2</sup>. RoboCup has steadily reduced the gap between robot and human performance over the past 10 years, in part because all participants are required to publish their code after each competition [Asada, et al, 2007]. Conversely, a closed model, in which competitive evaluations are conducted without a requirement for openness, reproducibility, and transparency, has hampered scientific progress in QA.

## **4 Challenge Problems**

One of the principal goals of the OAQA is to collaboratively develop a code-base of QA technologies that generalizes over time and may be effectively applied to a broadening collection of QA problems. This goal is driven by the concern that historically, QA systems have been built for a single, specific task and often lack clear and repeatable methods for generalization and adaptation to new domains and problems. We believe that developing integrated end-to-end systems from a common collection of reusable component technologies, in order to target an explicitly crafted set of **Challenge Problems**, would help ensure more general solutions and greater adaptability.

---

<sup>2</sup> <http://www.robocup.org>

**Challenge Problems** are applications of QA technologies designed to test and drive these technologies along key performance dimensions, important for a wide range of business applications of QA. **Performance Dimensions** include measures like Accuracy, Speed, Question Difficulty, Usability, Breadth of Domain etc. We describe a collection of Performance Dimensions below. Performance Dimensions should ultimately be associated with well-defined metrics and evaluations for any given set of challenge problems.

**Selecting good challenge problems:** In addition to performance dimensions and their metrics, useful challenge problems should captivate the imagination of the scientific community and help to persuade the broader community that advances in automatic question answering can lead to significant and positive impact on science, business and society.

**The Challenge Problem Set Hypothesis:** While, in isolation, a solution to a specific challenge problem may not address a real-world application in its entirety, collaboratively developed solutions to a balanced set of challenge problems will lead to a general QA capability that can be effectively adapted to address a range of business problems with predictable and manageable cost.

## **4.1 Performance Dimensions**

In this section we propose a set of Performance Dimensions. We do not, in this white paper, attempt to work out precise quantitative metrics for each of these dimensions, but we acknowledge that this work would be an important contribution for the proposed collaboration on the OAQA.

### **Accuracy @ N**

Accuracy @ N is the percentage of questions for which the QA system provides a correct answer in the  $M^{\text{th}}$  rank in the answer list, where  $M$  is  $\leq N$  for any given question set.  $N$  is typically 1 in classic QA evaluations.

A Challenge Problem whose success metrics required a high degree of accuracy would rank highly on this dimension whereas one that was tolerant to lower accuracy would rank lower.

Accuracy is generally important for building user confidence in the QA application and for reducing the amount of time users spend in vetting wrong answers. Accuracy at 1 is particularly important if the QA system were embedded as a subroutine in a larger system. In these cases a human would not be vetting a collection of alternative answers and their contexts or justification.

## **Confidence Accuracy**

Confidence Accuracy is the probability that the QA system is right about being right. This feature of a QA system assumes that the system provides a Confidence Score, for example, a number between 0 and 1 that indicates a self-assessed probability that the answer(s) it is providing is correct. The degree to which a high Confidence Score correlates with right answers gives the probability that the QA system is right about being right – its Confidence Accuracy.

Since not all answers will be right nor will they be equally justified by the available content, Confidence Accuracy is a critical feature in the general utility of QA applications. It gives the user or program calling the QA system a realistic indicator about the accuracy of its results. This enables the caller to make better and/or faster decisions. In the case of an embedded QA capability where an application is programmatically calling the QA system directly, Accuracy and Confidence Accuracy are critical metrics.

As Accuracy@1 approaches 100%, Confidence Accuracy becomes less important. However, even in the most optimistic settings, QA systems will not operate at 100% accuracy and Confidence Accuracy will be important in improving the utility of the QA system.

Good Confidence Accuracy implies that the system is developing accurate *internal* justifications for its answers. We distinguish *internal justifications* from *external or user justifications*. The former may be effective at predicting the correctness of an answer independently of how it would be understood or evaluated by an end user.

*External or user justifications* are ultimately judged by the user as an acceptable explanation for the correctness of an answer. Although we consider the provision of good user justifications to be a feature of system Usability, a QA system may also consider the presence or absence of external justification(s) as a parameter when calculating answer confidence.

## **Broad Domain**

Broad Domain refers to the breadth of topics for which the QA system can achieve acceptable levels of accuracy.

We do not propose here a formal definition for Broad Domain. However, one can imagine questions where the topics cover a wide variety of domains of knowledge ranging from history and geography to science and medicine to literature, pop culture and current events.

We also consider a Broad Domain to imply the scope of the domain is not bounded, but rather that it is “open”. This implies that the types of things that maybe be asked about in a Challenge Problem that ranks highly in this dimension is considered open-ended and not limited to an *a priori* fixed set of concepts.

While we do not propose precise metrics for Broad Domain in this white paper, we can imagine that the breadth of a domain may be measured by the number of specific concepts asked about. The openness of the domain may be measured by the rate in which new concepts grow with the size of the question sets.

A challenge problem that focuses narrowly on answering questions about a single topic from a single text would rank low on this dimension, a challenge problem that admitted questions on any topic would rank high.

### **Question Difficulty**

Question Difficulty refers to the complexity of inference required to determine and justify answers from the available content.

The complexity of inference may range from simple string matching to complex logical inference that relies on a formalization of the content and a precisely constructed domain theory of axiomatic knowledge.

Question difficulty is relative to the dataset used and in most cases cannot be assessed a priori. Moreover, a question that is “easy” against one dataset may be “hard” against another one. In general, known-answer extraction should be considered easier than when an answer needs to be derived or assembled from separate pieces of information retrieved from the dataset. Ideally, a QA system could achieve a high level of accuracy for difficult questions over a broad domain, however, practically and in the near term we expect a tradeoff between Broad Domain and Question Difficulty.

In other words, challenge problems that require complex reasoning and inference to answer difficult questions will likely be more effective over narrow domains defined by specific, relatively limited volumes of content. This tradeoff is precipitated, in part, by the human knowledge engineering effort required to formalize the domain theory and axiomatic knowledge required to support complex inference.

A desirable property of a QA system is that as richer knowledge representations are added, the ability to handle more difficult questions increases without reducing the system’s domain breadth.

### **Query Language Complexity**

Query Language Complexity refers to the ambiguity and structural complexity of the questions in a QA problem. It does not refer to the formal computational complexity or expressivity of the language. Rather, it relates to the difficulty of extracting the intended meaning of the question from its linguistic expression.

So for example, challenge problems that require answering ambiguous, grammatically complex natural language questions would rank higher along this dimension. Challenge

problems that require the user to express queries in an unambiguous formal language like SQL or SPARCL or admit only questions that instantiate pre-determined templates would rank lower on this dimension.

A challenge problem that requires answering ill-formed natural language questions or cross-lingual natural language questions would rank higher on this dimension.

Solutions that dialog with the user to map from the user's expression to a formal unambiguous internal representation or that ensure a correct interpretation of a complex user query would score higher in this dimension and likely do well in terms of usability as well. Furthermore, a solution that took this approach might also score higher on the accuracy dimension, given that it is more likely to determine the right interpretation of the query and therefore get the right answer(s). In interactive QA contexts, the need for interactive refinement of the information need to improve accuracy must be balanced against a measure of the overall usability of the system.

### **Content Language Complexity**

Content Language Complexity is similar to Query Language Complexity but varies independently in Challenge Problems. Challenge Problems defined over well-formed formal language representations of content would rank lower on this dimension. Challenge Problems that must deal with natural language blog data or email data for example may rank higher on this dimension.

Note that there is less opportunity to engage the user in interpreting the content than in interpreting questions, given that the volumes are huge and it is the principal task of the QA system to automatically interpret the content in order to find the most likely answers and their justifications.

### **Speed (Response Time)**

Speed refers to the time required to answer a question. A Challenge Problem may range from requiring sub-second response time to batch oriented problems that may allow for 10's of minutes if not hours per question.

A Challenge Problem that required very fast response time would rank high along this dimension. Fast response is ultimately a relative measure. For example, it may be judged relative to how long an average user would take to answer a given set of questions with similar levels of accuracy and confidence but without the use of a QA system.

Consider the case where a typical user with the equivalent of an off-the-shelf text search engine took an average of 2 minutes per question on some test set and the QA system took on average 20 minutes per question over the same set with similar levels of accuracy. In this case the QA system's response time would be judged poor. If on the other hand, the QA system answered with an average of 2 seconds per question on the same test set

with the same or better accuracy as the human, then the QA system would be clearly adding some value.

The impact of a significant improvement in speed on any particular business application is yet another question. For example, in a batch research-oriented workflow it may not matter, while in a real-time technical support scenario the rapid response time may be the primary deciding factor for the application of the technology.

### **User Interaction/Usability**

This performance dimension is intended to describe the degree of user interaction required to succeed on the challenge problem. This dimension can include a mixed bag of important metrics.

Most business applications will require features like visualizations/explanations of the query, interactive query refinement, alternative answers, answer contexts, user-acceptable justifications etc. Along this dimension, challenge problems might also address managing the overall information-seeking process of hypothesis generation, query formulation, answer/result management, hypothesis testing and prediction generation.

A Challenge Problem that only requires “question in/answer out” would rank lowly on this dimension. A Challenge Problem whose solution required performing a range of interactions with the user would rank highly on this dimension.

### **Additional Performance Dimensions to Consider**

In the sample challenge set profile presented below, the performance dimensions listed above are considered. We expect that the list of performance dimensions will be subject to further refinement as new dimensions are proposed and different measures for existing dimensions are explored. The following additional measures have already been noted for future consideration:

- **Answer Specificity / Complexity.** A challenge problem could be measured based on the specificity and multiplicity of the answers required (document(s), passages(s), sentence(s), word(s)) and/or the complexity of what counts as an acceptable answer (e.g., a cause-and-effect question like “What caused the Sudanese civil war?” might best be answered by a set of sentences that describe an interlocking set of events). Note that simple questions can require very complex answers.

## **4.2 Sample Challenge Set Profile**

A **Challenge Set Profile** scores a **set** of challenge problems in terms of the relative degree to which they would test QA solutions along a key set of performance dimensions.

A challenge problem’s scores are meaningful within the set, since the scores for any one challenge problem are intended to be relative to the scores considered for the other

problems in the set. These profiles are qualitative and subjective, but none-the-less help to describe and communicate a relative assessment of different Challenge Problems.

A research program should develop crisp metrics for each of its challenge problems and plot solution performance with respect to these metrics along the same performance dimensions used to judge and select the challenge problems themselves.

The remainder of this section is an exercise in developing a sample Challenge Set Profile to explore the value of assessing different challenges along the performance dimensions we discussed above. For each problem we consider, we assign a score to each performance dimension with a number between 0 and 10. We plot the results on “radar graphs” to facilitate a quick visual comparison of the scope of the different challenge problems in the set. Additional work would be required under the OAQA program to more completely and formally describe and rationalize a well-balanced challenge problem set.

We describe five challenge problems:

1. TREC QA
2. TAC QA
3. Jeopardy!
4. Learning By Reading
5. Sustained Investigation

## 4.2.1 TREC QA

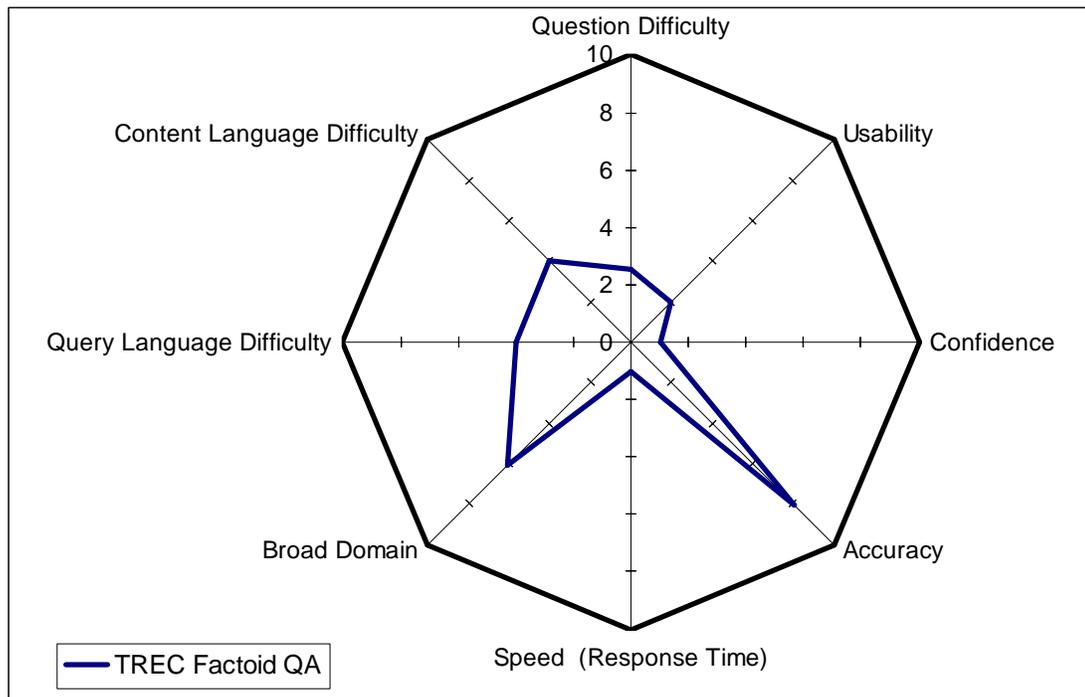


Figure 1: TREC QA Challenge Profile

In the TREC QA problem, the task is to answer 500 unseen natural questions from a primary source corpus of roughly 1 million news articles in one week's time, where the primary source had been available for possibly years. The QA system does not interact with a user. It provides one or more answers per question (more than 1 for list questions or in some cases a ranked list of 5 top answers). A solution to TREC QA is not prohibited from using other sources but the answer provided must ultimately come from the primary source.

1. **Query Language Difficulty:** The questions are free form natural language but are fairly simple in their expression suggesting a relatively low score in Query Language Difficulty.
2. **Content Language Difficulty:** The primary source is English newswire text stored in structured XML documents. So the content language difficulty is medium to low.
3. **Question Difficulty:** Most if not all questions may be answered directly by the primary corpora and do not require synthesis or complex inference which would suggest a relatively low score in Question Difficulty.
4. **Usability:** The TREC QA problem is batch oriented. It requires that a solution provide textual answers to each question but does not allow any interaction with a user, so its Usability score is low.
5. **Accuracy:** Accuracy is the primary metric in the TREC QA problem. Generally speaking, the more answers judged correct, the better the QA system scores

relative to others in the challenge. The TREC QA problem ranks relatively high in this dimension.

6. **Confidence:** With the exception of one year's use of the Confidence Weighted Score, the TREC QA problem generally does not require an accurate confidence score for a solution to excel and therefore ranks lower on this dimension.
7. **Speed:** The TREC QA problem is batch oriented and systems may utilize up to a week's time to answer 500 questions; this problem ranks lower on the Speed dimension.
8. **Broad Domain:** The TREC QA problem does not in theory limit itself to any topic or domain and as a challenge problem should score highly along this dimension. The caveat here is that in practice, a historical analysis of the past TREC QA question sets might reveal a broad but limited set of concepts. A formal analysis with more precise metrics might better assess the actual domain breadth of past TREC QA problem instances.

#### 4.2.2 TAC QA

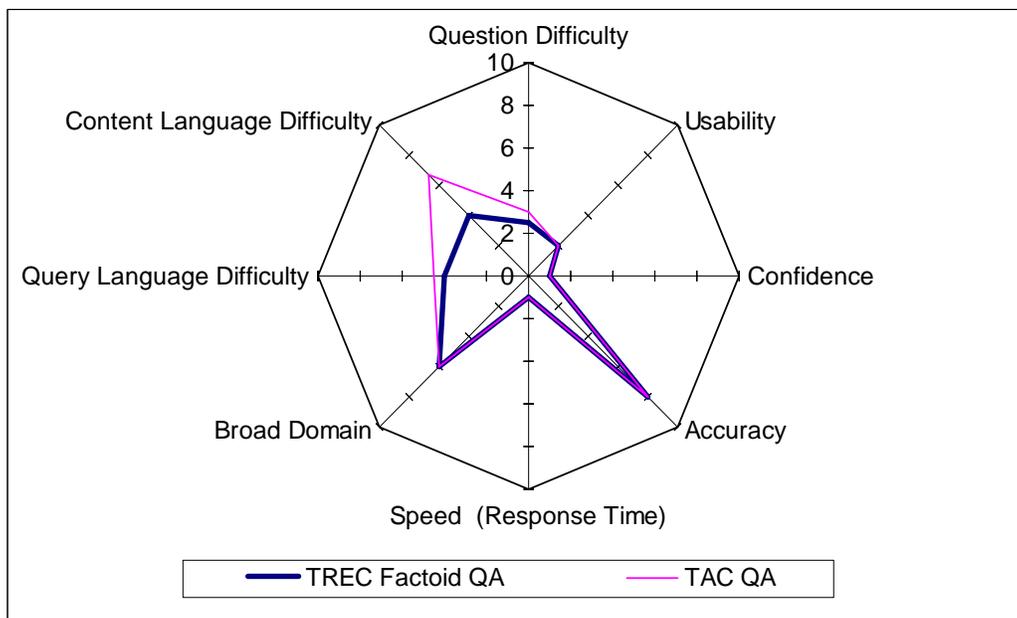


Figure 2: TREC QA and TAC QA

As we currently understand the TAC QA (<http://www.nist.gov/tac/tracks/2008/qa/>) problem (to be evaluated for the first time in the context of TREC 2008), the task is to answer 500 unseen natural questions from a primary source corpus of roughly 3.2 million permalink articles in one week's time, where the primary source has been available since 2006. Illustrated in Figure 2, TAC will differ from TREC primarily in the difficulty of the questions, which will include asking about lists of sentiments or opinions, and in the

content language difficulty which will be directed toward blog posts rather than newswire text.

As in the TREC problem, the QA system does not interact with a user. It provides one or more answers per question (more for “list questions” for example). It is allowed to consider other sources but the answer must ultimately be shown to come from the primary source.

1. **Query Language Difficulty:** The questions are fairly simple in their expression which would lead to a relatively low score in Query Language Difficulty.
2. **Content Language Difficulty:** TAC is significantly more difficult than TREC in terms of Content Language Difficulty. The source documents are threaded blog posts, which have more complex structure than newswire texts. The content of blog posts is also written in colloquial language, and is often telegraphic and/or ungrammatical. As a result, TAC ranks more highly on this dimension.
3. **Question Difficulty:** The questions in TAC Challenge Problem will be more complex than those in TREC QA. They will require identifying spans of text that represent opinions, opinions targets, and opinion holders; question may ask for lists of opinions, targets or holders as well. As a result, TAC ranks more highly on this dimension than TREC.
4. **Usability:** The TAC QA Problem is batch oriented. It requires that a solution provide textual answers to each question but does not allow any interaction with a user, so its Usability score is low.
5. **Accuracy:** Accuracy is the primary metric in the TAC QA problem. The more answers judged correct the better the QA system ranks relative to others in the challenge. The TAC QA problem ranks relatively highly in this dimension.
6. **Confidence:** Like TREC, the TAC QA problem generally does not seem to require an accurate confidence score for a solution to excel and ranks lower on this dimension.
7. **Speed:** The TAC QA problem is batch oriented and systems may utilize up to a week’s time to answer 500 questions; this problem ranks lower on the Speed dimension.
8. **Broad Domain:** The TAC QA problem does not in theory limit itself to any topic or domain and as a challenge problem should score highly along this dimension. The caveat here is that in practice, an analysis of the corpus might reveal a broad but limited set of concepts that are likely to be the focus of test questions (e.g. consumer products and services).

### 4.2.3 Jeopardy!

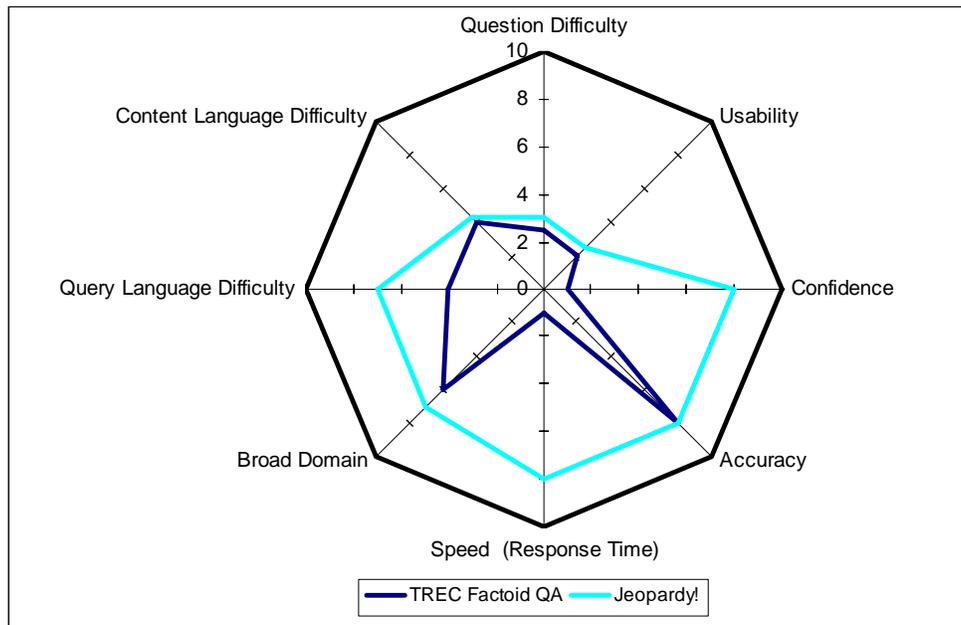


Figure 3: Jeopardy! and TREC QA

In the Jeopardy! Challenge Problem, the Jeopardy! quiz show is used as a model for integrating a set of metrics into a single QA challenge problem. The task may be simply stated as winning a series of Jeopardy!-style games against a real or simulated champion human player. Succeeding on this challenge requires a combination of high degrees of accuracy, speed and confidence over a broad domain. Moreover, Jeopardy! questions often contain additional details beyond the core question. These details may help find the answer, or they might enhance the entertainment or educational value of the question but not lead directly to the answer. Distinguishing the core question from the tangential details is a hard challenge for an automatic QA system, and requires sophistication in natural language parsing.

Jeopardy! provides a QA evaluation framework for additional dimensions of performance and their associated metrics. The game dynamics require high degrees of confidence since the final metric (i.e., winning the game) requires that the system avoids losing questions it chooses to answer, since there is a stiff penalty for getting a question wrong. There is also a requirement to answer interactively and in less than 3-5 seconds. Very high levels of performance will require considering the answers of other players and adjusting confidence levels and follow-up answers according. This may be considered a primitive form of dialog or collaboration. By varying the style and domain of the questions, the same framework could be used to evaluate collaborative and/or competitive question answering in non-Jeopardy! domains.

In addition, The Jeopardy! challenge has significant potential to capture the imagination and interest of a broader community because of the long-standing public popularity of the

Jeopardy! quiz show in the U.S. and its general perception as a challenging test of human intelligence (or at least breadth of knowledge). Reaching a larger audience, this challenge problem has the potential to stimulate a broader dialog on automatic Question Answering and a deeper appreciation for its potential.

In comparison to the TREC QA problem, the Jeopardy! problem broadens the challenge along key dimensions as shown in the radar graph in Figure 3.

1. **Query Language Difficulty:** The questions in the Jeopardy! challenge problem are generally more complex than those in TREC QA. They are longer and contain richer language demanding more of parsing technology. They contain multiple cues about the answer as well as tangential information demanding deeper semantic analysis to distinguish reinforcing predicates from irrelevant information.
2. **Content Language Difficulty:** Jeopardy! ranks similar to or a bit higher than TREC QA on this scale as the necessary sources to achieve higher degrees of accuracy and justification/confidence are broader and more varied. These include general web, news, dictionaries, encyclopedias and hosts of other sources.
3. **Question Difficulty:** The Jeopardy! Challenge is similar to the TREC QA challenge with respect to question difficulty. There are some questions that require deeper reasoning, and there are a small percentage of list questions. The majority of questions are factoid – these are questions that may be answered by one or more entities that satisfy the constraints expressed in the natural language question.
4. **Usability:** Jeopardy! ranks a bit higher than TREC QA on this dimension. The interaction is largely confined to question in and answer out; it is however interactive rather than batch, and reacting to other players' answers is required for high-levels of performance. Also, it should be noted that because the challenge requires a high degree of Confidence, the QA systems generation and use of some sort of justification is implied. This is internal justification, however, and its suitability for consumption by the user is not directly tested by the challenge. This is in contrast to the LBR challenge problem (see below) which requires, for example, humans to judge explanations well in order to succeed.
5. **Accuracy:** Accuracy is a primary metric for success in this challenge as it is in TREC QA. It focuses on Accuracy@1-3, since at most the system will have three chances to deliver the correct answer. Accuracy at 1 is strongly favored.
6. **Confidence:** Jeopardy! requires a very high degree of confidence, higher than TREC QA or the other challenge problems in the set since the final metric (winning the game) requires that you do not lose on questions you choose to answer. The QA system must compute an accurate confidence in less than 3-5 seconds.
7. **Speed:** Speed (interactive question answering response time) is critical in this challenge problem. The system must answer in less than 5 seconds. So this problem set ranks highly on this dimension.
8. **Broad Domain:** The domain in Jeopardy! is arguably broader than TREC QA, requiring access a larger variety of knowledge that goes beyond what might be

found in a collection of 1 million news articles. We ranked it higher on this dimension.

#### 4.2.4 Learning by Reading

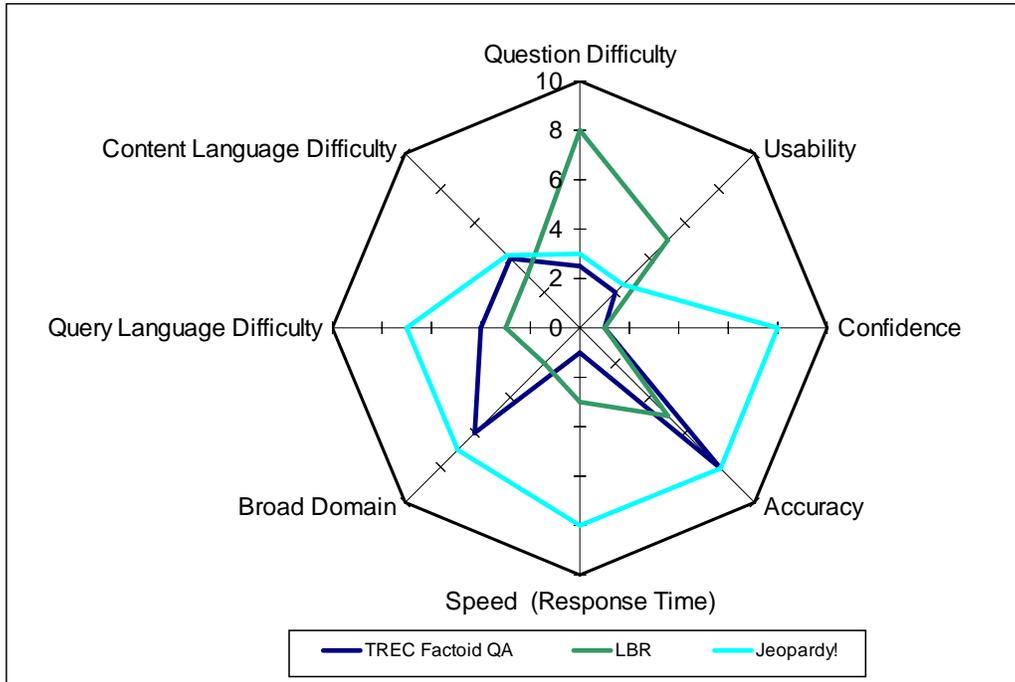


Figure 4: LBR, TREC QA and Jeopardy!

The Learning by Reading (LBR) challenge problem discussed here is inspired and based on our experience with HALO and the LBR pilot DARPA project. A formal LBR program is currently under development by DARPA. We do not intend to represent what the formal DARPA LBR program may become but rather to use our experience with the HALO and the LBR pilot to propose another interesting and useful QA challenge problem.

In the LBR challenge problem, a subject matter text book (i.e., the *target text*) is selected. This may be, for example, an in depth text on Chemistry, Biology or an even more specific subject like Cardiology. The QA system is built to answer complex questions about the material in the text and must provide answers as well as human understandable justifications. This challenge problem addresses two key dimensions that are weakly addressed by the other problems discussed so far, namely **Question Difficulty** and **Usability**.

Note that the intended goal of LBR is to demonstrate a system can “learn by reading”. However, this is demonstrated by an LBR solution’s overt behavior to answer users’ questions. LBR solutions must answer questions that require inference over the logical content of the target text. It is expected that answers can not simply be derived by linguistic inference, but rather would require some domain theory and the ability to map

from text into a domain theory, apply axiomatic rules, for example, and infer answers that are not directly represented in the text even with the use of different language. For this reason, the LBR challenge ranks significantly higher on the Question Difficulty performance dimension.

1. **Query Language Difficulty:** Assuming the LBR challenge uses template questions it scores lower than the other challenge problems discussed.
2. **Content Language Difficulty:** LBR targets a single well written natural language text. It scores marginally lower than TREC QA and Jeopardy! given the greater variety of language they must deal with and commensurately lower than TAC QA which emphasizes dealing with blogs. Note that the subject matter for the LBR challenge is likely harder and the ability to assimilate the information effectively to meet the challenge problem's success criteria is perhaps better reflected in the question difficulty dimension where LBR ranks very high.
3. **Question Difficulty:** The LBR challenge ranks very high relative to the others in this dimension.
4. **Usability:** The requirement for human-judged explanations ranks the LBR challenge higher than the others in this dimension.
5. **Accuracy:** Accuracy is not as critical in this challenge problem as it is in the others, primarily because it leans toward a more interactive experience with the user and provides deep explanations of multiple alternatives.
6. **Confidence:** The user interaction in the LBR challenge allows for the user to consider multiple answers and their explanations and therefore is more tolerant to lower confidence accuracy than for example the Jeopardy! problem. Its scores equal with TREC and TAC QA on this dimension but scores much higher than all the other challenge problems in the Usability dimension where explanations play a larger role.
7. **Speed:** The LBR challenge problem must provide answers as part of an interactive session and therefore ranks higher than TREC and TAC QA but does not demand the rapid response times that the Jeopardy! challenge problem requires.
8. **Broad Domain:** The domain for the LBR challenge is relatively narrow focusing on a specific subject represented by a specific text.

LBR solutions must provide explanations for their answers; to earn a high evaluation score, these explanations must earn the best human user judgments. Arguably, none of the challenge problems we discuss in this paper requires a level of Usability that is realistic for most real-world end-user applications of QA. A good Challenge Problem Set should include a member problem that does a better job at stressing the Usability dimension.

## 4.2.5 Sustained Investigation

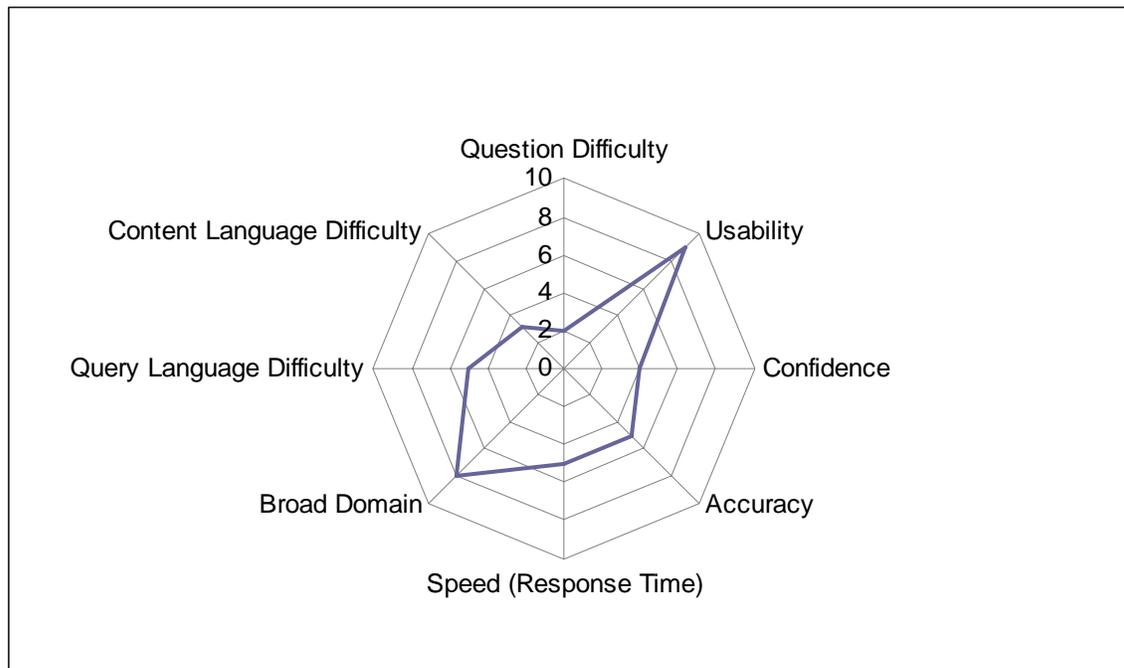


Figure 5: **Sustained Investigation**

In this challenge problem, the user's task requires more than answering a single question in isolation. A variety of applications and circumstances require the user to prepare an in-depth report while researching a complex topic from different perspectives. When preparing a detailed intelligence report or a comprehensive market assessment, the user (an intelligence or business analyst) must make deductions or provide logical conclusions based on a set of collected data. This entails a series of questions asked by the user as part of an ongoing dialog with the system. The system therefore is called upon to act as a guide through the data discovery process.

An interactive question answering system is tasked with understanding the user's information need, which may indeed be unclear when the user begins their task, becoming clearer only as the user navigates through the available data. The system should negotiate with the user regarding the relevance of associated concepts that may be connected to the concepts presented in the query, and assist the user in completing the task. The system operates to support and sustain the user's investigation until the information need has been satisfactorily met. As a consequence, this challenge problem stresses the Usability performance dimension much more than the other challenge problems we have described thus far.

1. **Query Language Difficulty:** Questions are in the form of natural language queries. Query language difficulty thus is assigned a low score.
2. **Content Language Difficulty:** The primary data source is a collection of articles in free form English. So the content language difficulty is medium to low. However, there is no specific limitation imposed on the data sources; in fact, we can consider the open web as a source, thus this dimension could be high for

specific challenge problems using the web or similarly chaotic, unstructured sources (such as blogs).

3. **Question Difficulty:** Questions are presented in the form of natural language queries. While some questions may focus on specific answers to factual questions, other questions will be more general in nature, in an attempt to gauge the extent and depth of the available data. Question difficulty is thus assigned a medium score due to the unpredictable nature of the questions anticipated. In general, based on observations from analytic tryouts, the questions asked during interactive sessions can vary from quite simple (e.g., factual recall) to very complex (e.g., hypothesis validation).
4. **Usability:** The Usability score is quite high, as the level of user interaction required is high. On this dimension metrics such as user satisfaction, task completion, user's confidence in results, effort expended, and various subjective measures of interaction adequacy are relevant.
5. **Accuracy:** The notion of accuracy must be defined carefully for this task, as not every user input represents a single question with a single, well-defined correct answer (or set of answers). For example, dialog events that represent a clarification between the system and user may be better evaluated by an Appropriateness measure (see 5.1 below). Nevertheless, the user's information need must be met by some high proportion of the system's responses for it to be judged accurate, so the score on this dimension is medium to high.
  - a. **Appropriateness** – in addition to accuracy, there is a need to assess the appropriateness of system responses and in this dimension the performance must be high. In other words, even if interaction loosens the requirement of strict accuracy for each system response, the response given must still be appropriate from the user view point (e.g., correcting a misconception about the data contents) as well as from the overall task view point (e.g., advances the user toward task completion).
6. **Confidence:** This task does not require computation of confidence scores and ranks lower on this dimension. However, since interaction may also involve a visual rendering of an answer space, a rough assessment of system's confidence may be useful in streamlining the interaction, e.g. via color coding of answers.
7. **Speed:** While it may not be essential for the average response time to be less than 5 seconds, users expect a reasonable turn-around time. Speed is scored low to medium importance when measured over an interactive session. Nonetheless, each interaction exchange should be quite fast, with responses in 3 seconds or less considered appropriate in order to sustain viable dialogue with the user (based on user studies within the Intelligence Community).
8. **Broad Domain:** The domain in Sustained Investigation is broad, entailing detailed information on a variety of topics. It is ranked high on this dimension.

### 4.3 Challenge Problem Set Integration and Generalization

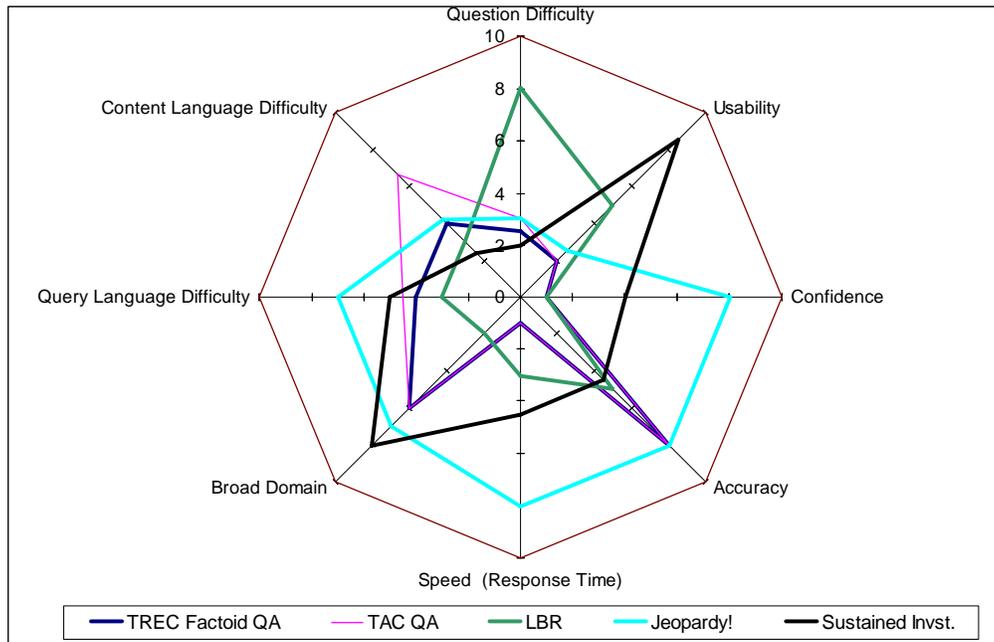


Figure 6: Challenge Set Profile

The composite radar graph for the sample challenge problems we discussed in this paper is illustrated in Figure 6. It shows that collectively the problems cover a much larger area of the graph than any one problem alone. It shows that TREC QA is subsumed by the other challenge problems.

The scores on this Challenge Set Profile are intended to be relative, the gap left between the highest scoring problem and the boundaries of the radar graph along any performance dimension are meant to suggest that none of these challenge problems maximally address any of the proposed performance dimensions. In this discussion we did not explicitly attempt to describe ideal or maximal performance along any dimension. A few observations are worth noting, however.

**Query and Content Language Difficulty:** None of these challenge problems address dealing with anything other than text. Additional modalities including image, speech, music and video are out of the scope of this paper, yet we acknowledge that QA systems could deal directly in these modalities. Furthermore, these dimensions should be expanded to encompass trans-lingual QA: none of the challenge problems we discussed address the requirement.

**Usability:** Arguably one of this set's biggest gaps is in Usability, where there is an array of issues to explore in satisfying the requirements of a broad class of users ranging from technical support agents, to web self-service users to business and national intelligence analysts.

However one or more Challenge Problem Sets are crafted, ideally, and according to the **Challenge Problem Set Hypothesis** submitted in the beginning section, collaboration on a well-balanced Challenge Problem Set should lead to an integrated QA capability that can perform maximally on all the challenge problems in the set and adapt to new problems that demand similar profiles. This adaptation should occur with lower and more predictable costs than it might have for any independently developed solution.

Collaboration under OAQA should facilitate the transfer of QA technologies among the solutions within a Challenge Problem Set. For example, consider Confidence; if a solution to the Jeopardy! problem performs well at judging questions as correctly answerable from a given primary corpora, that technology should be transferred to solutions to other challenge problems within the set. Similarly, if a solution to LBR, for example, can answer complex questions over a narrow domain in Biology, it should be shown that that capability can be added to a system capable of high performance on a broader domain to result in a hybrid capability.

**Adaptability:** We can imagine describing **adaptability** as a performance dimension for *2<sup>nd</sup>-Order* challenge problems that require a solution to adapt in a limited amount of time/effort, to, for example, new source corpora, additional knowledge, a new or expanded domain, or an increase in question difficulty, while maintaining similar levels of accuracy and/or speed. We believe that *2<sup>nd</sup>-Order* challenge problems that directly address adaptability are necessary to ensure the generalization of QA technologies.

## 5 Approach to Open Collaboration

The open advancement approach is based on an iterative, collaborative research process with the following main use cases:

- **Establish a Shared Logical Architecture.** The collaborators share a common set of data object definitions and modular interfaces, which will be used in the construction of both individual text processing modules (which implement the module interfaces by consuming/producing standard data objects) and end-to-end dataflows. The collaborators also share a common framework for specifying and representing a QA problem domain (corpus, questions, answers, etc.).
- **Define Formal Metrics.** The collaborators share a common set of metrics which will be used to measure the performance of individual modules and end-to-end dataflows;
- **Define Challenge Problems.** The collaborators share a set of challenge problems using the common framework; a problem definition must include a description indicating how this challenge problem is intended to drive innovation along particular dimensions (e.g. metrics and measurements).
- **Design Experiments.** The collaborators share a common process for open advancement, which specifies the steps to be taken in configuring an experiment, conducting an experiment, gathering measurements, and reporting system performance based on those measurements.
- **Manage Development.** The collaborators follow the common process in advancing the state of the art on the selected challenge problem(s), while continuing to refine the processing modules, end-to-end dataflow(s), and the common development process itself continuously over time. Effort is invested in those component technologies and dataflows which provide the best improvements on performance for selected problems.

Collaboration on the open advancement of QA will make it possible for stakeholders to answer important questions from prospective users of the technology, for example:

- What is the best QA dataflow for my problem domain? Using which components?
- What's the expected upper bound on performance in my problem domain?
- How can performance improve if we invest in improving the technology?
- What's the overall cost of adapting and tuning current technology for my problem domain?
- What are the most important component technologies we should invest in, given current performance levels vs. targets?

The ability to answer such questions would certainly improve the perceived business value of ongoing QA research and development.

## **6 Open Collaboration Model**

A common issue that arises for externally-funded academic research is the assignment of intellectual property rights. Many universities seek to retain ownership of intellectual property created through external sponsorship, while sponsors expect to receive ownership or at least preferential access in return for their support.

The proposed approach for collaborative, open advancement addresses this issue directly by adopting an open-source model. All rights in software remain the property of the developing party, and software developed as part of the funded collaboration is released according to open-source licensing terms agreed upon in advance. A master agreement is negotiated with each academic partner, establishing the open collaboration agreement terms in advance. Individual research projects can be funded in the form of amendments to the master agreement, streamlining the paperwork involved to get the work funded at each cycle. IBM and Carnegie Mellon University have established an open collaboration agreement which satisfies these requirements.

Partners from industry, academia and government can collaborate in a variety of ways: by funding seedling projects or larger-scale research programs; by contributing technology; by consuming emerging technologies and testing them in real-world applications. For example, IBM has proposed to sponsor seedling projects, contribute technologies, and utilize technologies emerging from the open collaboration in broader business applications. Carnegie Mellon intends to both contribute and utilize technologies as part of ongoing research and development of end-to-end systems for external sponsors.

More specifically, IBM proposes to provide an open-source infrastructure (possibly including but not limited to data models and tools for system configuration, integration, testing and results analysis). Carnegie Mellon's QA research team is working to adapt both existing and emerging QA components to the open collaboration model. Goals for the coming year include building a collaborative system that allows extensive testing of different combinations of technologies on different QA problems.

IBM is also working to establish similar agreements with other academic research partners, in hopes of achieving broader collaboration.

## **7 Next Steps**

Possible next steps for the open advancement of QA include the following:

- Workshop participants and other academic organizations form a working group to propose collaborations on technology development, challenge problem design, etc.

- The working group continues to improve and refine this document for broader dissemination in the community (to get feedback about ideas, direction, funding etc.).
- The working group identifies possible sources for external funding (government, industry, etc.).
- IBM establishes seed funding for academia under the Open Collaborative Research (OCR) agreements to help demonstrate the potential of OAQA by working solving multiple challenge problems using a single extensible QA architecture, with intent to attract broader collaboration and potential follow-on funding. (As of December 2008, University of Massachusetts, University of Texas at Austin, Carnegie Mellon University and the University of Southern California have received grants to openly advance the science of Question Answering).
- IBM and CMU continue to develop emerging infrastructure for collaborative development, including new partners as they join the open collaboration effort.

## 8 References

Asada, M. et al. 2007. Middle Size Robot League Rules and Regulations for 2008. Published for the Robocup 2008 middle-sized robot competition. Available from <http://www.er.ams.eng.osaka-u.ac.jp/robocup-mid/index.cgi?page=Rules+and+Regulations>

Burger, John et al., 2003. “Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)”, unpublished manuscript, [http://www-lpir.nist.gov/projects/duc/papers/qa.Roadmap-paper\\_v2.doc](http://www-lpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc)

Maybury, M. 2002. Working notes of the Workshop on Question Answering: Strategy and Resources. May 28, 2002. In conjunction with the Third International Conference on Language Resources and Evaluation (LREC). Las Palmas, Canary Islands, Spain. 29-31 May, 2002. [www.lrec-conf.org/lrec2002/lrec/wksh/QuestionAnswering.html](http://www.lrec-conf.org/lrec2002/lrec/wksh/QuestionAnswering.html).