

# **Evaluation of Restricted-Domain Question-Answering Systems**

**Anne R. Diekema  
Ozgur Yilmazel  
Elizabeth D. Liddy**

**Center for Natural Language Processing  
School of Information Studies  
Syracuse University  
Syracuse, New York, USA**



# Goals of Presentation

- **Introduce a set of possible dimensions for evaluating restricted-domain QA systems**
  - Based on criteria identified by users of an operational restricted-domain QA system
- **Open a discussion on feasibility of these for common evaluation dimensions**
  - For both open-domain and restricted-domain QA systems



## What R-D QA Systems ARE NOT:

- **TREC-style QA systems**
  - General interest questions
  - Short, factoid answers
  - Questions classified into a few question-types to determine what answer-type is needed
  - Redundancy on web can determine best answer
- **Early toy restricted-domain QA systems**
  - Hand-encoded domain knowledge bases
  - Lack portability to new domains



## What R-D QA systems ARE:

- **Based to some extent on IR / QA principles**
- **Specialized for domain-specific needs**
- **Need to provide a wider range of responses**
  - Factual answers
  - Comparisons
  - Reasons
  - Conditioned answers
- **Task-oriented**



# Situatedness

- **QA within / for a specific:**
  - Domain
  - User community
  - Task
- **Where QA system must function:**
  - In real time, not batch mode
  - On real users' real world questions
  - With real, not surrogate assessments of relevance



# Situatedness

- **QA within / for a specific:**
  - Domain
  - User community
  - Task
- **Where QA system must function:**
  - In **real** time, not batch mode
  - On **real** users' **real** world questions
  - With **real**, not surrogate assessments of relevance



# Situatedness

- **QA within / for a specific:**
  - Domain
  - User community
  - Task
- **Where QA system must function:**
  - In **real** time, not batch mode
  - On **real** users' **real** world questions
  - With **real**, not surrogate assessments of relevance
- **Therefore, evaluation must reflect the situation**



# KAAS

- **Knowledge Acquisition & Access System**
- **Funded by NASA, New York State, AT&T**
- **QA system for use within a distance-based, collaborative learning environment (AIDE)**
- **For undergrad students from 2 universities majoring in Aerospace Engineering**
- **Students using AIDE for a course can ask questions while working in teams or on own**



## **KAAS** (cont'd)

- **Collection**
  - Textbooks, technical papers / reports, websites
  - Pre-selected for relevance and pedagogical value
- **Two-stage QA model**
  - 1<sup>st</sup> – passage retrieval using expanded query representation
  - 2<sup>nd</sup> – selection of answer-providing passages based on generic + specialized entities & relations
- **Answer Format**
  - Students & professors want support / justification



# Analysis of NASA Questions

- **Examined 406 questions asked by students**
- **Questions quite different from TREC queries**
  - Ambiguous
  - Contain many syntax and spelling errors
  - More complex
    - Reflect domain knowledge
    - Require long answers, perhaps from multiple documents



# Linguistic Features of Questions

- **Large number of domain-specific verb & noun phrases,**
  - including some Proper Noun phrases, but many more common noun phrases
- **Longer questions, more complex syntax, multiple prepositional phrases**
- **Focuses are complex, not necessarily singular**
- **Predominance of 8 major classes of questions**

<b>Wh-</b>	<b>Simpler factoid questions of the <i>what, when, where</i> type.</b>	<b><i>When was the idea of glass transition temperatures discovered?</i></b>

<b>Wh-</b>	<b>Simpler factoid questions of the <i>what, when, where</i> type.</b>	<i>When was the idea of glass transition temperatures discovered?</i>
<b>Yes / No</b>	<b>Require a yes or no answer, but may mask a complex inquiry.</b>	<i>Doesn't simplification of honeycomb design for the TPS of the RLV jeopardize accuracy of the results?</i>

<b>Wh-</b>	<b>Simpler factoid questions of the <i>what, when, where</i> type.</b>	<i>When was the idea of glass transition temperatures discovered?</i>
<b>Yes / No</b>	<b>Require a yes or no answer, but may mask a complex inquiry.</b>	<i>Doesn't simplification of honeycomb design for the TPS of the RLV jeopardize accuracy of the results?</i>
<b>How</b>	<b>Require an explanation.</b>	<i>How are layers in TABI bonded together?</i>

<b>Wh-</b>	<b>Simpler factoid questions of the <i>what, when, where</i> type.</b>	<i>When was the idea of glass transition temperatures discovered?</i>
<b>Yes / No</b>	<b>Require a yes or no answer, but may mask a complex inquiry.</b>	<i>Doesn't simplification of honeycomb design for the TPS of the RLV jeopardize accuracy of the results?</i>
<b>How</b>	<b>Require an explanation.</b>	<i>How are layers in TABI bonded together?</i>
<b>Quantification</b>	<b>Looking for a specific amount, such as cost, weight, volume, number, maximum.</b>	<i>What is the highest temperature the space shuttle undersurface gets to during its mission?</i>

<b>Conditional</b>	<b>Indicate a condition that the answer needs to take into account, as indicated by phrases such as: <i>in addition to, aside from, other than, etc.</i></b>	<i>Aside from contact of two tiles, are there other reasons why insulating tiles on RLVs should be insulated from one another?</i>

<b>Conditional</b>	<b>Indicate a condition that the answer needs to take into account, as indicated by phrases such as: <i>in addition to, aside from, other than, etc.</i></b>	<i>Aside from contact of two tiles, are there other reasons why insulating tiles on RLVs should be insulated from one another?</i>
<b>Alternative</b>	<b>User provides several alternatives, one of which needs to be proven true.</b>	<i>Are thermal protection systems commonly composed of one panel or a collection of smaller tiles?</i>

<p><b>Conditional</b></p>	<p>Indicate a condition that the answer needs to take into account, as indicated by phrases such as: <i>in addition to, aside from, other than, etc.</i></p>	<p><i>Aside from contact of two tiles, are there other reasons why insulating tiles on RLVs should be insulated from one another?</i></p>
<p><b>Alternative</b></p>	<p>User provides several alternatives, one of which needs to be proven true.</p>	<p><i>Are thermal protection systems commonly composed of one panel or a collection of smaller tiles?</i></p>
<p><b>Why</b></p>	<p><b>Require an explanation</b></p>	<p><i>Why all shear loads and twisting moments set to zero for the preliminary design phase of TPS?</i></p>

<b>Conditional</b>	Indicate a condition that the answer needs to take into account, as indicated by phrases such as: <i>in addition to, aside from, other than, etc.</i>	<i>Aside from contact of two tiles, are there other reasons why insulating tiles on RLVs should be insulated from one another?</i>
<b>Alternative</b>	User provides several alternatives, one of which needs to be proven true.	<i>Are thermal protection systems commonly composed of one panel or a collection of smaller tiles?</i>
<b>Why</b>	Require an explanation	<i>Why all shear loads and twisting moments set to zero for the preliminary design phase of TPS?</i>
<b>Definition</b>	Looking for a formal or semi-formal definition of an element, process, etc	<i>What is a liquid metal?</i>



# Attempted a TREC-Style Evaluation



# Attempted a TREC-Style Evaluation

- **But such evaluations are based on:**
  - Short, factoid test questions
  - Mined from question logs
  - Large, public test collection
  - Paid assessors
  - Adjudicated answers
  - System-comparable results



# Why this Doesn't Work in R-D QA

- Developing a realistic **question set** is not easy



# Why this Doesn't Work in R-D QA

- **Developing a realistic question set is not easy**
  - Tried with students - based on prior class projects, but later questions looked nothing like the synthetic ones



# Why this Doesn't Work in R-D QA

- **Developing a realistic question set is not easy**
  - Tried with students based on prior class projects, but later questions looked nothing like the synthetic ones
- **Establishing **answers** for question set is hard**



# Why this Doesn't Work in R-D QA

- **Developing a realistic question set is not easy**
  - Tried with students based on prior class projects, but later questions looked nothing like the synthetic ones
- **Establishing answers for question set is hard**
  - Had PhD students evaluate system's answers, but slow
  - Answer-patterns not standardized enough for automatic evaluation



# Why this Doesn't Work in R-D QA

- **Developing a realistic question set is not easy**
  - Tried with students based on prior class projects, but later questions looked nothing like the synthetic ones
- **Establishing answers for question set is hard**
  - Had PhD students evaluate system's answers, but slow
  - Answer patterns not standardized enough for automatic evaluation
- **Test collections not readily available**
  - Gathering, converting, determining coverage is costly



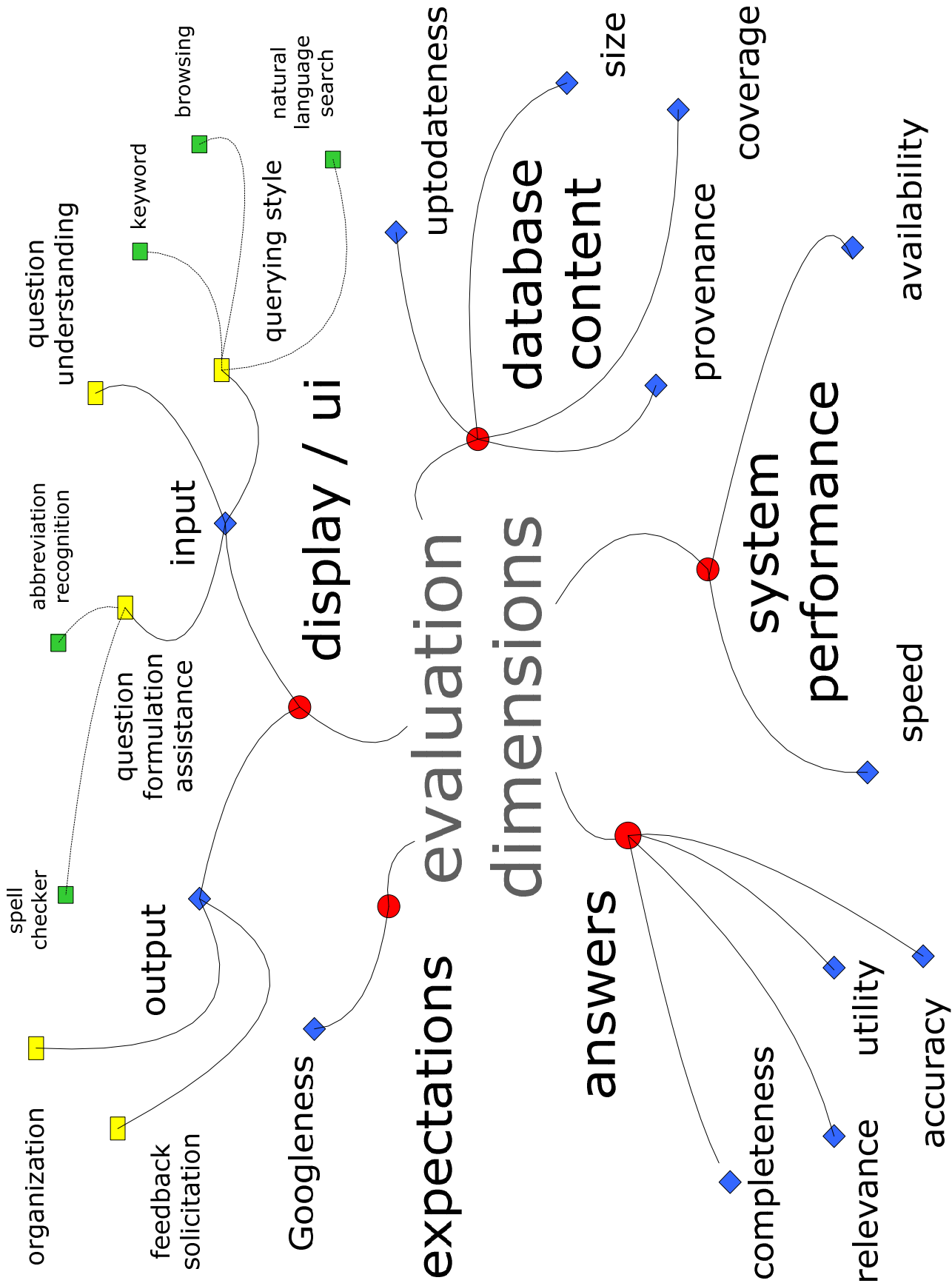
## Why this Doesn't Work in R-D QA

- **Developing a realistic question set is not easy**
  - Tried with students based on prior class projects, but later questions looked nothing like the synthetic ones
- **Establishing answers for question set is hard**
  - Had PhD students evaluate system's answers, but slow
  - Answer patterns not standardized enough for automatic evaluation
- **Test collections not readily available**
  - Gathering, converting, determining coverage, costly
- **Doesn't say anything about how the system would contribute to real users' tasks**
  - Contextual needs have major impact on user judgments



# User-based Evaluations

- **Conducted Surveys of KAAS Users**
  - Asking about their experiences with the AIDE
  - 25 to 30 students each for two semesters
  - Open-ended survey items on KAAS about:
    - Quality of the responses
    - Usefulness
- **Content analysis of responses by 3 researchers**
- **Identified a combined set of 5 major and 23 minor evaluation dimensions**





# User-Based Evaluation Dimensions

## 1. System Performance

- Speed
- Availability / reliability / upness



# User-Based Evaluation Dimensions

## 2. Answers

- Accuracy
- Completeness
- Relevance to question
- Applicability to task / utility / usefulness



# User-Based Evaluation Dimensions

## 3. Database Content

- Provenance / reliableness / accuracy
- Up-to-dateness
- Coverage / extensiveness
- Size



# User-Based Evaluation Dimensions

## 4. User Interface

- **Input**
  - **Question / information need understanding**
  - **Querying style**
    - **Natural Language search**
    - **Keywords**
    - **Browsing**
  - **Question formulation assistance**
    - **Spell Checking**
    - **Abbreviation Recognition**
  
- **Output**
  - **Organization**
  - **Feedback solicitation**



# User-Based Evaluation Dimensions

## 5. Expectations

- *Googleness*
  - Users' had little initial patience with a system that had a different look and feel



# Google-Fixation Fix

- **One Professor posed a short task in the AIDE for class teams**
- **Proceeded to:**
  1. **Demo 2 answer-finding approaches:**
    - **Google search**
    - **KAAS questions**
  2. **Measure the time it took to complete task**
  3. **Evaluate quality of the results with class**
  4. **Show how much each contributed to task**
- **Resulted immediately in dramatic increase in student use of KAAS**



# User-Based Evaluation Metrics

## 1. System Performance Metrics

- **ANSWER RETURN SPEED** - how long it takes to return an answer
- **UP-TIME** - how often the system is available



# User-Based Evaluation Metrics

## 1. System Performance Metrics

- **ANSWER RETURN SPEED** - how long it takes to return an answer
- **UP-TIME** - how often the system is available

## 2. Answer Metrics

- **ACCURACY or CORRECTNESS**
- **COMPLETENESS**
- **RELEVANCE**
- **TASK UTILITY** – whether the system's answer enables the user to complete their task



# User-Based Evaluation Metrics

## 1. System Performance Metrics

- **ANSWER RETURN SPEED** - how long it takes to return an answer
- **UP-TIME** - how often the system is available

## 2. Answer Metrics

- **ACCURACY** or **CORRECTNESS**
- **COMPLETENESS**
- **RELEVANCE**
- **TASK UTILITY** – whether the system's answer enables the user to complete their task

## 3. Database Content Metrics

- **SOURCE QUALITY** - Domain-knowledgeable evaluation
- **SCOPE** - Coverage metric of the field of interest



# User-Based Evaluation Metrics (cont'd)

## 4. User Interface Metrics

- **ADAPTABILITY** – range of forms of input accepted
- **ASSISTANCE** – degree of helpfulness of clarification & expansion provided
- **EASE OF USE** - amount of effort required from user



# User-Based Evaluation Metrics (cont'd)

## 4. User Interface Metrics

- **ADAPTABILITY** – range of forms of input accepted
- **ASSISTANCE** – degree of helpfulness of clarification & expansion provided
- **EASE OF USE** - amount of effort required from user

## 5. Expectation Metrics

- **EXPECTATION DIVERGENCE** – degree of discrepancy between user's expectancy & system
- **DEMONSTRABILITY** – ability to show extent to which system meets user's task needs



# Cross-Fertilization of Evaluations?



# Cross-Fertilization of Evaluations?

- **Separate / same dimensions for open-domain & restricted-domain QA systems?**



# Cross-Fertilization of Evaluations?

- **Separate / same dimensions for open-domain & restricted-domain QA systems?**
- **Which dimensions make sense for each / both?**



# Cross-Fertilization of Evaluations?

- **Separate / same dimensions for open-domain & restricted-domain QA systems?**
- **Which dimensions make sense for each / both?**
- **How to determine comparability of various restricted-domain system evaluations?**



# Cross-Fertilization of Evaluations?

- **Separate / same dimensions for open-domain & restricted-domain QA systems?**
- **Which dimensions make sense for each / both?**
- **How to determine comparability of various restricted-domain system evaluations?**
- **Cost / effort of realistic, task-focused restricted-domain QA system evaluations?**



# Task-focused Utility!

- **Key metric for KAAS users**
  - Are not QA researchers
  - Just want to do their work



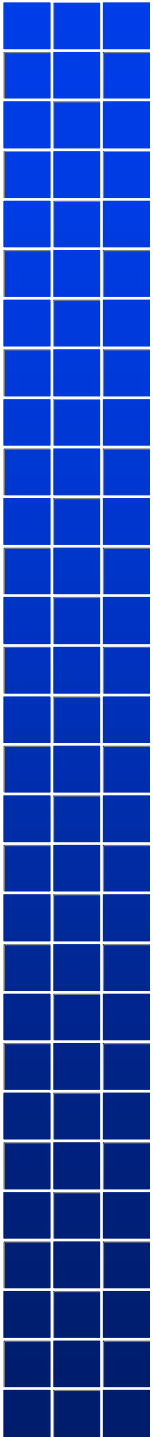
# Task-focused Utility!

- **Key metric for KAAS users**
  - Are not QA researchers
  - Just want to do their work
- **Task Scenarios to evaluate:**
  - System's contribution to users' tasks
  - Appropriate length & detail of answers



# Task-focused Utility!

- **Key metric for KAAS users**
  - Are not QA researchers
  - Just want to do their work
- **Task Scenarios to evaluate:**
  - System's contribution to users' tasks
  - Appropriate length & detail of answers
- **If tasks are different for individual R-D QA systems, how do we compare systems?**
  - Does comparability matter to users?



**Thank You!**

**Questions?**



# Sample Questions from Real Users

- *How difficult is it to mold and shape graphite-epoxies compared with alloys or ceramics that may be used for thermal protective applications?*
- *How does the shuttle fly?*
- *Do welding sites yield any structural weaknesses that could be a threat for failure?*
- *Are Thermal Protection systems of spacecrafts commonly composed of one panel or a collection of smaller tiles?*
- *How can the aerogels be used in insulation of holes in TPS?*