ALTSS Course

CENTRE FOR LANGUAGE TECHNOLOGY
MACQUARIE UNIVERSITY ~ SYDNEY

Information Extraction and Question Answering

Lecture 1

Introduction

Diego Mollá
diego@ics.mq.edu.au

Contents

Contents

Information Extraction – Reading

- Appelt & Israel (1999) *Introduction to Information Extraction Technology*.
  http://www.ai.sri.com/~appelt/ie-tutorial/
- Cowie & Lehnert (1996) "Information Extraction". *Communications of the ACM*, Vol.39, No.1, pages 80-91.
- Grisham & Sundheim (1996) "Message Understanding Conference –6: A Brief History". In *Proceedings MUC-6.*
- Hobbs *Generic Information Extraction System*.
  http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/gen_ie.htm

Question Answering – Reading

- Burger et al. *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)* http://www-nlpir.nist.gov/projects/duc/roadmapping.html
- Harabagiu et al. (2000) "FALCON: Boosting Knowledge for Answer Engines", in *Proceedings of the Text Retrieval Conference (TREC-9).*
- Lin (2002), "The Web as a Resource for Question Answering: Perspectives and Challenges", in *Proceedings LREC2002*
- Lin and Katz (2003), *Question Answering Techniques for the World Wide Web*. Tutorial presented at EACL 2003
- Moldovan et al. (1999) "LASSO: A Tool for Surfing the Answer Net", in *Proceedings of the Text Retrieval Conference (TREC-8).*
- Molla et al. (2003) "Answer Extraction from Technical Texts". *IEEE Intelligent Systems* 18(4):12-17.
- NLE (2001) *Natural language Engineering*, special issue on question answering. Vol. 7 Number 4.

Outline of This Lecture

- What is Information Extraction and Question Answering
- History of Information Extraction Systems
- History of Question Answering Systems
- Issues about IE and QA

The Motivation

- Most of the information in most companies and organizations is contained in text in human languages (reports, memos, email, web pages, …), and not in databases or similar structured formats
  - Estimate: about 70% [all depends how you measure, of course]
- Most of that information is now available in digital form:
  - Estimate: about 60% [CAP Ventures/Fuji Xerox, 1998]
- Most of the information in the World Wide Web is in form of HTML documents

Conclusion:

- You're missing out on a lot of good stuff if you can't get answers from all that digital information written in human languages

What is Information Retrieval?

- Retrieving information from document repositories
- Query-based IR
  - Document (ad-hoc) retrieval
    - Retrieve documents that are relevant to the user query
  - Passage retrieval
    - Retrieve passages that are relevant to the user query
  - Answer extraction
    - Retrieve the exact text that answers the question
  - Information extraction
    - Find all the useful information in the text
  - Question answering
    - Answer the user question

What Information Extraction is About

- The problem:
  - extract well-defined pieces of information from collections of documents
- The goal:
  - to populate a database
- Typically, most of the information in a document is ignored
- IE can be contrasted with earlier goals of building story understanding systems, where broad and deep coverage is needed

An Example Document

San Salvador, 19 Apr 89 (ACAN-EFE) -- [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

...

Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador.

...

Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle.

A Corresponding Filled Template

Incident: Date	19 Apr 89
Incident: Location	El Salvador: San Salvador (CITY)
Incident: Type	Bombing
Perpetrator: Individual ID	urban guerrillas
Perpetrator: Organization ID	FMLN
Perpetrator: Confidence	Suspected or Accused by Authorities: FMLN
Physical Target: Description	vehicle
Physical Target: Effect	Some Damage: vehicle
Human Target: Name	Roberto Garcia Alvarado
Human Target: Description	attorney general: Roberto Garcia Alvarado
Human Target: Effect	Death: Roberto Garcia Alvarado

Information Extraction: Inputs and Outputs

- The inputs to IE:
  - A knowledge source
  - A communicative or information need
  - Perhaps a user model
- The output of IE:
  - Structured, semantically explicit information, most commonly:
    - Attribute-value/Slot-filler templates
    - Database tables

The General Approach

1. Locate sentences or fragments that contain relevant information
2. Ignore information that is not relevant
3. Extract the information
4. Output information in a predetermined form

Information Extraction vs NL Understanding

- We don't care about subtleties in author's intentions
- We don't need to be able to answer general questions about the text
- We do need to be able to extract specific predetermined information from the text
- We can settle for a less expressive representation of the 'meaning' of the document (i.e., templates)

Task Specification for Information Extraction

- Compared to other NLP tasks, very well defined:
  - What domain information do we want to extract?
  - How do we encode the output information?
- We can easily compare computer and human performance on the same task

Target Applications

- Converting unstructured texts to databases
- Providing input to summarization systems
- Creating indexes for information retrieval systems
- Creating the basic elements to be returned by question answering systems

Question Answering Systems

1. Analyse the question
2. Consult resources
   - Knowledge base
   - Database
   - Text documents
   - The Web
3. Present the answer
   - Text form
   - Possibly with supporting material

Applications

- Based on the source of the answers:
  - Structured data (databases)
  - Semi-structured data (comments in database fields)
  - Free text
- Based on the supporting documents
  - Search over a single text (e.g. reading comprehension tests)
  - Search over a fixed set of documents (e.g. TREC)
  - Search over the Web
- Based on the domain
  - Domain independent
  - Domain dependent (e.g. help systems)

The User Factor

- Trained user
  - Can use specialised query languages
    - Boolean queries
    - SQL
- First time user, casual user
  - Needs to be informed about the system's limitations
  - Natural language queries
  - Simplified queries (e.g. keywords)
- Frequent user
  - The system can maintain a user model

Types of questions

- By answer type
  - Factoid answers
    - Wh_ questions
  - Summary
  - Opinion
- Yes/no questions
- Commands
- Specially difficult questions:
  - "Why", "how" (understanding of causality or instrumental relations)
  - "What" (little constraint on the answer type)
  - Definitions (the answer may depend on the user)

Evaluation

- What makes an answer good?
- Is answer justification good?
  - What happens if the answer is correct but the justification is wrong?
  - If there isn't answer justification, how would we know the system made a mistake?

Answer presentation

- Interactive system
  1. The user starts with a general question
  2. The user narrows the search, engaging in a "dialogue" with the system
- Use speech input
  - Conversational access to question answering systems

IR, IE and QA

- Information Retrieval
  - The term IR typically means "document retrieval"
  - Find the documents that are relevant to a user query
  - Use of bag of words approaches
  - Strength in numbers: statistical and corpus-based approaches
  - TREC: Text REtrieval Conferences

IR, IE and QA

- Information Extraction
  - Fill predefined templates from natural language texts
  - Templates indicate keyroles in stereotypical events
    - Corporate takeovers:
      - Acquired company
      - Date of acquisition
      - Amount paid
      - …
  - Use of pattern-matching techniques and domain-dependent information
  - MUC: Message Understanding Conferences

IR, IE and QA

- Question Answering
  - Return an answer to the user question
  - We cannot predetermine what the user is going to ask
  - Integration of information retrieval and information extraction
  - General approach:
    1. Analyse the question and determine the answer type
    2. Preselect passages that may contain the answer
    3. Process the passages to find answer candidates of the correct type
    4. Rank the answer candidates and return the most likely answer
  - Question Answering Track of TREC

Outline of This Lecture

- What is Information Extraction and Question Answering
- History of Information Extraction Systems
- History of Question Answering Systems
- Issues about IE and QA

History of Information Extraction Systems

- DeJong's FRUMP news monitoring system reported in late 1970s
  - Based on Conceptual Dependency
    - An early attempt to achieve full understanding of text
- 1980 DaSilva and Dwiggins extracted satellite flight information
- 1981 Cowie developed a system to extract information on plants and animals from field guide descriptions
- 1987-1998 The Message Understanding Conference

The Message Understanding Conferences

- MUC-1 (1987); MUC-2 (1989)
  - Naval operations messages
- MUC-3 (1991); MUC-4 (1992)
  - Terrorism in Latin American Countries
- MUC-5 (1993)
  - Joint ventures and microelectronics domain
- MUC-6 (1995)
  - News articles on management changes
- MUC-7 (1998)
  - Satellite launch reports

MUC 6 Tasks

- Named Entity Recognition
  - Recognition of entity names (for people and organizations), place names, temporal expressions, and certain types of numerical expressions
- Coreference
  - Identification of coreference relations among noun phrases
- Information Extraction
  - Extraction of information about a specified class of events and the filling of a template for each instance of such an event

Named Entity Recognition

```
Mr.  <ENAMEX TYPE="PERSON">Dooner</ENAMEX> met with <ENAMEX TYPE="PERSON">Martin
Puris</ENAMEX>, president and chief executive officer of <ENAMEX
TYPE="ORGANIZATION">Ammirati & Puris</ENAMEX>, about <ENAMEX
TYPE="ORGANIZATION">McCann</ENAMEX>'s acquiring the agency with billings of <NUMEX
TYPE="MONEY">$400 million</NUMEX>, but nothing has materialized.
```


Coreference

```
Maybe <COREF ID="136" REF="134">he</COREF>'ll even leave something from <COREF
ID="138" REF="139"><COREF ID="137" REF="136">his</COREF> office</COREF> for <COREF
ID="140" REF="91">Mr.  Dooner</COREF>.  Perhaps <COREF ID="144">a framed page from
the New York Times, dated Dec.  8, 1987, showing a year-end chart of the stock market
crash earlier that year</COREF>.  <COREF ID="141" REF="137">Mr.  James</COREF> says
<COREF ID="142" REF="141">he</COREF> framed <COREF ID="143" REF="144"
STATUS="OPT">it</COREF> and kept <COREF ID="145" REF="144">it</COREF> by <COREF
ID="146" REF="142">his</COREF> desk as a "personal reminder.  It can all be gone like
that."
```


Information Extraction

McCann has initiated a new so-called global collaborative system, composed of world-wide account directors paired with creative partners. In addition, Peter Kim was hired from WPP Group's J. Walter Thompson last September as vice chairman, chief strategy officer, world-wide.

```
<SUCCESSION_EVENT-9402240133-3> :=
    SUCCESSION_ORG: <ORGANIZATION-9402240133-1>
    POST: "vice chairman, chief strategy
           officer,  world-wide"
    IN_AND_OUT: <IN_AND_OUT-9402240133-5>
    VACANCY_REASON: OTH_UNK
<IN_AND_OUT-9402240133-5> :=
    IO_PERSON: <PERSON-9402240133-5>
    NEW_STATUS: IN
    ON_THE_JOB: YES
    OTHER_ORG: <ORGANIZATION-9402240133-8>
    REL_OTHER_ORG: OUTSIDE_ORG
<ORGANIZATION-9402240133-1> :=
    ORG_NAME: "McCann"
    ORG_TYPE: COMPANY
<ORGANIZATION-9402240133-8> :=
    ORG_NAME: "J. Walter Thompson"
    ORG_TYPE: COMPANY
<PERSON-9402240133-5> :=
    PER_NAME: "Peter Kim"
```

Evaluation

- Evaluation can be based on how many slots are filled correctly
- Precision
  - Correct slots filled/Total slots filled
- Recall
  - Correct slots filled/Total Possible Correct slots
- The F Measure: a weighted combination of Precision and Recall

$$F = \frac{(\beta^2 + 1)PR}{(\beta^2 P + R)}$$

β represents relative importance of P and R

Evaluation

- For complex IE tasks, people agree on slot fills in the 60-80% range
- Current state-of-the-art:
  - the F = 0.60 barrier
  - Systems achieve about 60% of human performance

Benefits of MUCs

- A common task for everyone
- Development of large corpora with associated "key templates"
- Set format for the templates, methods of automatically scoring program output to hand-created key templates
- Methods for evaluating system performance

Outline of This Lecture

- What is Information Extraction and Question Answering
- History of Information Extraction Systems
- History of Question Answering Systems
- Issues about IE and QA

History of Question Answering Systems

- NL Front ends to databases
  - BASEBALL, LUNAR, Chat-80
- Dialogue interactive advisory systems
  - Eliza, SHRDLU
- Question answering and story comprehension
  - QUALM
- Large Projects
  - Unix Consultant, LILOG
- The Question Answering Track of TREC
  - TREC-QA

NL Front Ends to Databases

- BASEBALL (1961)
  - QA about baseball games
    - *Who did the Red Sox lose to on July 5?*
    - *How many games did the Yankees play in July?*
    - *On how many days in July did eight teams play?*
- LUNAR (1971)
  - QA on the chemical analysis data on lunar rock and soil composition accummulated from the Apollo mission
    - *What is the average concentration of aluminium in high alkali rocks?*
    - *How many Brescias contain Olivine?*
  - Was able to answer 90% of questions from untrained users in a 1971 lunar science convention

NL Front Ends to Databases

- Chat-80 (1979-1982)
  - QA on geography
    - *What is the total area of countries south of the Equator and not in Australasia?*
    - *Which is the largest African country?*
  - Prolog program
  - A more modern version converts the question to SQL
- Common features of these NL DB front ends:
  - Use of a structured DB
  - A limited domain

Dialogue Interactive Advisory Systems

- Eliza (1966)
  - Simulation of a Rogerian psychologist
  - Simple keyword selection and pattern matching techniques
  - Many people thought that the machine could understand
- SHRDLU (1968-1970)
  - http://hci.stanford.edu/~winograd/shrdlu/
  - Dialogue with a robot in a toy world
  - Dialogue management and inferences
  - Handling of the laws of physics in the toy world

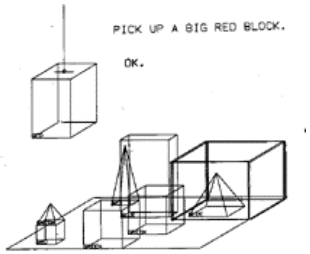
ELIZA – Sample dialogue

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I am depressed much of the time.
I AM SORRY TO HEAR THAT YOU ARE DEPRESSED
It's true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
My mother takes care of me.
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
My father.

ELIZA Sample (cont.)

My father.
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO THINK I AM AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

SHRDLU – Sample dialogue



http://www.cs.cf.ac.uk/Dave/AI1/COPY/shrdlu.html

1 pick up a red block
OK
in order to pick up the red block it had to find a space to place the gren one on top.

2 grasp the pyramid
I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN
there are three pyramids so nothing is done

3 find a block which is taller than the one you are holding and put it in the box
BY "IT" I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING
OK
the system interprets one and it and puts down the red block and picks up the taller blue block reach figure 4

4 what does the box contain?
THE BLUE PYRAMID AND THE BLUE BLOCK

5 what is the pyramid supported by?
THE BOX
earlier the system did not know which pyramid now it assumes the one in the box

QA and Story Comprehension

- QUALM (1977)
  - QA on an arbitrary text
  - Text stored as Conceptual dependencies
    - Aim at full text understanding
  - Question classification into conceptual categories
  - Use of scripts to model typical scenarios
    - Inferences and default reasoning

Large Projects

- Berkeley Unix Consultant (1988)
  - QA on the Unix operating system
  - Complex system that included, among others:
    - Language Analyser
    - Goal Analyser
      - Predict the user goals and plans
    - User modeling
    - Domain planner
    - Knowledge Representation
    - An agent

Large Projects

- LILOG (1991) – IBM Germany
  - QA on tourist information in a German city
  - Conceived more as research on NLP rather than as a practical system
  - Topics of research
    - Linguistic processing
      - parsing, lexicon, syntax, semantics
    - Knowledge representation
      - representation, DB engineering
    - Spatial knowledge
    - Generation

TREC

- trec.nist.gov
- Co-sponsored by
  - NIST
    "National Institute of Standards and Technology"
  - DARPA
    "Defense Advanced Research Projects Agency"
  - ARDA
    "Advanced Research and Development Activity"

TREC-QA

- TREC-QA
  - Question & Answering track of the Text REtrieval Conference
- Question Answering based on Large Volumes of Text
- Competition-based (like MUC)
- Quantitative Evaluation
- Increasing difficulty in every edition
  - TREC-QA Roadmap

TREC-8 QA Track

- The first TREC-QA track
- 1,900 MB of data
- 200 fact-based, short-answer questions
  - constructed by examining the data
  - explicit answer is guaranteed to be in at least one document
- For each question, systems return five answers with pointers to the documents
- Each answer evaluated by 2 human assessors
  - Introduction of the Mean Reciprocal Rank (MRR) score
    - Score of one question = 1/Position of the first correct answer

TREC-9 QA Track

- 3,000 MB of data
- 693 fact-based, short-answer questions
- The questions are logs (and variations) of real systems
  - Encarta
  - Excite
- The answer is guaranteed to be in the data
- Each answer assessed by one human assessor
  - MRR score
  - Many assessment "blunders" appeared

TREC-10 QA Track

- 3,000 MB of data
  - Newspaper and newswire documents
- Main task: Only 50-byte runs
  - Evaluation like TREC-8
- Addition of "list" and "context" tasks

TREC-10 QA Track

Main task

- similar to previous years' task
- no guarantee of an answer

List task

- assemble a set of instances
- need to detect repetition of same instance across documents

Context task

- track discourse objects across questions

TREC-11 QA

- This time, return the exact answer
- Return only one answer per question
- Rank the question-answers according to confidence

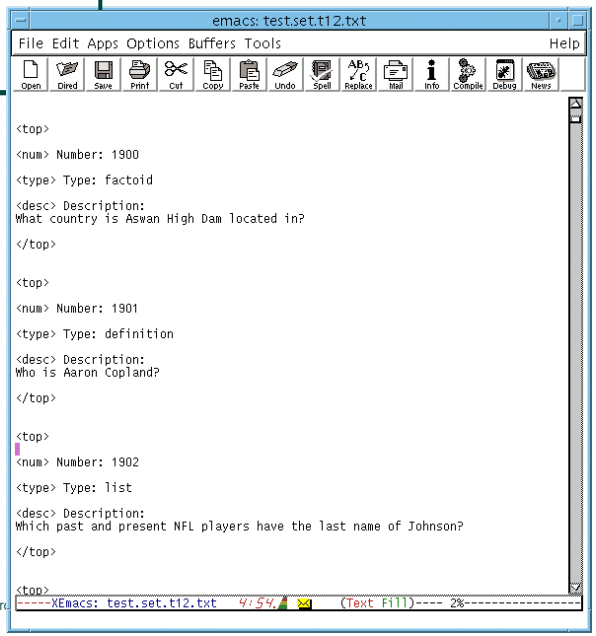
$$\frac{1}{Q}\sum_{i=1}^{Q}\frac{\#\text{ correct in first } i \text{ ranks}}{i}$$

- List task mainly as TREC-10
  - Evaluation in terms of accuracy: "the number of distinct, correct responses divided by the target number of instances"
- No context task

TREC-12 QA

- Same data as TREC11
- Three types of questions in the main task:
  - Factoid
  - List
  - Definitions
- No ranking according to system confidence
- Each type of question is evaluated independently, and a combined score is produced

TREC12 – Main Task questions



Outline of This Lecture

- What is Information Extraction and Question Answering
- History of Information Extraction Systems
- History of Question Answering Systems
- Issues about IE and QA

IE from Conference Announcements

- Mailboxes get filled with announcements for conferences and other events
- We want to get to the most important elements of information required for making decisions
- We'd like to have the information available in structured form for easier manipulation and searching

Our Scenario

- We want to build an application that will interface to a mail client
- We assume someone else will provide the software that filters out conference announcements for us
- Our job is to construct the information extraction functionality that will populate a conference database

A Conference Announcement Fragment

```
From: icgi94  <netmail!icgi94@iti.upv.es>
To: Non Receipt Notification Requested  <elsnet-
list@cogsci.edinburgh.ac.uk>
Date: Friday, 11 February 1994 7:05PM
---------------------------------------------------------------
         SECOND & FINAL ANNOUNCEMENT & CALL FOR PAPERS
   SECOND INTERNATIONAL COLLOQUIUM ON GRAMMATICAL INFERENCE
                         ICGI - 94
              Pueblo Acantilado, Alicante, Spain
                    September 21-23, 1994

Co-sponsored by the Universidad Politecnica de Valencia (UPV),
the Universidad de  Alicante  (UA),  the  Asociacion  Espa7ola
de  Reconocimiento de
```


The Anatomy of a Conference Announcement

- Mail headers
- Optionally, some introductory remarks from the sender
- A header block, marked out typographically in some way
- Multiple paragraphs describing the purpose of the event and the topics of the papers requested
- Somewhere in those paragraphs, various dates and the URL if there is one

A Conference Announcement Fragment

```
LOCAL ORGANIZERS
  Frank Wolter, Leipzig        Holger Sturm, Leipzig

IMPORTANT DATES
  Submission deadline: May 15, 2000
  Notification: July 15, 2000
  Workshop: October 4-7, 2000
  Preliminary version for workshop volume due:  at the workshop
  Notification of acceptance for publication:   December 1, 2000

FURTHER INFORMATION
E-mail enquiries about AiML-ICTL 2000 should be directed to
<wolter@informatik.uni-leipzig.de>.  Information about AiML can
be obtained on the World-Wide Web at
<http://www.illc.uva.nl/~mdr/AiML/>, and about AiML-ICTL 2000 at
<http://www.informatik.uni-leipzig.de/~wolter/aiml.html>.
```


What's in a Data Fill

- ID
- Title
- City
- Country
- Start date
- End date
- Due date
- URL

An Example Filled Template

```
ID: 031
Title: SECOND INTERNATIONAL COLLOQUIUM ON GRAMMATICAL
INFERENCE
Acronym: ICGI-94
Start Date: 21-SEP-1994
End Date: 23-SEP-1994
City: ALICANTE
Country: SPAIN
Due Date: 10-JAN-1995
```


What's Involved for Each Data Field?

- ID: trivial
- Title: very hard – what counts as a title?
- City: requires some geographical knowledge
- Country: requires some geographical knowledge
- Start date: relatively straightforward but ...
- End date: similar issues to start date
- Due date: need to look for cues
- URL: usually straightforward

Issues about QA

- Roadmap for Question Answering

  http://www-nlpir.nist.gov/projects/duc/roadmapping.html
- Increasing difficulty
- A very ambitious program overall
- A very fast pace for TREC-10 to TREC-14

Performance Criteria

- Timeliness
  - The answer must be provided in real-time
  - New data sources must be incorporated in the QA system as soon as they become available
- Accuracy
  - Incorrect answers are worse than no answers
  - Need to incorporate world knowledge and mechanisms that mimic common sense inference

Performance Criteria

- Usability
  - Knowledge in a QA system must be tailored to the specific needs of a user
  - A QA system must be able to mine answers regardless of the data source format
  - The answer must be delivered in any format desired by the user
  - The QA system must allow the user to describe the context of the question

Performance Criteria

- Completeness
  - Complete answers to a user's question is desirable
  - Answer fusion in a coherent information is required
- Relevance
  - The answer to a user's question must be relevant within a specific context
  - Interactive QA, in which a sequence of questions helps clarify the information need, may be necessary
  - Evaluation must be user-centered: humans are the ultimate judges

The Issues

- Question classes: Need for question taxonomies
- Question processing: Understanding, ambiguities, implicatures and reformulations
- Context and Q&A
- Data sources for Q&A
- Answer extraction: Extraction of simple and distributed answers; answer justification and evaluation of answer correctness

The Issues

- Answer formulation
- Real time question answering
- Multi-lingual question answering
- Interactive Q&A
- Advanced reasoning for Q&A
- User profiling for Q&A
- Collaborative Q&A