Task first, please

Valentin Jijkoun and Maarten de Rijke

Intelligent Systems Lab Amsterdam
University of Amsterdam

SIGIR'07 Workshop on Focused IR, July 27, 2007

Overview

- Question Answering @ TREC and CLEF
- QA task parameters
- Parameter optimization: user or tools?
- Example QA scenarios

Question Answering

- Focused information need, focused response
- Emerged in 1960s as natural language interfaces to databases
- Revived in 1990s as text-based QA
- Tasks: TREC (1999), NTCIR (2002), CLEF (2003)

QA @ TREC

- Started with short "factoid" questions and a news text collection
  - *Who killed Lee Harvey Oswald?*
  - *When was Mozart born?*
- Currently: topic-driven, in news and blogs
  - Topic: *John William King convicted of murder*
  - *How many non-white members of the jury were there?*
  - *Who was the foreman for the jury*?
  - *Who was the victim of the murder? …*
  - *What OTHER bits of information are interesting?*

QA @ CLEF

- Similar to QA @ TREC, but:
  - Mono- and cross-lingual (9 languages)
  - Temporal restrictions (*Who was X during Y?*)
  - Systems provide supporting snippets
  - Real-Time Exercise, Answer Validation
  - Wikipedias added in 2007

Parameters for a QA task

- Response format and answer "exactness"
- Number of responses
- "NIL" questions
- Question types
- Question generation, answer assessment
- Data collection: match with information needs?
- Multi-/cross- linguality
- Evaluation measures

Exact answers

- Early QA@TREC: retrieving passage with answer
- Now: retrieving answer string
- QA@CLEF: plus supporting snippet(s)
  - Surely saves assessor's time
  - Users prefer context (Lin et al., 2003)

Exact answers - 2

- All tasks require "exactness" judgments
- *Where is IMF headquartered?*
  - *Washington* (exact)
- *Where was Mozart born?*
  - *Salzburg* (inexact)
  - *Salzburg, Austria* (exact)
- Highly depends on assessor's (=user's!) background and expectations
  - Should be included in the information need?

Number of answers

- Top 1, in earlier editions top 3
  - Compromise between the load on assessors and usability
  - OK for apps like Mobile QA, not OK for information analysts
- What about "list" questions?
  - *Name all airports in London.*
    - "closed" list, precision/recall? All-or-nothing answer?
  - *Name several French composers.*
    - "open" list, recall not important

"NIL" questions

- Questions that are "known" to have no answers in the collection
- Would user appreciate an empty reply?
  - "I don't know" vs. "I'm not sure, but maybe…" responses
  - Confidence weighted score?

Question types

- What questions should be included?
  - Question from search engine logs:
    - procedural 38% *(How to cook ham?)*
    - description 13% *(Who is Victoria Gott?)*
    - explanation 10% *(Why people do good deeds?)*
    - factoid 10% *(What did caribs eat?)*
- Is QA addressing real questions?

Question generation and answer assessment

- TREC: questions from search logs
  - We don't necessarily know what the user was expecting
- CLEF: still back-generated from collections
  - Responses assessed by "authors"
  - Many not quite natural q's like: *In which European capital is the Eiffel tower?*

Matching sources and information needs

- Most textual QA activities on news texts
  - TREC: +blogs to provide text diversity
    - But why answer *"In what country is Luxor?"* from newspapers or blogs?
  - CLEF: +Wikipedia as source of "general" facts
- Should not sources be "natural" for questions?
  - General factoids: encyclopedias
  - Smaller events: news
  - Opinions, preferences: blogs

Cross-lingual tasks

- One of key points for QA@CLEF
  - Currently: translating questions to create cross-lingual task
  - Why answer *"Wie is Flaubert?"* agains Spanish collection?

Confusion in task definition?

- At least two distinct modes for QA:
  - QA as a user-driven task
    - User background and expectations
    - Interaction (interface)
  - QA as a framework to evaluate NLP tools
    - Make questions on which your tool may help
    - ~ reading comprehension
- Much vagueness and inconsistency come from confusing the two modes

Task: Intelligence gathering

- "Analytical" QA
- Exploratory questions
  - *What has been Russia's reaction to the US bombing of Kosovo?*
- Frame-based questions
  - Events: time, location, injured,…
- Source: news, blogs(?)
- The scenario of ciQA (Complex Interactive QA)

Task: Event-based QA

- User: journalist or history student
  - Starts with an article describing an event and follows with questions around the topic
- Sources: news, encyclopedia's
- Response: ranked list of answers with justification
- Limited cross-linguality

Task: Trivia game show

- No explicit user, but explicit task
- Factoid questions, not necessarily natural
- Answerhood, evaluation measures - straighforward
- No restriction on sources
- Lots of test data

Conclusions

- QA needs a clear definition
  - Need justification of the setup choices
  - Essential for meaningful evaluation
- Many possible options
  - Application-driven
  - Tool-driven
- Task comes first!