Text REtrieval Conference (TREC) Question Answering Tasks and Evaluation Methods

Hoa Trang Dang

National Institute of Standards and Technology

April 27, 2007

NIST
National Institute of Standards and Technology

Evolution of QA Tasks

- factoid (1999): fact-based short answer ("How many calories are there in a Big Mac?")
- list (2003): "List the names of chewing gums"
- definition (2003): "Who is Vlad the Impaler?"
- static question series about a target (2004)
  - time-dependent questions about events (2005)
- complex "relationship" questions (2005)
  - interaction/clarification (2006)
- question series over blogs and newswire (2007)

Overview

- Two TREC 2007 QA tasks:
  1. Main task: return answers to questions in series
  2. Complex Interactive QA: return answers to relationship questions; allow limited interaction

Main Task:
Question Series

Question Series

- Series are abstraction of “user sessions”
- Each series is about a specified target (Person, Organization, Event, Thing)
- Goal is to gather info about target
- Series contains factoid, list, and final “Other” question requesting additional (unspecified) interesting facts
- Questions tagged as to type (factoid, list, other)
- Questions can depend on previous answers

Example Question Series

TARGET:	"John William King convicted of murder"
145.1 FACTOID	How many non-white members of the jury were there?
145.2 FACTOID	Who was the foreman for the jury?
145.3 FACTOID	Where was the trial held?
145.4 FACTOID	When was King convicted?
145.5 FACTOID	Who was the victim of the murder?
145.6 LIST	What defense and prosecution attorneys participated in the trial?
145.7 OTHER	Other

Document Set

- Combined collection:
  - Newswire documents (freely distributed by NIST)
    - 3 GB data; 1 million documents
  - Blogs (purchased from U. Glasgow: 400 pounds)
    - 136 GB data; 3.2 million permalink documents
- Responses to all questions must be supported by documents from corpus
- NIST provides ~50 “top docs” for each topic/target

Evaluation of Factoid Questions

- Response is single [docid, answer-string] or NIL
- Human assessors judged response as one of {wrong, unsupported, inexact, *locally* correct, *globally* correct}
- NIL is globally correct iff no answer in collection
- Score of Factoid question is 1 if response is judged as globally correct, 0 otherwise
- FactoidScore = Accuracy = fraction of factoid questions judged as globally correct

List Questions

- Questions seek multiple instances of a specific type
- Response is a list of [docid, answer-string] pairs
- Each pair is judged as for factoids
- One answer-string marked as distinct for each set of equivalent globally correct answer-strings

List Scoring

- Single assessor created final list of known, distinct, globally correct answers
- Precision = #distinct / #returned
- Recall = #distinct / #total
- Combine precision and recall: F = (2*P*R)/(P+R)
- ListScore = F score of list question

"Other" Questions

- Response is a list of [docid, answer-string] pairs
- Response should contain additional interesting information about target (not in previous questions in series)
- Primary assessor determines the set of "atomic" information nuggets that a good response should contain
  - distinction between *vital* and *okay* nuggets
- Primary assessor marks which nuggets appear in system response

Example Nugget List for “Other”

Target: “John William King convicted of murder”

vital	KKK and New Black Panthers gathered in town where trial held
vital	Only 1 white man had ever been executed in Texas for killing a black
vital	Governor Bush took no position on a proposed Texas hate-crimes law
okay	King had shirt with victim's DNA in his apartment
okay	King was sentenced to death
okay	Two other men were implicated in same crime
okay	King was a white supremacist
okay	Victim was dragged for 3 miles along road

"Other" Scoring

- Using assessor judgments, compute nugget recall and approximation of nugget precision (a function of response length)
- Score for question is F(beta=3), which gives more weight to recall than to precision
- Compute two variants of "Other" score:
  - primary-assessor F-score (1 assessor)
  - pyramid F-score (multiple assessors)

Primary “Other” Scoring

weight of nugget is 1 if vital, 0 if okay

numVitalMatches = sum of weights of all nuggets retrieved
numVital = sum of weights of all nuggets in list
numTotalMatches = # of vital and okay nuggets retrieved
C = character allowance per match (C=100)

Recall = numVitalMatches / numVital

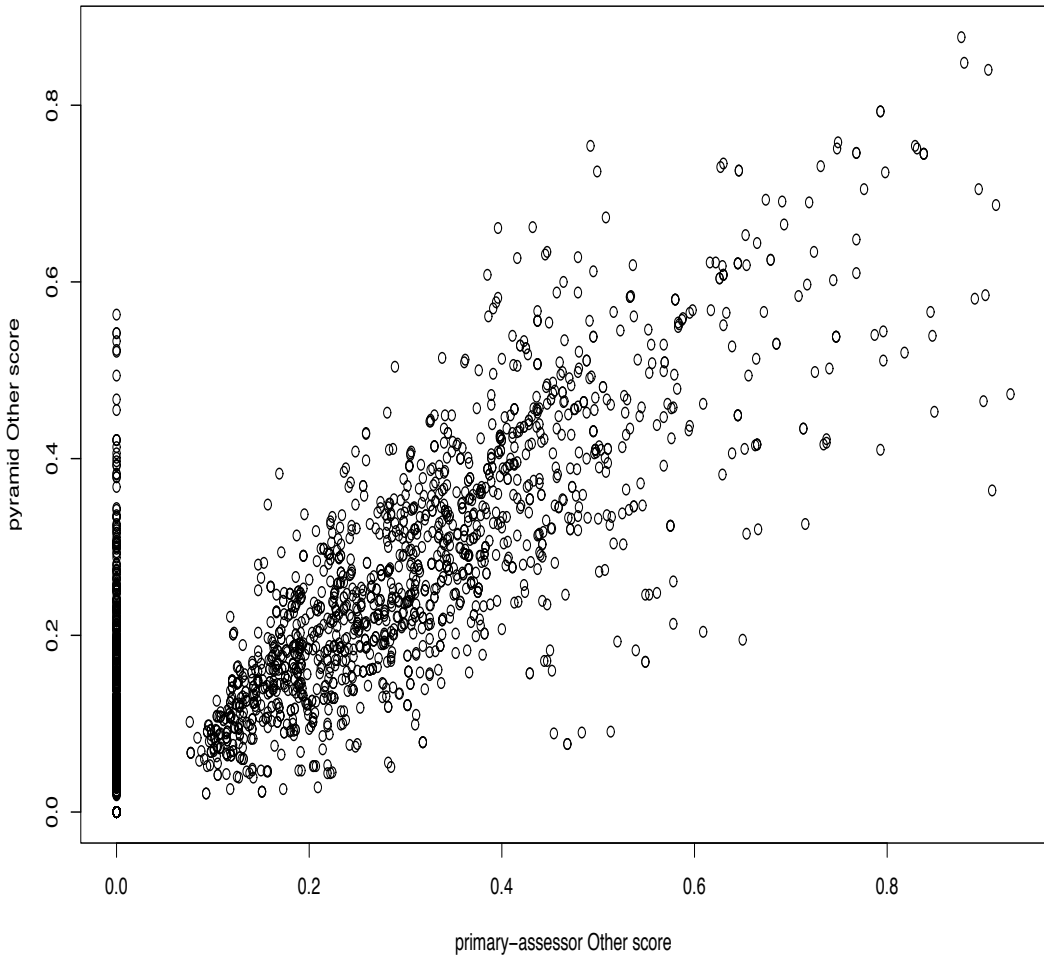
Approximated Precision:
set *okayLength = C * numTotalMatches*
if (*length < okayLength*) then *Precision = 1*
else *Precision* = 1 - ((*length - okayLength) / length*)

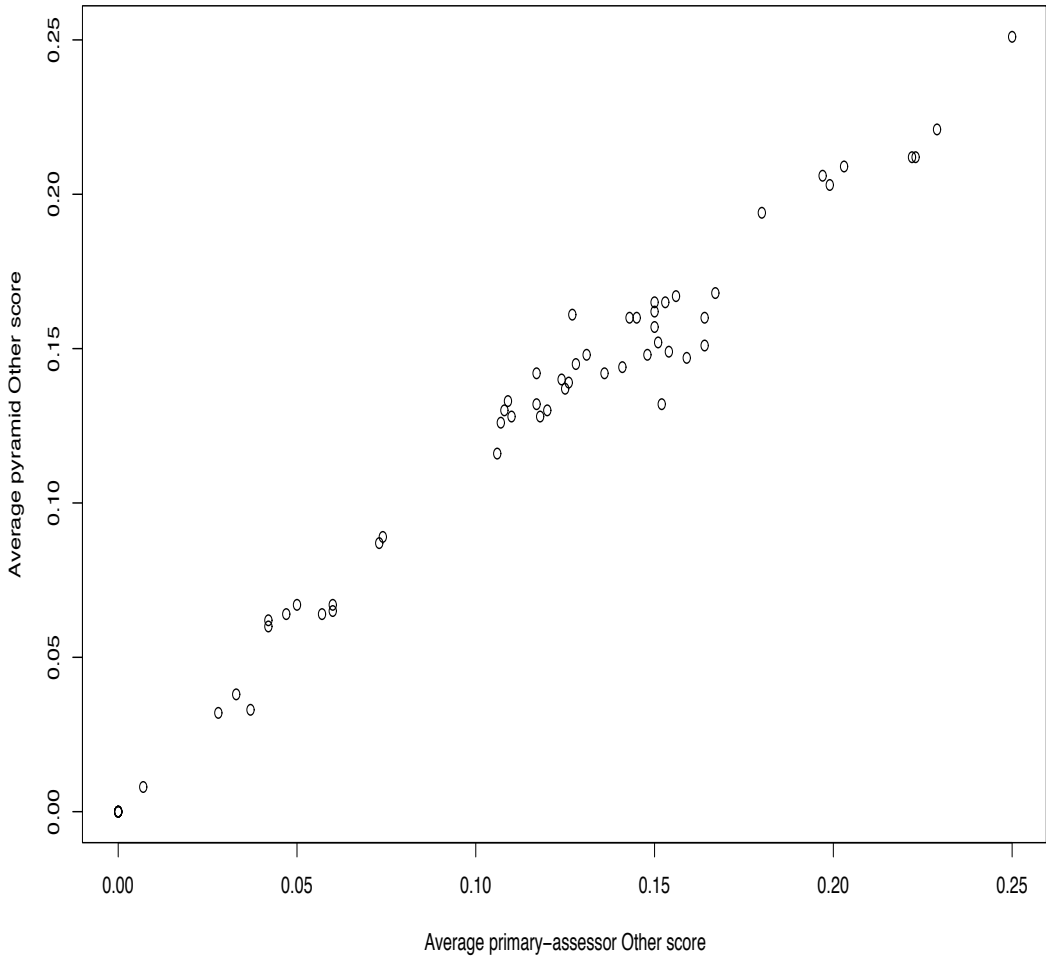
*F(beta=3) = 10 * Recall * Precision / (9 * Precision + Recall)*

Nugget Pyramid "Other" Scoring

- Based on Lin and Demner-Fushman (HLT 2006)
- 9 judgments of vital/okay from 8 different assessors, using nugget list from primary assessor
- Nugget weight in [0.0, 1.0] instead of {0.0, 1.0}
  - weight is fraction of judgments of vital for the nugget, normalized so maximum nugget weight is 1.0
- Precision, Recall, F same as for primary-assessor scoring

Primary F vs. Pyramid F





P = 0.870 [0.863,1.00]

P = 0.987 [0.980, 1.00]

Challenge: Fragmented Text

- How many Oscars has she [Judi Dench] won? **one**
  - NYT19990321.0226: *Judi Dench.... Oscar history: This is her first win*
- How many Oscars did Hitchcock win? **none**
  - NYT19990808.0092: *Hitchcock.... Oscar considerations: Five nominations for best director....No wins.*

Challenge: Temporal Inference

- In what year was Moon born? **1956**
  - NYT19980721.0033: *Moon's age (42 in November)*
- What year was he [Barry Manilow] born? **1946**
  - APW19990616.0281: *Today's birthdays.... Barry Manilow is 53*

Challenge: Temporal Inference

- What year was she [Patsy Cline] inducted into the Hollywood Walk of Fame: **1999**
  - APW19990804.0218: *More than three decades after her death, country legend Patsy Cline got a star on the Hollywood Walk of Fame. About 150 fans gathered Tuesday to witness the unveiling of the star...*
- When was King convicted? **23 February 1999**
  - NYT19990225.0385: *voting on Tuesday to convict King*

Challenge: Identifying the Event

- How old was Elian at the time of the shipwreck? **five years old**
  - NYT20000126.0214: Elain Gonzalez, who was 5 at the time, was found clinging to an inner tube off the coast of Florida on Nov. 25 after the boat carrying him to the United States capsized....
- Who was the women's winner of the 1999 Chicago Marathon? **Joyce Chepchumba**
  - XIE19991028.0042: *Chepchumba, fresh from a gusty victory in Chicago on Sunday....*

Challenge: Common Sense Reasoning

- How many non-white members of the jury were there? **one**
  - APW19990301.0168: *A jury of eleven whites and one black sentenced John William King, 24, to death*
- How many judges were in the pageant? **7**
  - *She and six other celebrities will pick Miss America 2000*

Complex Interactive QA (ciQA)

ciQA Task

- Complex question comprises a template and free narrative
- Response format and evaluation is the same as for "Other" question
- Allow optional 5-minute interaction with assessor using web-based forms created and hosted by participant
- Allow second, post-interaction, submission of answers
- Search entire newswire collection

Question Templates

1. What evidence is there for transport of [goods] from [entity] to [entity]?
2. What [RELATIONSHIP] exist between [entity] and [entity]?
3. What effect does [entity] have on [entity]?
4. What is the position of [entity] with respect to [issue]?
5. Is there evidence to support the involvement of [entity] in [entity/event]?

Example Topic

- Template 2: What [financial relationships] exist between [drug companies] and [universities]?
- Narrative: The analyst is concerned about universities which do research on medical subjects slanting their findings, especially concerning drugs, towards drug companies which have provided money to the universities.

Example Topic

- Template 4: What is the position of [Richard Seed] with respect to [human cloning]?
- Narrative: The analyst would like to know how Richard Seed felt about human cloning. Specifically, the analyst would like to know what his feelings were regarding human cloning and what actions he took as a result.

TREC QA vs. DUC summarization

- Similarities:
  - Complex answers ("Other" and ciQA)
  - Nugget-based evaluation ("Other" and ciQA)
- Differences:
  - QA requires searching large corpus
  - QA evaluation requires exact answers where possible (factoid, list); list task could be useful for synthesis and abstraction in summarization
  - TREC QA doesn't evaluate fluency