CS674 Natural Language Processing

Question Answering

Eric Breck

Cornell University

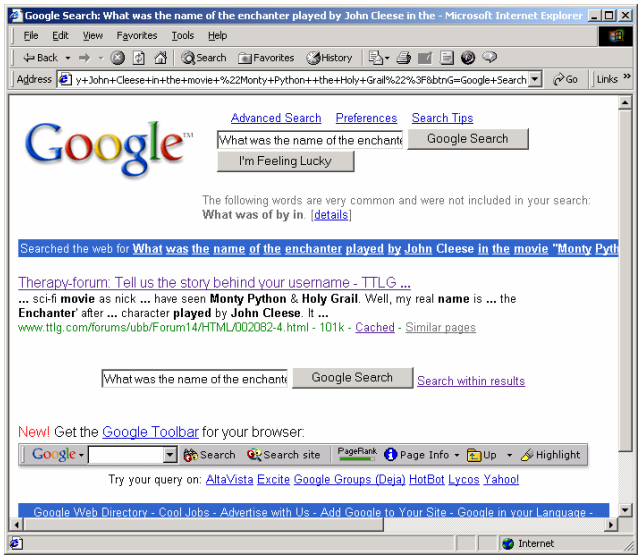
Slides based on Claire Cardie (Cornell),
Ellen Voorhees (NIST),
Pasca & Harabagiu (SIGIR 2001),
J. Callan and K. Czuba (CMU)

A question

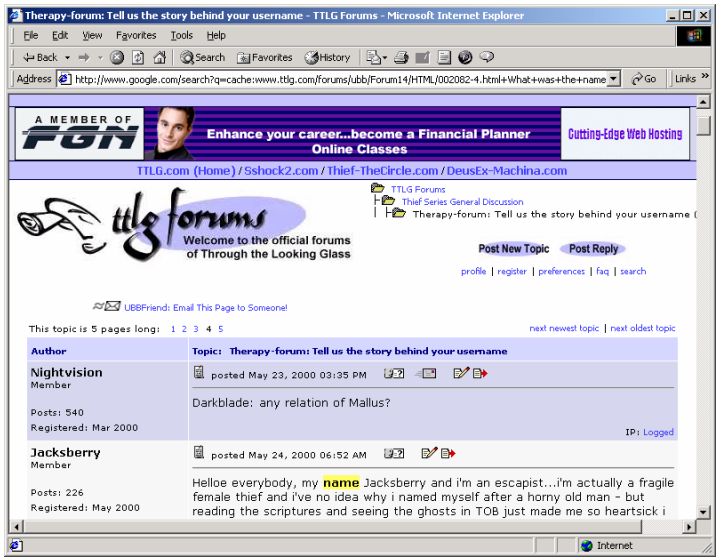
- What was the name of the enchanter played by John Cleese in the movie "Monty Python and the Holy Grail"?

Trivia: John Cleese was a Cornell A.D. White professor-at-large.

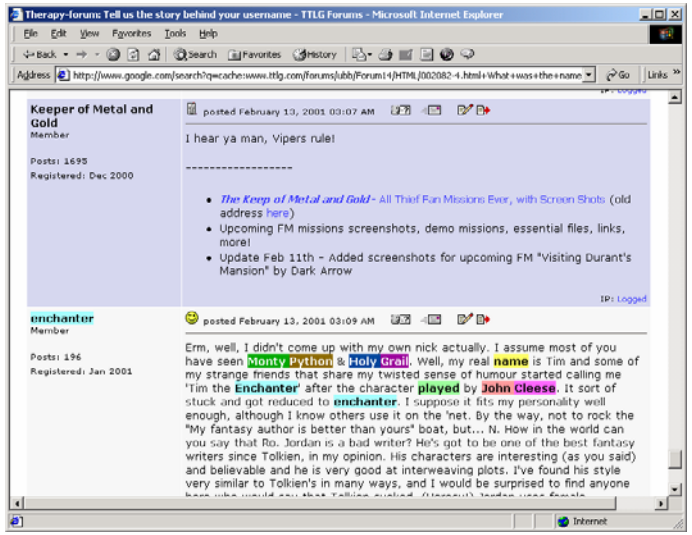
IR solution



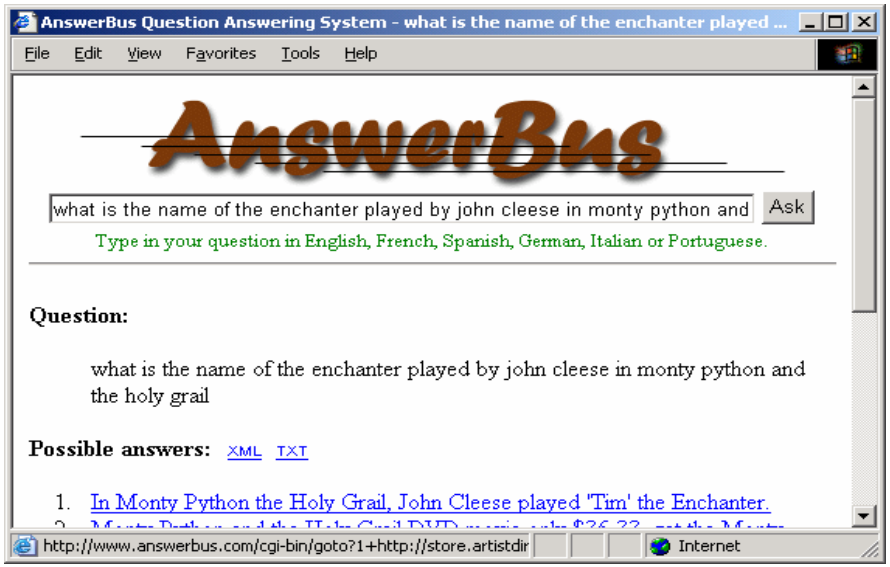
The document



The answer



QA solution



Question answering task

- **Goal: User types a question, system produces the correct answer**
  - System treats question as more than a bag of words
  - System returns a short response
- **Dimensions of QA task determine its difficulty**
  - Closed domain vs. open domain
  - Searching structured data vs. unstructured data
  - Extracted answers vs. generated or compiled answers
  - Answer length

Outline

- **A bit of history**
- **Evaluation**
- **Three systems**

History

- **Closed-domain QA systems**
  - LUNAR [Woods & Kaplan, 1977]
  - WOLFIE [Thompson & Mooney, 1998]
  - Q/A [Lehnert, 1978]
- **Open-domain QA systems**
  - MURAX (Julian Kupiec, 1993)
    - Trivial Pursuit, answered from Grolier's
  - TREC QA evaluations [1999-2003, …]

LUNAR

- **Answered questions about moon rocks and soil gathered by the Apollo 11 mission**
  - Data base of information for all collected samples
- **Architecture**
  - Parse English question into a data base query
  - Run query on data base to produce answer
- **Sample questions**
  - What is the average concentration of aluminum in high alkali rocks?
  - What samples contain P205?

Towards open-domain QA

Which country has the largest part of the Amazon rain forest?

The chaotic development that is gobbling up the Amazon rain forest could finally be reined in with a new plan developed by officials of Amazon countries and leading scientists from around the world.

"That's some of the most encouraging news about the Amazon rain forest in recent years," said Thomas Lovejoy, a tropical ecologist at the Smithsonian Institution and an Amazon specialist.

"It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon, especially in Brazil, except bad news," Lovejoy said in a recent interview.

Sixty percent of the Amazon, the world's largest tropical rain forest, lies in Brazil, but the forest also covers parts of the eight surrounding countries.

Lovejoy was one of the organizers of an unusual workshop held in mid-January in Manaus, Brazil, a sprawling city of 1 million people in the heart of the Amazon. It was the center of Brazil's once-thriving rubber trade.

Evaluation: TREC

- **Annual Information Retrieval Conference**
- **Multiple tracks**
  - Ad hoc retrieval
  - Interactive
  - Cross-language
  - As of 1999 (TREC-8): QA
- **Run by NIST**

Open-domain QA for TREC

- **Open-domain QA is too hard**
- **Task influenced by what can be (reasonably) evaluated**
  - Given a relatively unambiguous English question, find a fact-based answer
  - Still hard…
    - e.g. "Where is the Taj Mahal?"
      - Agra, India
      - New Jersey
    - What if answer doesn't exist? Or requires piecing together information from different documents?

[TREC8-12 1999-2003]

TREC QA evaluations

- **Restrictions**
  - answer exists in the collection
  - supporting info can be found in a single document
  - answer expressed as (length limited) text fragments
    - E.g. "…at the city of Agra in the State of Uttar Pradesh, the…"
    - assume that the answer itself is short (less than 50 bytes)
    - lengths evaluated: 250 byte, 50 byte, exact
  - can return up to 5 guesses per question to the user
    - For TREC-11 (2002), only 1 guess
    - Use "NIL" to denote that no answer exists in the corpus

TREC questions

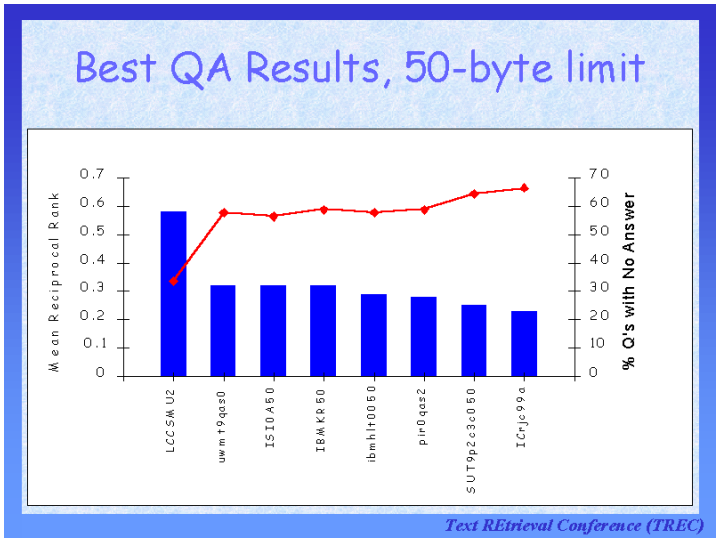
- **Who was the first American in space?**
- **Where is the Taj Mahal?**
- **How did Socrates die?**
- **Who invented the paper clip?**
- **Why did David Koresh ask the FBI for a tape recorder?**
- **Who is Colin Powell?**

- **Note that these are open-domain topics, but closed-class question form**
  - Questions conform to predictable language patterns
  - Most questions can be answered with little or no reasoning

TREC QA: evaluation

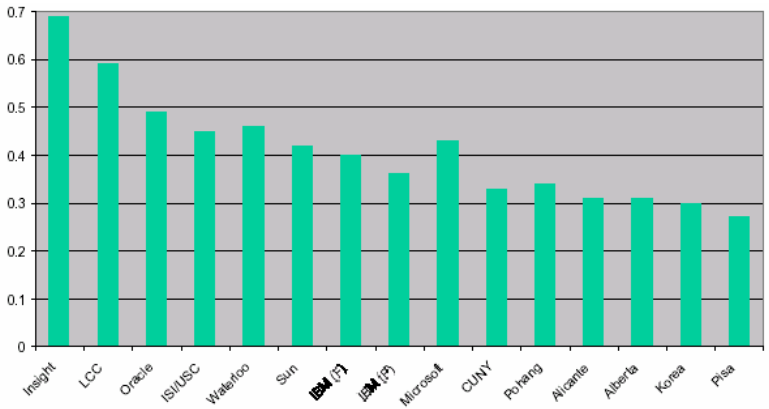
- **Human assessors judge the answers**
  - Allowed to accept multiple answers
- **Systems scored on *mean reciprocal rank* of first correct answer**
  - First answer correct = 1 point, second answer correct = ½ point, third answer correct = 1/3 point, …
  - 0 if none of the *n* answers are correct
  - Average across all questions
- **Also reported on the number of questions answered correctly**
- **Later evaluations: list questions and definition questions**

State of the art: TREC 2000

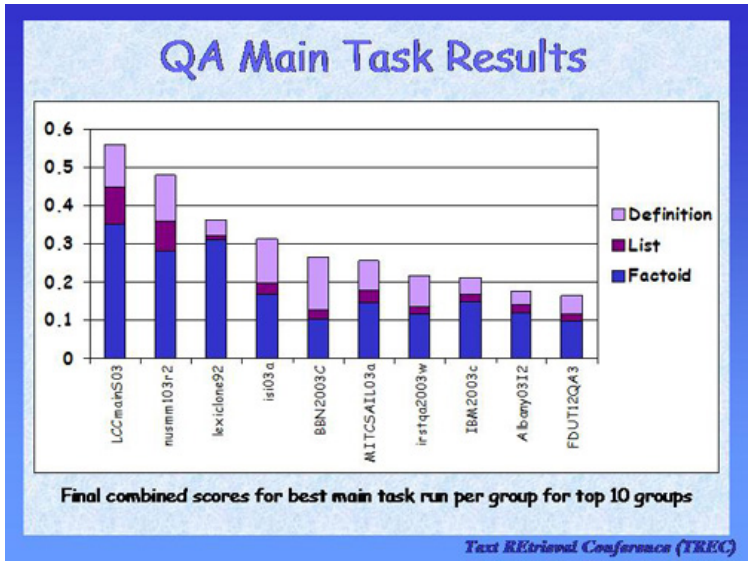


Courtesy of E. Voorhees

State of the art: TREC 2001



State of the art: TREC 2003

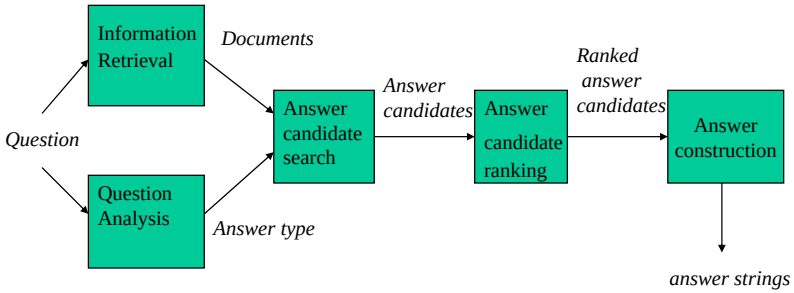


TREC QA evaluation: issues and problems

- **No penalties for answers that are correct but not helpful**
  - E.g. "…,Agra, India, New Jersey,…"
- **No penalties for wrong answers**
  - In a real system, these might be confusing to a person
- **No reward for multiple, complementary answers**
- **No user model or user interaction, no context**
  - So no guidance when question is ambiguous
- **Ambiguity about what is allowed**
  - Is it fair to find the answer on the Web, instead of the supplied corpus?
  - Is it fair to use a Gazeteer?

[Czuba, 2002] © 2002, Callan

A simple system: MITRE's Qanda

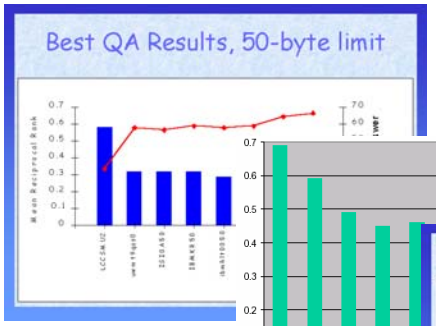


A gadfly: AskMSR

- **Eric Brill et al. at Microsoft Research**
- **Exploit massive redundancy of the Web (vs TREC collection)**
  - Where is the Louvre Museum located?
    - +The Louvre Museum +is located
    - +The Louvre Museum +is +in
    - +The Louvre Museum +is near
    - …
    - Louvre AND Museum AND near
  - Tiling
- **Competitive TREC-9,10 performance**

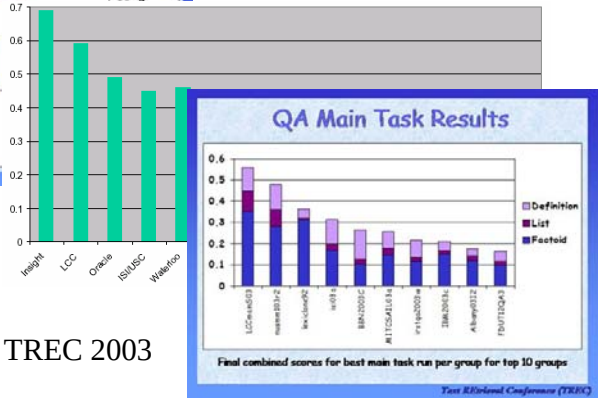
LCC system

Sanda Harabagiu, Dan Moldovan et al. (SMU, UT Dallas)



TREC 2001

TREC 2000



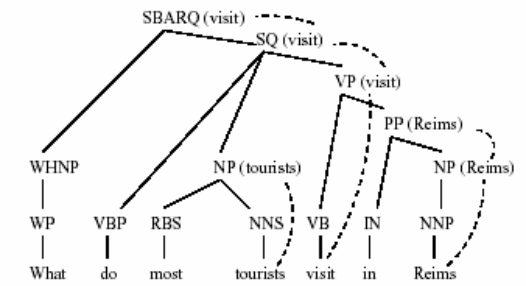
TREC 2003

Multi-strategy approach

- **State of the art in QA is the LCC system**
  - Employs informed use of standard IR techniques
  - Use of broad ontology (eXtended WordNet)
  - Lots of NLP
  - Answer verification
- **Similar to most other systems in architecture except for**
  - Much more careful tuning of algorithms and resources
  - More sophisticated control of IR and NLP
  - Feedback loops (these may be gone now)

Question analysis

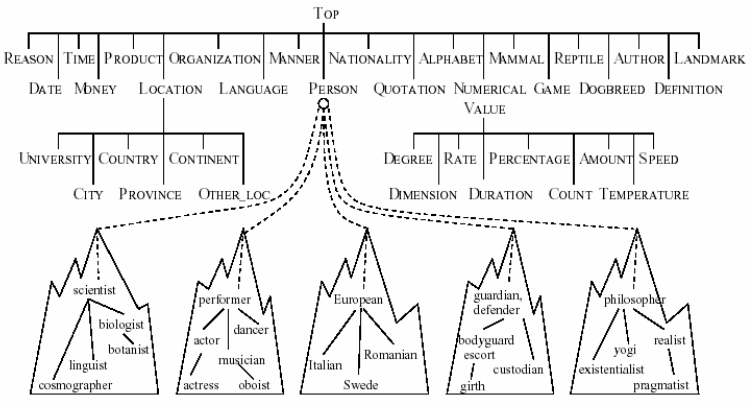
- **Parsing and named entity recognition**
- **Expected answer type determined by parsing**



- **Exceptions for "special cases"**

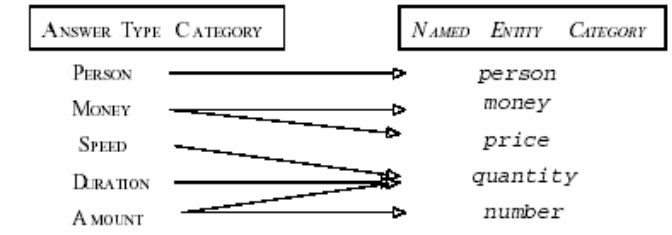
(Q-P1): *What {is|are} <phrase_to_define>?*
(Q-P2): *What is the definition of <phrase_to_define>?*
(Q-P3): *Who {is|was|are|were} < person_name(s)>?*

Expected answer types



Expected answer types

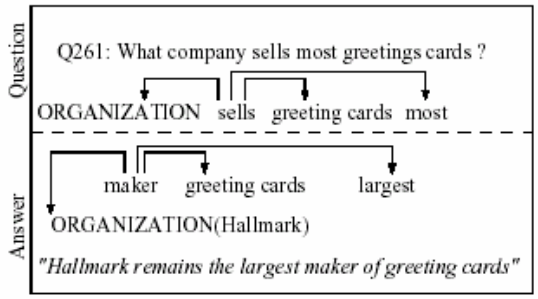
- **Answer types are mapped to named-entity categories that can be recognized in text**



- **Answer types drive processing of paragraphs**
  - Passages need to contain the expected answer type

Answer verification

- **Goal: increase precision**
- **Parse passages to create a dependency tree among words**
- **Attempt to unify logical forms of question and answer text**



Assessment

- **Strengths**
  - Controlled use of IR system
    - Query expansion via lexical and semantic equivalents
    - Believed to be the major power of the system
  - Tailored resources (see paper)
    - WordNet, parser, NE identifier, etc.
  - Answer verification
    - Initially thought to be the key component of the system
    - Now…not so clear
- **Weaknesses**
  - Complex system, contribution of each component unclear