

# Statistical NLP

Spring 2011

Berkeley



N L P

## Lecture 26: Question Answering

Dan Klein – UC Berkeley

## Question Answering

---

- Following largely from Chris Manning's slides, which includes slides originally borrowed from Sanda Harabagiu, ISI, Nicholas Kushmerick.

# Large-Scale NLP: Watson

The screenshot shows a Wiktionary page for the phrase "a camel is a horse designed by". The page title is "a camel is a horse designed by a committee". The main content is a discussion forum post by Ruben P. Mendez on April 16, 2004, asking for the origin of the maxim. The post includes a search bar and a list of related entries. The page also features a sidebar with navigation links and a "The Phrase Finder" header.

## People *want* to ask questions?

### Examples of search queries

- who invented surf music?
- how to make stink bombs
- where are the snowdens of yesteryear?
- which english translation of the bible is used in official catholic liturgies?
- how to do clayart
- how to copy psx
- how tall is the sears tower?
- how can i find someone in texas
- where can i find information on puritan religion?
- what are the 7 wonders of the world
- how can i eliminate stress
- What vacuum cleaner does Consumers Guide recommend

Around 10–15% of query logs

## AskJeeves (Classic)

---

- Probably the most hyped example of “question answering”
- It largely did pattern matching to match your question to their own knowledge base of questions
- If that works, you get the human-curated answers to that known question (which are presumably good)
- If that fails, it falls back to regular web search
- A potentially interesting middle ground, but not full QA

## A Brief (Academic) History

---

- Question answering is not a new research area
- Question answering systems can be found in many areas of NLP research, including:
  - Natural language database systems
    - A lot of early NLP work on these
  - Spoken dialog systems
    - Currently very active and commercially relevant
- The focus on open-domain QA is (relatively) new
  - MURAX (Kupiec 1993): Encyclopedia answers
  - Hirschman: Reading comprehension tests
  - TREC QA competition: 1999–

## Question Answering at TREC

---

- Question answering competition at TREC consists of answering a set of 500 fact-based questions, e.g., “*When was Mozart born?*”.
- For the first three years systems were allowed to return 5 ranked answer snippets (50/250 bytes) to each question.
  - IR think
  - Mean Reciprocal Rank (MRR) scoring:
    - 1, 0.5, 0.33, 0.25, 0.2, 0 for 1, 2, 3, 4, 5, 6+ doc
  - Mainly Named Entity answers (person, place, date, ...)
- From 2002 the systems are only allowed to return a single *exact* answer and the notion of confidence has been introduced.

## The TREC Document Collection

---

- One round about 10 years ago: news articles from:
  - AP newswire, 1998-2000
  - New York Times newswire, 1998-2000
  - Xinhua News Agency newswire, 1996-2000
- In total 1,033,461 documents in the collection.
- 3GB of text
- While small in some sense, still too much text to process using advanced NLP techniques (on the fly at least)
- Systems usually have initial information retrieval followed by advanced processing.
- Many supplement this text with use of the web, and other knowledge bases

## Sample TREC questions

---

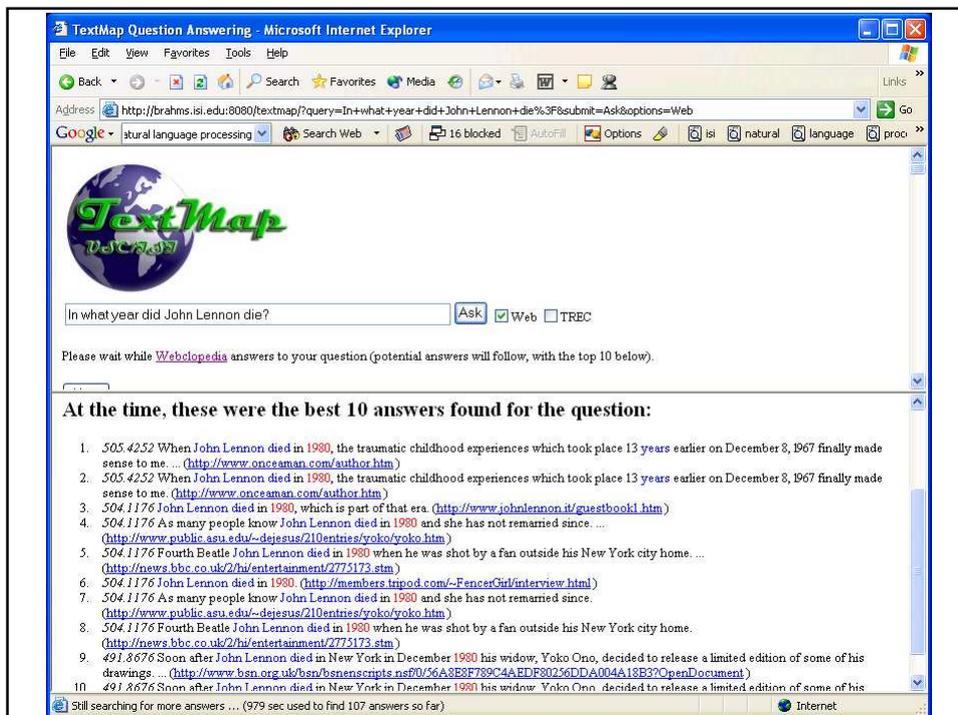
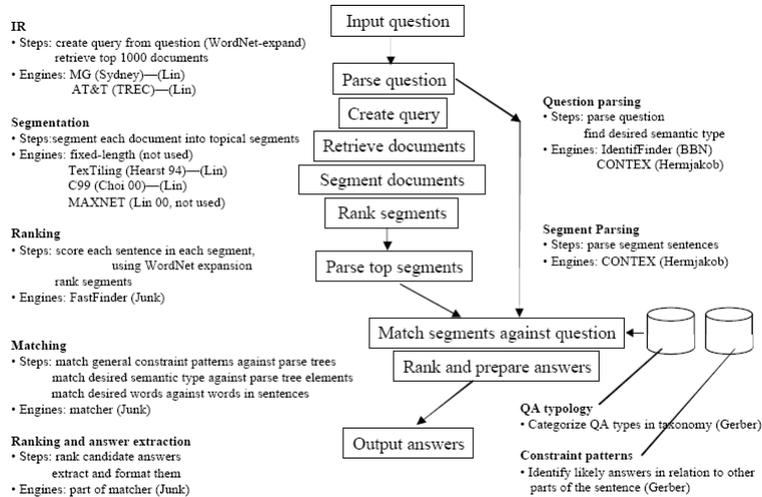
1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Qintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?

## Top Performing Systems

---

- Currently the best performing systems at TREC can answer approximately 70% of the questions
- Approaches and successes have varied a fair deal
  - Knowledge-rich approaches, using a vast array of NLP techniques stole the show in 2000, 2001, still do well
    - Notably Harabagiu, Moldovan et al. – SMU/UTD/LCC
  - AskMSR system stressed how much could be achieved by very simple methods with enough text (and now various copycats)
  - Middle ground is to use large collection of surface matching patterns (ISI)

# Webclopedia Architecture



TextMap Question Answering - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://brahms.isi.edu:8080/textmap/?query=Who+was+the+prime+minister+of+Australia+in+1990%3F&options=Web>

Who was the prime minister of Australia in 1990?   Web  TRECC

Please wait while [Weblogpedia](#) answers to your question (potential answers will follow, with the top 10 below).

**Current top 10 (of 109) for "Who was the prime minister of Australia in 1990?" - still finding more...**

- 393.0737 ... 1986, 1987, 1990) Prime Minister Paul Keating (1992, 1994, 1995) Prime Minister John Howard (1996, 1999, 2000, 2001, 2003). (2) Visits to Australia from Japan ... (<http://www.mofa.go.jp/region/asia-paci/australia/>)
- 354.7754 ... sound file Prime Minister R3 Memmes opens 'Australia Calling' - 1939 ... 1990 BACK. image file Ian McNamara, presenter of Australia All Over - c. 1990 ... (<http://www.sbc.net.au/ra/gallery/gallery.htm>)
- 352.0389 John Howard is the prime minister of Australia, Niall is something he unfortunately stepped in, though he managed to scape most of it off on the curb ([http://splenerville.com/~blith33/blog/ml-comments.cgi?entry\\_id=5158](http://splenerville.com/~blith33/blog/ml-comments.cgi?entry_id=5158))
- 352.0389 Mr. Howard is the prime minister of Australia (<http://www.opinionjournal.com/editorial/feature.html?id=110003123>)
- 347.6498 Visit to Japan by Prime Minister John Winston Howard of Australia (July 2003) (<http://www.mofa.go.jp/region/asia-paci/australia/>)
- 345.3444 You can send your letter to The Honorable John Howard, MP Prime Minister of Australia, Parliament House, Canberra, ACT 2600 Australia. (<http://www.planetary.org/html/news/articles/headers/1997/heads/041497.html>)
- 336.8657 Prime Minister Junichiro Koizumi held talks with Mr. John Howard, the Prime Minister of Australia, at the Prime Minister's Official Residence. ([http://www.kantei.go.jp/foreign/koizumiphoto/2003/07/16australia\\_e.html](http://www.kantei.go.jp/foreign/koizumiphoto/2003/07/16australia_e.html))
- 334.0130 Summit Meeting Between Prime Minister Junichiro Koizumi and Prime Minister John Winston Howard of Australia (Summary) (May 2002) (<http://www.mofa.go.jp/region/asia-paci/australia/>)

Still searching for more answers ... (397 sec used to find 116 answers so far)

TextMap Question Answering - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://brahms.isi.edu:8080/textmap/?query=What+is+gum+arabic+made+from%3F&options=Web>

What is gum arabic made from?   Web  TRECC

Please wait while [Weblogpedia](#) answers to your question (potential answers will follow, with the top 10 below).

Searching for answer sentences on the subject: **gum arabic**

Expecting an answer of type: **Q-COMPONENTS, S-NP (0.8)**

[View analysis of question](#)

- 37.0068 From a North African conifer a sap was extracted and boiled to make an alcohol-based ... 1 ounce gum arabic (about 2 tablespoons) 1/2 cup water 2 teaspoons honey 1 ... (<http://www.care2.com/channels/solutions/home/1329>)
- 34.2060 You have your gum arabic solution made in five minutes of not so hard work, and all you need now is to allow some time for all the dirt to precipitate to the ... (<http://www.microscopy-uk.org.uk/mag/artmag03/wdpart3c.html>)
- 34.2060 You have your gum arabic solution made in five minutes of not so hard work, and all you need now is to allow some time for all the dirt to precipitate to the bottom (perhaps overnight.) 5) Decant the supernatant liquid and adjust the density to suit your preferences. (<http://www.microscopy-uk.org.uk/mag/artmag03/wdpart3c.html>)
- 34.0466 Unfortunately, dried (crystallized) pure gum arabic is very brittle: paints made with a high proportion of gum binder often lift easily (redissolve) from the paper, but also will bronze (appear shiny or leathery) and crack or flake if applied in a thick or masstone layer. ([http://ttt.uwv.es/~gbenet/teoria%20de%20color/water\\_color/pigmit.html](http://ttt.uwv.es/~gbenet/teoria%20de%20color/water_color/pigmit.html))
- 34.0219 Made from the dried sap of the Sahelian acacia tree, gum arabic sales totaled two-thirds of all U.S. imports from Sudan in 1996. ([http://www.pponline.org/ppi\\_ci.cfm?lnlgAreaID=108&subsecID=900003&contentID=251860](http://www.pponline.org/ppi_ci.cfm?lnlgAreaID=108&subsecID=900003&contentID=251860))
- 33.5295 ... gum arabic or gum acacia - Hardened sap ... gyotaku - In Japanese tradition, a relief print made from an actual fish, and sometimes from a shell, leaf, or other ... (<http://www.artlex.com/ArtLex/Gm.html>)
- 33.1687 gum arabic n. : gum from an acacia tree; used as a thickener (especially in candies and pharmaceuticals) [syn: gum acacia] (<http://dict.die.net/gum%20arabic/>)
- 32.8439 Sugar-coated confections made by the panning process employ gum arabic solutions to provide an adhesive and film coating for nuts, candy corn, jelly beans, bridge mixes, and others. (<http://www.jumbo.th.com/application.html>)

Still searching for more answers ... (149 sec used to find 29 answers so far)

## Ravichandran and Hovy 2002 Learning Surface Patterns

---

- Use of Characteristic Phrases
- "When was <person> born"
  - Typical answers
    - "Mozart was born in 1756."
    - "Gandhi (1869-1948)..."
  - Suggests phrases like
    - "<NAME> was born in <BIRTHDATE>"
    - "<NAME> (<BIRTHDATE>-"
  - Regular expressions

## Use Pattern Learning

---

- Example: Start with "Mozart 1756"
  - Results:
    - "The great composer Mozart (1756-1791) achieved fame at a young age"
    - "Mozart (1756-1791) was a genius"
    - "The whole world would always be indebted to the great music of Mozart (1756-1791)"
  - Longest matching substring for all 3 sentences is "Mozart (1756-1791)"
  - Suffix tree would extract "Mozart (1756-1791)" as an output, with score of 3
- Reminiscent of IE pattern learning

## Pattern Learning (cont.)

---

- Repeat with different examples of same question type
  - “Gandhi 1869”, “Newton 1642”, etc.
- Some patterns learned for BIRTHDATE
  - a. born in <ANSWER>, <NAME>
  - b. <NAME> was born on <ANSWER> ,
  - c. <NAME> ( <ANSWER> -
  - d. <NAME> ( <ANSWER> - )

## Experiments: (R+H, 2002)

---

- Some question types from Webclopedia QA Typology (Hovy et al., 2002a)
  - BIRTHDATE
  - LOCATION
  - INVENTOR
  - DISCOVERER
  - DEFINITION
  - WHY-FAMOUS

## Experiments: Pattern Precision

---

- **BIRTHDATE table:**
  - 1.0 <NAME> ( <ANSWER> - )
  - 0.85 <NAME> was born on <ANSWER> ,
  - 0.6 <NAME> was born in <ANSWER>
  - 0.59 <NAME> was born <ANSWER>
  - 0.53 <ANSWER> <NAME> was born
  - 0.50 - <NAME> ( <ANSWER>
  - 0.36 <NAME> ( <ANSWER> -
- **INVENTOR**
  - 1.0 <ANSWER> invents <NAME>
  - 1.0 the <NAME> was invented by <ANSWER>
  - 1.0 <ANSWER> invented the <NAME> in

## Experiments (cont.)

---

- **WHY-FAMOUS**
  - 1.0 <ANSWER> <NAME> called
  - 1.0 laurate <ANSWER> <NAME>
  - 0.71 <NAME> is the <ANSWER> of
- **LOCATION**
  - 1.0 <ANSWER>'s <NAME>
  - 1.0 regional : <ANSWER> : <NAME>
  - 0.92 near <NAME> in <ANSWER>
- Depending on question type, get high MRR (0.6–0.9), with higher results from use of Web than TREC QA collection

## Shortcomings & Extensions

---

- Need for POS &/or semantic types
  - "Where are the Rocky Mountains?"
  - "Denver's new airport, topped with white fiberglass cones in imitation of the Rocky Mountains in the background , continues to lie empty"
  - <NAME> in <ANSWER>
- NE tagger or ontology could enable system to determine "background" is not a location

## Shortcomings... (cont.)

---

- Long distance dependencies
  - "Where is London?"
  - "London, which has one of the busiest airports in the world, lies on the banks of the river Thames"
  - would require pattern like:  
<QUESTION>, (<any\_word>)\*, lies on <ANSWER>
- But: abundance & variety of Web data helps system to find an instance of patterns w/o losing answers to long distance dependencies

## Shortcomings... (cont.)

- Their system uses only one anchor word
  - Doesn't work for Q types requiring multiple words from question to be in answer
    - "In which county does the city of Long Beach lie?"
    - "Long Beach is situated in Los Angeles County"
    - required pattern:  
<Q\_TERM\_1> is situated in <ANSWER> <Q\_TERM\_2>
- Does not use case
  - "What is a micron?"
  - "...a spokesman for Micron, a maker of semiconductors, said SIMMs are..."

## AskMSR

- **Web Question Answering: Is More Always Better?**
  - Dumais, Banko, Brill, Lin, Ng (Microsoft, MIT, Berkeley)
- Q: "Where is the Louvre located?"
- Want "Paris" or "France" or "75058 Paris Cedex 01" or a map
- Don't just want URLs

The screenshot shows a Google search interface with the query "Where is the Louvre museum located?". The search results page displays several entries, including a PDF document titled "An Analysis of the AskMSR Question-Answering System" and a hotel listing for "Hotel Montpensier" located near the Louvre Museum. A sidebar on the right contains suggestions for "Paris Metro" and "Paris hotels".

# AskMSR: Shallow approach

- *In what year did Abraham Lincoln die?*
- Ignore hard documents and find easy ones

**Abraham Lincoln, 1809-1865**

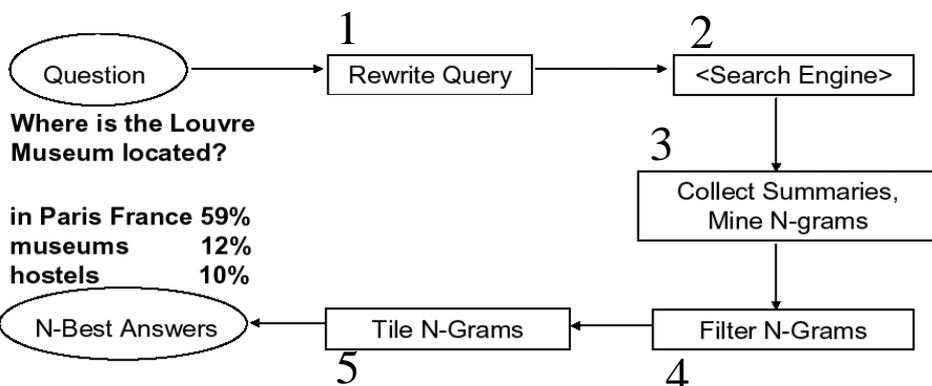
**ABRAHAM LINCOLN**  
Sixteenth President of the United States  
Born in 1809 - Died in 1865

**Sixteenth President**  
1861-1865  
Married to Mary Todd Lincoln

**Abraham Lincoln**  
16th President of the United States (March 4, 1861 to April 15, 1865)  
Born: February 12, 1809, in Hardin County, Kentucky  
Died: April 15, 1865, at Petersen's Boarding House in Washington, D.C.

"I was born February 12, 1809, in Hardin County, Kentucky. My parents were both born in Virginia, of undistinguished families, perhaps I should say. My mother, who died in my tenth year, was of a family of the name of Lincoln."

# AskMSR: Details



## Step 1: Rewrite queries

- Intuition: The user's question is often syntactically quite close to sentences that contain the answer
  - Where is the Louvre Museum located?
  - The Louvre Museum is located in **Paris**
  - Who created the character of Scrooge?
  - **Charles Dickens** created the character of Scrooge.

## Query Rewriting: Variations

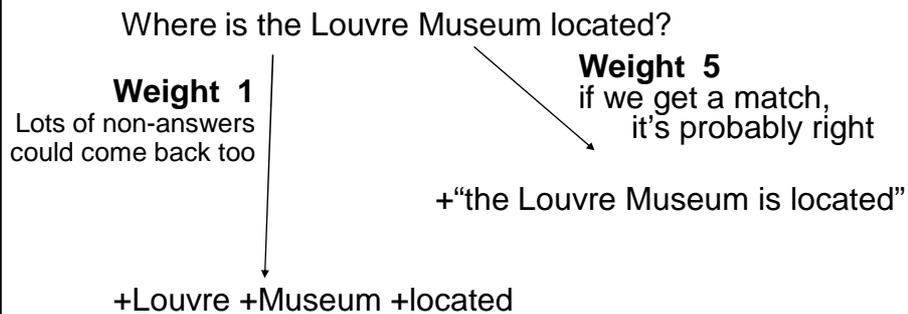
- Classify question into seven categories
  - **Who** is/was/are/were...?
  - **When** is/did/will/are/were ...?
  - **Where** is/are/were ...?
    - a. Category-specific transformation rules  
eg "For Where questions, move 'is' to all possible locations"  
"Where is the Louvre Museum located"
      - "is the Louvre Museum located"
      - "the is Louvre Museum located"
      - "the Louvre is Museum located"
      - "the Louvre Museum is located"
      - "the Louvre Museum located is"
    - b. Expected answer "Datatype" (eg, Date, Person, Location, ...)  
**When** was the French Revolution? → DATE
- Hand-crafted classification/rewrite/datatype rules  
(Could they be automatically learned?)

Nonsense,  
but who  
cares? It's  
only a few  
more queries

## Query Rewriting: Weights

---

- One wrinkle: Some query rewrites are more reliable than others



## Step 2: Query search engine

---

- Send all rewrites to a search engine
- Retrieve top N answers (100?)
- For speed, rely just on search engine's “snippets”, not the full text of the actual document

## Step 3: Mining N-Grams

---

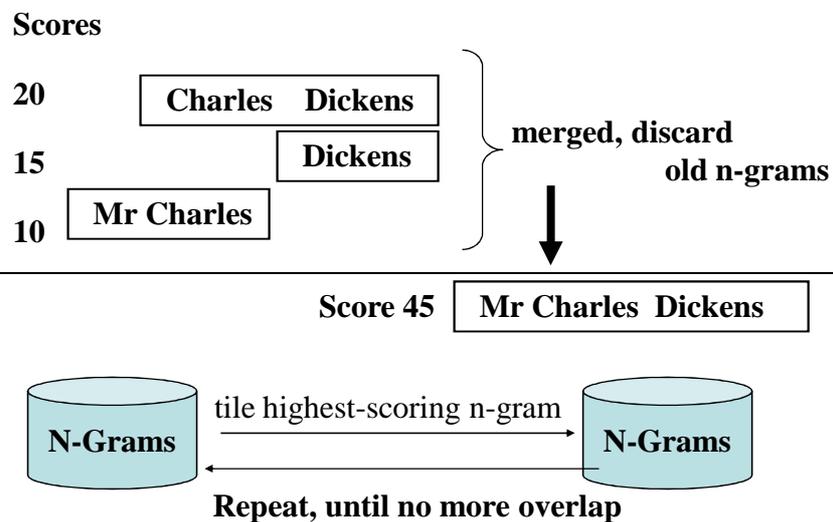
- Simple: Enumerate all N-grams (N=1,2,3 say) in all retrieved snippets
- Weight of an n-gram: occurrence count, each weighted by “reliability” (weight) of rewrite that fetched the document
- Example: “Who created the character of Scrooge?”
  - Dickens - 117
  - Christmas Carol - 78
  - Charles Dickens - 75
  - Disney - 72
  - Carl Banks - 54
  - A Christmas - 41
  - Christmas Carol - 45
  - Uncle - 31

## Step 4: Filtering N-Grams

---

- Each question type is associated with one or more “**data-type filters**” = regular expression
- When... → **Date**
- Where... → **Location**
- What ... → **Person**
- Who ... → **Person**
- Boost score of n-grams that do match regexp
- Lower score of n-grams that don't match regexp
- Details omitted from paper....

## Step 5: Tiling the Answers



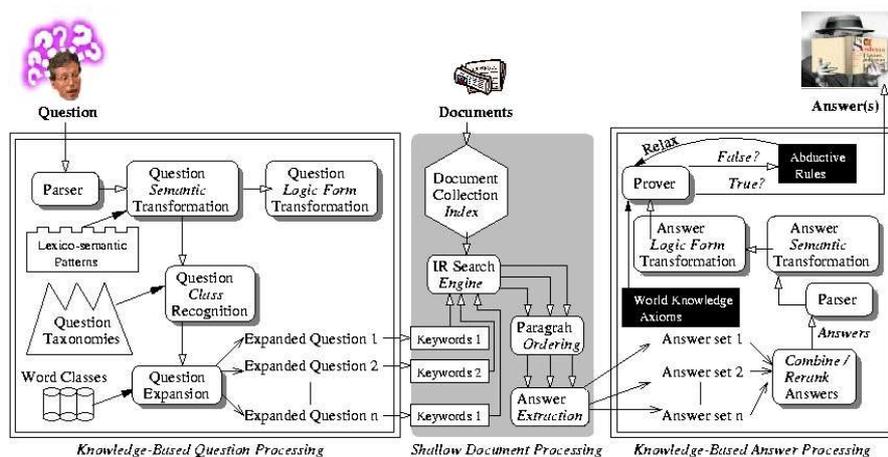
## Results

- Standard TREC contest test-bed:
  - ~1M documents; 900 questions
- Technique doesn't do too well (though would have placed in top 9 of ~30 participants!)
  - MRR = 0.262 (ie, right answered ranked about #4-#5 on average)
  - Why? Because it relies on the redundancy of the Web
- Using the Web as a whole, not just TREC's 1M documents... MRR = 0.42 (ie, on average, right answer is ranked about #2-#3)

# Issues

- In many scenarios (e.g., monitoring an individual's email...) we only have a small set of documents
- Works best/only for "Trivial Pursuit"-style fact-based questions
- Limited/brittle repertoire of
  - question categories
  - answer data types/filters
  - query rewriting rules

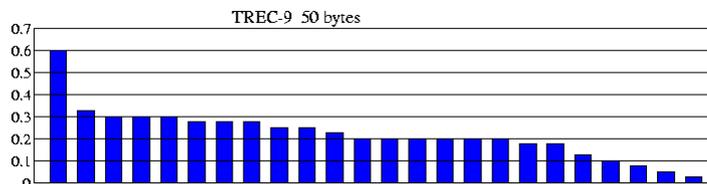
# LCC: Harabagiu, Moldovan et al.



## Value from Sophisticated NLP Pasca and Harabagiu (2001)

---

- Good IR is needed: SMART paragraph retrieval
- Large taxonomy of question types and expected answer types is crucial
- Statistical parser used to parse questions and relevant text for answers, and to build KB
- Query expansion loops (morphological, lexical synonyms, and semantic relations) important
- Answer ranking by simple ML method



## Abductive inference

---

- System attempts inference to justify an answer (often following lexical chains)
- Their inference is a kind of funny middle ground between logic and pattern matching
- But quite effective: 30% improvement
- Q: *When was the internal combustion engine invented?*
- A: *The first internal-combustion engine was built in 1867.*
- invent -> create\_mentally -> create -> build

## Question Answering Example

- How hot does the inside of an active volcano get?
- “lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit”
  - volcano ISA mountain
  - lava ISPARTOF volcano    ■ lava inside volcano
  - fragments of lava HAVEPROPERTIESOF lava
- The needed semantic information is in WordNet definitions, and was successfully translated into a form that was used for rough ‘proofs’

## Example of Complex Question

*How have thefts impacted on the safety of Russia’s nuclear navy, and has the theft problem been increased or reduced over time?*

Need of domain knowledge

*To what degree do different thefts put nuclear or radioactive materials at risk?*

Question decomposition

Definition questions:

- What is meant by nuclear navy?
- What does ‘impact’ mean?
- How does one define the increase or decrease of a problem?

Factoid questions:

- What is the number of thefts that are likely to be reported?
- What sort of items have been stolen?

Alternative questions:

- What is meant by Russia? Only Russia, or also former Soviet facilities in non-Russian republics?