

# Question Answering

## Question-Answering Systems

- Beyond retrieving relevant documents -- Do people want answers to particular questions?
- One of the oldest problem in NLP
- Different kinds of systems
  - Finding answers in document collections (www, encyclopaedia, books, manuals, medical literature, scientific papers, etc.)
  - Interfaces to relational databases
  - Mixed initiative dialog systems
- A scientific reason to do Q/A: the ability to answer questions about a story is the hallmark of understanding.

## Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S  
"AN ACCOUNT OF THE PRINCIPALITIES OF  
WALLACHIA AND MOLDOVIA"  
INSPIRED THIS AUTHOR'S  
MOST FAMOUS NOVEL



Bram Stoker

3



4

## Apple's Siri



5

## Multiple Document Question Answering

- A **multiple document Q/A task** involves questions posed against a collection of documents.
- The answer may appear in the collection multiple times, or may not appear at all! For this task, it doesn't matter where the answer is found.
- Applications include WWW search engines, and searching text repositories such as news archives, medical literature, or scientific articles.

## TREC-9 Q/A Task

Number of Documents:	979,000
Megabytes of Text:	3033
Document Sources:	AP, WSJ, Financial Times, San Jose Mercury News, LA Times, RBIS
Number of Questions:	682
Question Sources:	Encarta log, Excite log

### Sample questions:

- How much folic acid should an expectant mother get daily?
- Who invented the paper clip?
- What university was Woodrow Wilson president of?
- Where is Rider College located?
- Name a film in which Jude law acted.
- Where do lobsters like to live?

## TREC (Text REtrieval Conference) and ACQUAINT

- TREC-10: new questions from MSNSearch logs and AskJeeves, some of which have no answers, or require fusion across documents
- List questions: Name 32 countries Pope John Paul II has visited.
- “Dialogue” processing: Which museum in Florence was damaged by a major bomb explosion in 1993? On what day did this happen?
- ACQUAINT : Advanced Question and Answering for INTElligence (e.g., beyond factoids, the Multiple Perspective Q-A work at Pitt)
- **Current tracks at <http://trec.nist.gov/tracks.html>**

## Single Document Question Answering

- A **single document Q/A task** involves questions associated with one particular document.
- In most cases, the assumption is that the answer appears somewhere in the document and probably once.
- Applications involve searching an individual resource, such as a book, encyclopedia, or manual.
- Reading comprehension tests are also a form of single document question answering.

## Reading Comprehension Tests

### Mars Polar Lander- Where Are You?

(January 18, 2000) After more than a month of searching for a single sign from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life even existed on Mars. Polar Lander was to have touched down on December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. The last effort to communicate with the three-legged lander ended with frustration at 8 a.m. Monday. "We didn't see anything," said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion laboratory. The failed mission to the Red Planet cost the American government more the \$200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.

- When did the mission controllers lost Hope of communication with the Lander?  
(Answer:
- Who is the Polar Lander's project manager?  
(Answer:
- Where on Mars was the spacecraft supposed to touch down?  
(Answer:
- What was the mission of the Mars Polar Lander?  
(Answer:

## Reading Comprehension Tests

### Mars Polar Lander- Where Are You?

(January 18, 2000) After more than a month of searching for a single sign from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life even existed on Mars. Polar Lander was to have touched down on December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. The last effort to communicate with the three-legged lander ended with frustration at 8 a.m. Monday. "We didn't see anything," said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion laboratory. The failed mission to the Red Planet cost the American government more the \$200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.

- When did the mission controllers lost Hope of communication with the Lander?  
(Answer: 8AM, Monday Jan. 17)
- Who is the Polar Lander's project manager?  
(Answer: Richard Cook)
- Where on Mars was the spacecraft supposed to touch down?  
(Answer: near Mars' south pole)
- What was the mission of the Mars Polar Lander?  
(Answer: to study Mars' atmosphere and search for water)

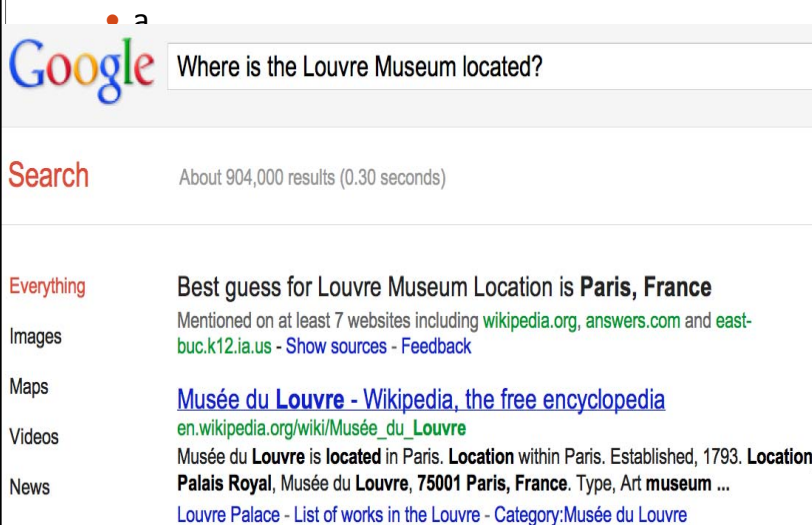
## Question Types

- Simple (factoid) questions (most commercial systems)
  - *Who wrote the Declaration of Independence?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions
  - *What do scholars think about Jefferson's position on dealing with pirates?*
  - *What is a Hajj?*
  - *In children with an acute febrile illness, what is the efficacy of single medication therapy with acetaminophen or ibuprofen in reducing fever?*
- Complex (opinion) questions
  - *Was the Gore/Bush election fair?*

## Commercial systems: mainly factoid questions

Where is the Louvre Museum located?	In Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What is the telephone number for Stanford University?	650-723-2300

## IR-based Question Answering (e.g., TREC, Google)



The screenshot shows a Google search interface. At the top, the Google logo is on the left, and a search bar contains the text "Where is the Louvre Museum located?". Below the search bar, the word "Search" is on the left, and "About 904,000 results (0.30 seconds)" is on the right. On the left side, there is a vertical list of filters: "Everything", "Images", "Maps", "Videos", and "News". The "Everything" filter is selected. The main content area displays the search results. The top result is "Best guess for Louvre Museum Location is Paris, France". Below this, it says "Mentioned on at least 7 websites including wikipedia.org, answers.com and east-buc.k12.ia.us - Show sources - Feedback". The next result is "Musée du Louvre - Wikipedia, the free encyclopedia" with a link to "en.wikipedia.org/wiki/Musée\_du\_Louvre". Below this, it says "Musée du Louvre is located in Paris. Location within Paris. Established, 1793. Location, Palais Royal, Musée du Louvre, 75001 Paris, France. Type, Art museum ...". The final result is "Louvre Palace - List of works in the Louvre - Category:Musée du Louvre".

Google Where is the Louvre Museum located?

Search About 904,000 results (0.30 seconds)

Everything Best guess for Louvre Museum Location is **Paris, France**  
Mentioned on at least 7 websites including [wikipedia.org](#), [answers.com](#) and [east-buc.k12.ia.us](#) - Show sources - Feedback

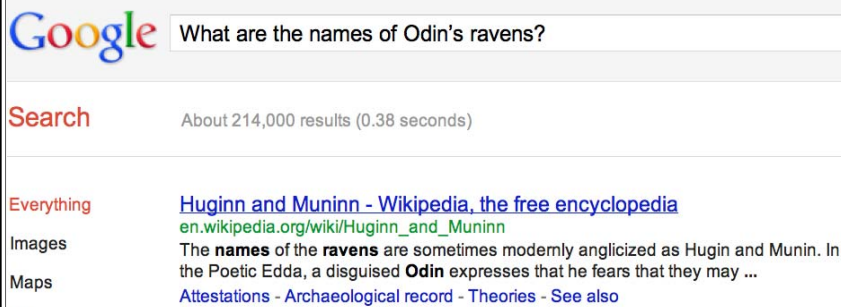
Images

Maps [Musée du Louvre - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Musée\\_du\\_Louvre](#)

Videos Musée du Louvre is located in Paris. Location within Paris. Established, 1793. Location, Palais Royal, Musée du Louvre, 75001 Paris, France. Type, Art museum ...

News Louvre Palace - List of works in the Louvre - Category:Musée du Louvre

Many questions can already be answered by web search



15

## IR-Based (Corpus-based) Approaches

(As opposed to knowledge-based ones)

1. We assume an Information Retrieval (IR) system with an index into documents that plausibly contain the answer(s) to likely questions.
2. And that the IR system can find plausibly relevant documents in that collection given the words in a user's question.

16



## Knowledge-based Approaches (e.g., Siri)

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet)
  - Restaurant review sources and reservation services
  - Scientific databases

17

## Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

18

## Corpus-Based Approaches

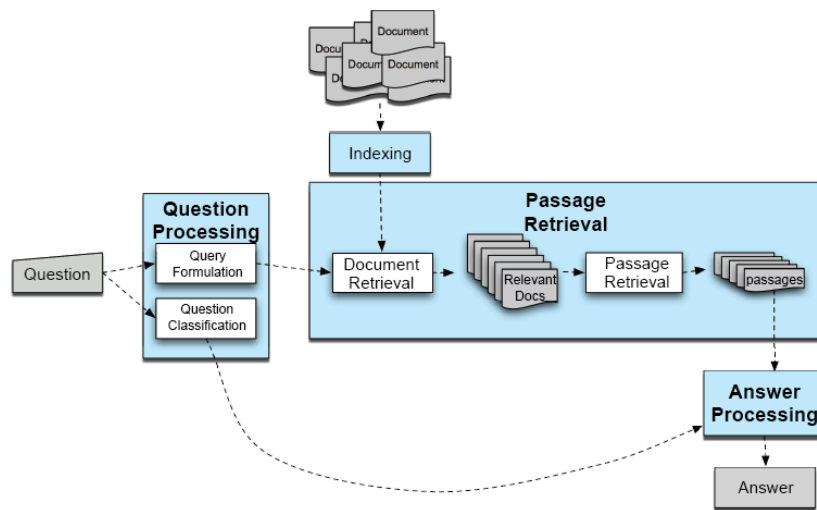
- **Factoid questions**
  - From a smallish collection of relevant documents
  - Extract the answer (or a snippet that contains the answer)
- **Complex questions**
  - From a smallish collection of relevant documents
  - *Summarize* those documents to address the question
    - Query-focused summarization

19

## Factoid questions

Question	Answer
Where is the Louvre Museum located?	in Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	the yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What's the official language of Algeria?	Arabic
What is the telephone number for the University of Colorado, Boulder?	(303)492-1411
How many pounds are there in a stone?	14

## IR-based Factoid Q/A



21

## Full-Blown System

- **QUESTION PROCESSING**
  - Parse and analyze question / Determine type of the answer
    - i.e., detect question type, answer type, focus, relations
  - Formulate queries suitable for use with IR/search engine
- **PASSAGE RETRIEVAL**
  - Retrieve ranked results
  - Break into suitable passages and rerank
- **ANSWER PROCESSING**
  - Perform NLP on those to extract candidate answers
  - Rank candidates based on NLP processing
    - Using evidence from text and external sources

## Question Processing

- Two tasks
  - Answer Type Detection
    - what kind of entity (person, place) is the answer?
  - Query Formulation
    - what is the query to the IR system
- We extract both of these from the question
  - keywords for **query** to the IR system
  - an **answer type**

23

## Things to extract from the question

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction
  - Find relations between entities in the question

24

## Question Processing

They're the two states you could be reentering if you're crossing Florida's northern border

- Answer Type:
- Query:
- Focus:
- Relations:

25

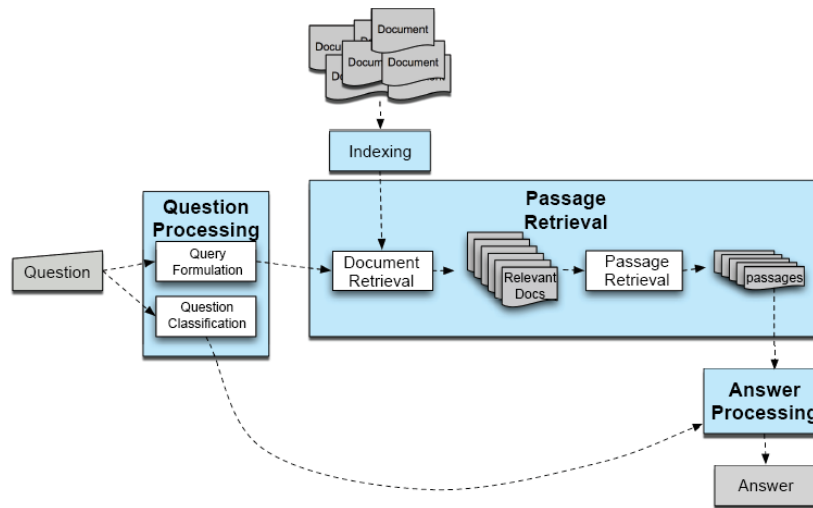
## Question Processing

They're the two states you could be reentering if you're crossing Florida's northern border

- Answer Type: US state
- Query: two states, border, Florida, north
- Focus: the two states
- Relations: borders(Florida, ?x, north)

26

## Factoid Q/A



27

## Answer Type Detection: Named Entities

- *Who founded Virgin Airlines?*
  - PERSON.
- *What Canadian city has the largest population?*
  - CITY.

28

## Answer Types Can Be Complicated

- **Who** questions can have organizations or countries as answers
  - Who sells the most hybrid cars?
  - Who exports the most wheat?

## Answer Types

- Factoid questions...
  - Who, where, when, how many...
    - **Who** questions are going to be answered by...
    - **Where** questions...
  - Simplest: use **Named Entities**

## Sample Text

Consider this sentence:

*President George Bush announced a new bill that would send \$1.2 million dollars to Miami Florida for a new hurricane tracking system.*

After applying a Named Entity Tagger, the text might look like this:

*<Person=“President George Bush”> announced a new bill that would send <MONEY=\$1.2 million dollars”> to <LOCATION=“Miami Florida:”> for a new hurricane tracking system.*

## Rules for Name Entity Tagging

Many Named Entity Taggers use simple rules that are developed by hand. Most rules use the following types of clues:

**Keywords:** Ex. “Mr.”, “Corp.”, “city”

**Common Lists:** Ex. Cities, countries, months of the year, common first names, common last names

**Special Symbols:** Ex. Dollar signs, percent signs

**Structured Phrases:** Ex. Dates often appear as “MONTH, DAY #, YEAR”

**Syntactic Patterns:** (more rarely) Ex. “LOCATIONS\_NP, LOCATION\_NP” is usually a single location (e.g. Boston, Massachusetts).



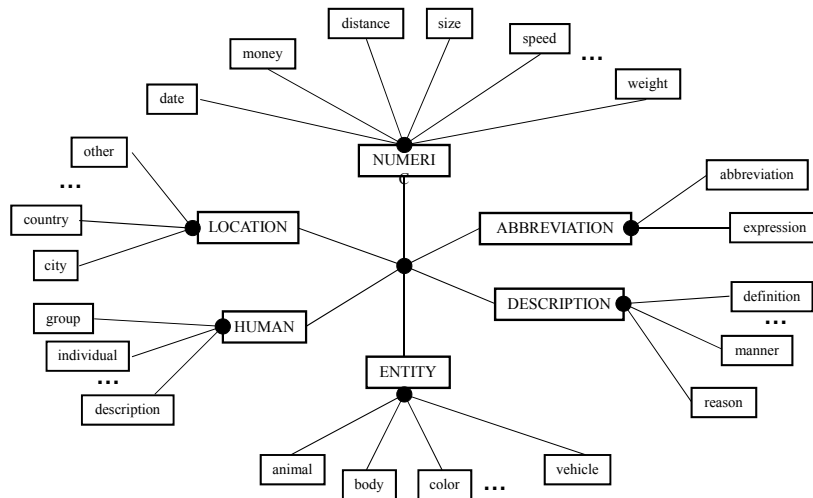
## Named Entities

- If we are looking for questions whose Answer Type is CITY
  - We have to have a “CITY” named-entity detector
- So if we have a rich set of answer types
  - We also have to build answer-detectors for each of those types!

## Answer Type Taxonomy (from Li & Roth, 2002)

- Two-layered taxonomy
- 6 coarse classes
  - ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC\_VALUE
- 50 fine classes
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency...
  - LOCATION: city, country, mountain...

## Answer Type Taxonomy (2/2)



35

## Answer Types

### ENTITY

animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

36

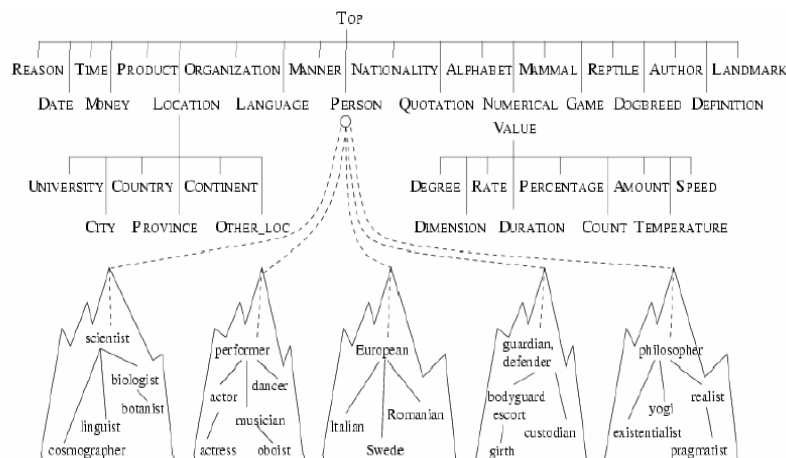
## More Answer Types

<b>HUMAN</b>	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
<b>LOCATION</b>	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
<b>NUMERIC</b>	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

37

## Another Taxonomy of Answer Types

- Contains ~9000 concepts reflecting expected answer types
- Merges NEs with the WordNet hierarchy



## Answer types in Jeopardy

Ferrucci et al. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. Fall 2010. 59-79.

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types  
he, country, city, man, film, state, she, author, group, here, company,  
president, capital, star, novel, character, woman, river, island, king,  
song, part, series, sport, singer, actor, play, team, show, actress,  
animal, presidential, composer, musical, nation, book, title,  
leader, game

## Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

## Answer Type Detection

- Regular expression-based rules can get some cases:
  - Who {is|was|are|were} PERSON
  - PERSON (YEAR – YEAR)
- Make use of answer-type word

## Answer Type Detection

- Other rules use the **question headword**:
  - = headword of first noun phrase after wh-word:
- Which **city** in China has the largest number of foreign financial companies?
- What is the state **flower** of California?

## Answer Type Detection

- Most often, we treat the problem as machine learning classification
  - **Define** a taxonomy of question types
  - **Annotate** training data for each question type
  - **Train** classifiers for each question class using a rich set of features.
    - these features can include those hand-written rules!

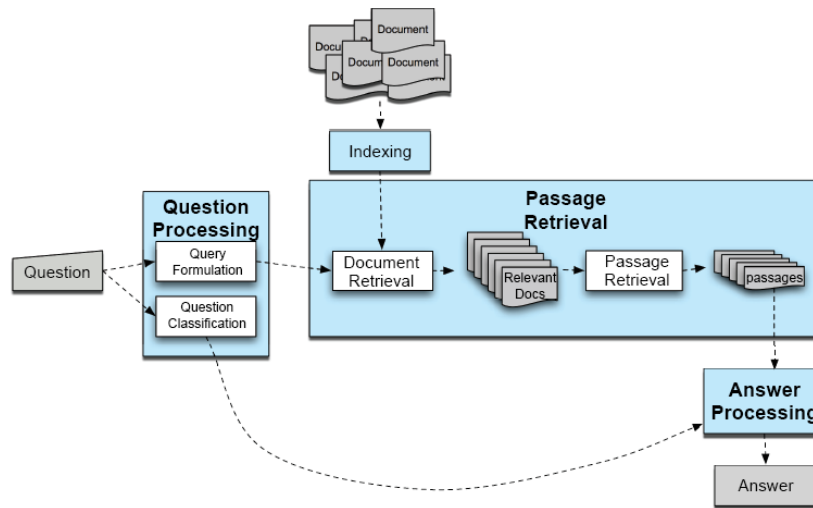
43

## Features for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

44

## Factoid Q/A



## Query Terms Extraction

- Grab terms from the query that will help us find answer passages

Question (from TREC QA track)	keywords
<b>Q002: What was the monetary value of the Nobel Peace Prize in 1989?</b>	monetary, value, Nobel, Peace, Prize, 1989
<b>Q003: What does the Peugeot company manufacture?</b>	Peugeot, company, manufacture
<b>Q004: How much did Mercury spend on advertising in 1993?</b>	Mercury, spend, advertising, 1993
<b>Q005: What is the name of the managing director of Apricot Computer?</b>	name, managing, director, Apricot, Computer

## Keyword Selection Algorithm

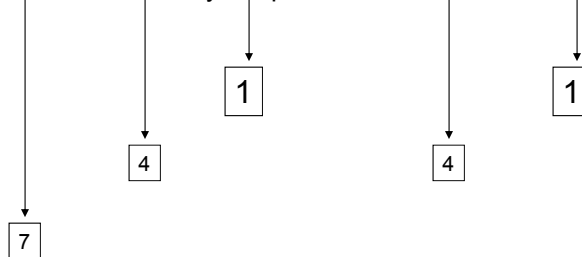
Dan Moldovan, Sanda Harabagiu, Marius Păcă, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. 1999. Proceedings of TREC-8.

1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words

## Choosing keywords from the query

Slide from Mihai Surdeanu

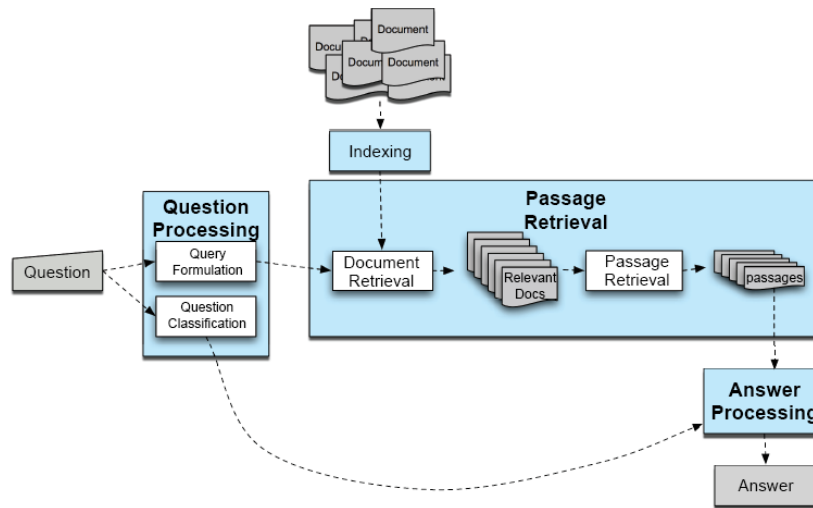
Who coined the term "cyberspace" in his novel "Neuromancer"?



cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7  
48



## Factoid Q/A



49

## Document/Passage Retrieval

- For some applications, the text collection that must be searched is very large. For example, the TREC Q/A collection is about 3 GB!
- Applying NLP techniques to large text collections is too expensive to do in real-time. So, information retrieval (IR) engines identify the most relevant texts, using the question words as key words.
- *Document Retrieval Systems* return the N documents that are most relevant to the question. *Passage retrieval systems* return the N passages that are most relevant to the question.
- Only the most relevant documents/passages are given to the remaining modules of the Q/A system. If the IR engine doesn't retrieve text(s) containing the answer, the Q/A system is out of luck!

## Passage Processing

- Output of IR:
  - Ranked Documents, according to similarity with keywords
- Problems:
  - Documents aren't best unit for QA
  - Ranking for IR may not be appropriate for QA
- So *passage retrieval*:
  - extracts and reranks shorter units from the returned set of documents

51

## Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help rerank passages

52

## Passage Extraction Loop

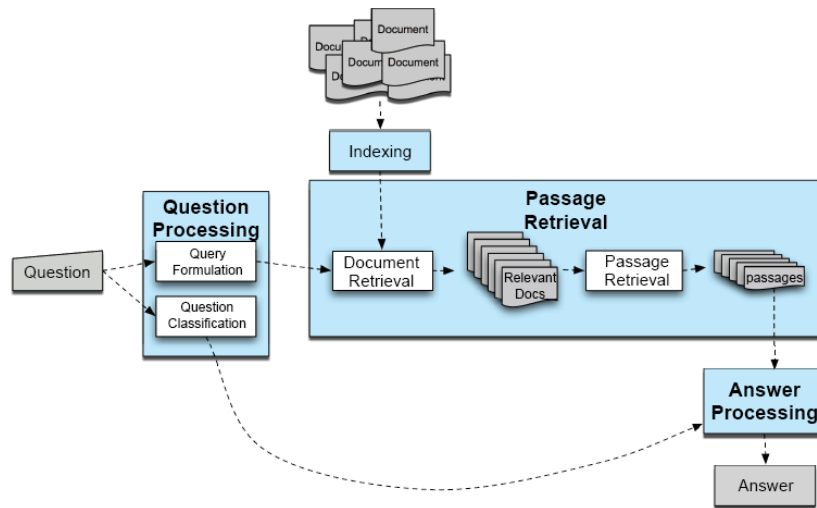
- **Passage Extraction Component**
  - Extracts passages that contain all selected keywords
- **Passage quality and keyword adjustment**
  - If the number of passages is lower than a threshold  $\Rightarrow$  query is too strict  $\Rightarrow$  drop a keyword
  - If the number of passages is higher than a threshold  $\Rightarrow$  query is too relaxed  $\Rightarrow$  add a keyword

## Features for Passage Ranking

Either in rule-based classifiers or with supervised machine learning

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

## Factoid Q/A



55

## Answer Identification/Extraction

At this point, we've assigned a type to the question and we've tagged the text with Named Entities. So we can now narrow down the candidate pool to entities of the right type.

Problem: There are often many objects of the right type, even in a single text.

- The **Answer Identification** module is responsible for finding the best answer to the question.
- For questions that have Named Entity types, this module must figure out which item of the right type is correct.
- For questions that do not have Named Entity types, this module is essentially starting from scratch.

## Answer Extraction

- Pattern-extraction methods:
  - Run an answer-type named-entity tagger on the passages
    - Each answer type requires a named-entity tagger that detects it
    - If answer type is CITY, tagger has to tag CITY
      - Can be full NER, simple regular expressions, or hybrid
  - Return the string that is the right type:
- *Who is the prime minister of India” (PERSON)*
  - Manmohan Singh, Prime Minister of India, had told left leaders that the deal would not be renegotiated.
- *“How tall is Mt. Everest? (LENGTH)*
  - The official height of Mount Everest is 29035 feet

## Answer Extraction

- Pattern-extraction methods:
  - Run an answer-type named-entity tagger on the passages
    - Each answer type requires a named-entity tagger that detects it
    - If answer type is CITY, tagger has to tag CITY
      - Can be full NER, simple regular expressions, or hybrid
  - Return the string that is the right type:
- *Who is the prime minister of India” (PERSON)*
  - **Manmohan Singh**, Prime Minister of India, had told left leaders that the deal would not be renegotiated.
- *“How tall is Mt. Everest? (LENGTH)*
  - The official height of Mount Everest is **29035 feet**

## Answer Extraction

- For answers that are not a particular named entity type:
  - Use regular expression patterns

Pattern	Question	Answer
<AP> such as <QP>	What is autism?	“, developmental disorders such as autism”
<QP>, a <AP>	What is a caldera?	“the Long Valley caldera, a <u>volcanic crater</u> 19 miles long”

- These can be written by hand or learned automatically

## Answer Extraction

- Sometimes pattern-extraction methods are insufficient
  - We can't write rules
  - There is more than one potential answer in the passage

## Ranking (multiple) Candidate Answers

Q066: Name the first private citizen to fly in space.

- Answer type: **Person**
- Text passage:

"Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in "Raiders of the Lost Ark", plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike\_Smith..."

## Ranking Candidate Answers

Q066: Name the first private citizen to fly in space.

- Answer type: **Person**
- Text passage:

"Among them was **Christa McAuliffe**, the first private citizen to fly in space. **Karen Allen**, best known for her starring role in "Raiders of the Lost Ark", plays **McAuliffe**. **Brian Kerwin** is featured as shuttle pilot **Mike Smith**..."
- Best candidate answer: **Christa McAuliffe**

## Word Overlap

One common method of Answer Identification is to measure the amount of **Word Overlap** between the question and an answer candidate.

**Basic Word Overlap:** Each answer candidate is scored by counting how many question words are present in or near the candidate.

**Stop Words:** sometimes closed class words (often called *Stop Words* in IR) are not included in the word overlap measure.

**Stemming:** sometimes morphological analysis is used to compare only the root forms or words (e.g. “walk” and “walked” would match).

**Weights:** some words may be weighted more heavily than others (e.g., verbs might be given more weight than nouns).

## Machine Learning

- In other cases we use machine learning to combine many rich features about which phrase is the answer



## Features for Answer Ranking

**Answer type match:** True if the candidate answer contains a phrase with the correct answer type.

**Pattern match:** The identity of a pattern that matches the candidate answer.

**Number of matched question keywords:** How many question keywords are contained in the candidate answer.

**Keyword distance:** The distance between the candidate answer and query keywords (measured in average number of words or as the number of keywords that occur in the same syntactic phrase as the candidate answer).

**Novelty factor:** True if at least one word in the candidate answer is novel, that is, not in the query.

**Apposition features:** True if the candidate answer is an apposition to a phrase containing many question terms. Can be approximated by the number of question terms separated from the candidate answer through at most three words and one comma (Pasca, 2003).

**Punctuation location:** True if the candidate answer is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

**Sequences of question terms:** The length of the longest sequence of question terms that occurs in the candidate answer.

65

## Apposition

### Apposition

from [English Grammar Today](#)

When we use two noun phrases (np) next to each other in a clause, and they refer to the same person or thing, we call this apposition:

*[NP 1]The living room, [NP 2]the biggest room in the house, looks out on to a beautiful garden.  
(The living room and the biggest room in the house are the same room.)*

*[NP 1]Timothy, [NP 2]their youngest child, is very musical.  
(Timothy and their youngest child are the same person.)*

## Candidate Answer scoring in Watson

- Each candidate answer gets scores from >50 components
  - (from unstructured text, semi-structured text, triple stores)
- logical form (parse) match between question and candidate
- passage source reliability
- geospatial location
  - California is "southwest of Montana"
- temporal relationships
- taxonomic classification

67

## Evaluation

- NIST (National Institute of Standards and Technology) has been running Q/A evaluations as part of it's TREC program. Both generic Q/A and application specific (bio- etc.).
  - (Also does DUC)
- **Accuracy** (does answer match gold answer?)
- Typical metric is **Mean Reciprocal Rank**.
  - Assumes that systems return a ranked list of M possible answers.
  - Your score is based on 1/Rank of the first right answer
    - If first answer is correct: 1
    - else if second answer is correct:  $\frac{1}{2}$
    - else if third answer is correct:  $\frac{1}{3}$ , etc.
    - Score is 0 if none of the M answers are correct
  - Take the mean over all queries

68

## Mean Reciprocal Rank

- What is the reciprocal rank for this question?
- Q: What is the capital of Utah?
  - A1: Ogden
  - A2: Salt Lake City
  - A3: Provo
  - A4: St. George
  - A5: Salt Lake

## Mean Reciprocal Rank

- What is the reciprocal rank for this question?
- Q: What is the capital of Utah?
  - A1: Ogden
  - A2: Salt Lake City
  - A3: Provo
  - A4: St. George
  - A5: Salt Lake
- The score for the question Q would be  $\frac{1}{2}$ .

## Is the Web Different?

- TREC and commercial applications:
  - retrieval performed against small closed collection of texts.
- The diversity/creativity in how people express themselves necessitates all that work to bring the question and the answer texts together.
- But...

## The Web is Different

- On the Web popular factoids are likely to be expressed in many different ways.
- At least a few of which will likely match the way the question was asked.
- So why not just grep the Web using all or pieces of the original question.

## Interim Summary

- factoid question answering
  - answer type detection
  - query formulation
  - passage retrieval
  - passage ranking
  - answer extraction
- web-based factoid question answering

## Question Answering

### Using Knowledge in QA

## Relation Extraction

- Answers: Databases of Relations
  - born-in("Emma Goldman", "June 27 1869")
  - author-of("Cao Xue Qin", "Dream of the Red Chamber")
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions

Whose granddaughter starred in E.T.?

(acted-in ?x "E.T.")

75 (granddaughter-of ?x ?y)

## Temporal Reasoning

- Relation databases
  - (and obituaries, biographical dictionaries, etc.)
- IBM Watson

"In 1594 he took a job as a tax collector in Andalusia"

Candidates:

  - Thoreau is a bad answer (born in 1817)
  - Cervantes is possible (was alive in 1594)

76

## Context and Conversation in Virtual Assistants like Siri

- Coreference helps resolve ambiguities
  - U: “Book a table at Il Fornaio at 7:00 with **my mom**”
  - U: “Also send **her** an email reminder”
- Clarification questions:
  - U: “Chicago pizza”
  - S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”

77

## IBM Watson

- A DeepQA computer system that can compete in real-time at the human champion level on the American TV quiz show Jeopardy
- Grand Challenge Problem for AI/NLP

## Jeopardy Requires a Broad Knowledge Base

- Factual knowledge
  - History, science, politics
- Commonsense knowledge
  - E.g., naïve physics and gender
- Vagueness, obfuscation, uncertainty
  - E.g., "KISS"ing music



## Geospatial knowledge (and earlier temporal knowlege example)

- **Beijing** is a good answer for "Asian city"
- **California** is "southwest of Montana"
- geonames.org:

www.geonames.org/search.html?q=palo+alto&country=

GeoNames Home | Postal Codes | Download / Webservice | About [login](#)

palo alto all countries

[search](#) [show on map](#) [advanced search](#)

459 records found for "palo alto"

	Name	Country	Feature class	Latitude	Longitude
1	<a href="#">Palo Alto</a> Palo Alto, Palo Alto, pa luo ao duo, paeruto, l'ano Amro, l'ano Amro, 帕洛阿尔托, 帕洛阿尔托	<a href="#">United States</a> , California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30"	W 122° 8' 34"
2	<a href="#">Palo Alto Township</a> Palo Alto Township	<a href="#">United States</a> , Iowa Jasper County	administrative division elevation 256m	N 41° 38' 15"	W 93° 2' 57"
3	<a href="#">Borough of Palo Alto</a>	<a href="#">United States</a> , Pennsylvania Schuylkill County	administrative division population 1,032, elevation 210m	N 40° 41' 21"	W 76° 10' 2"



## The Questions: Solution Methods

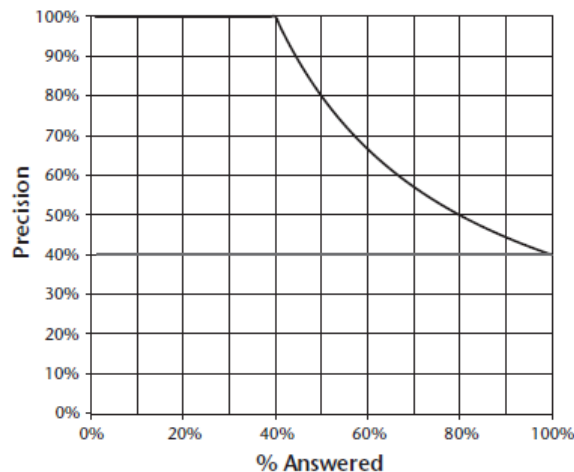
- Factoid questions
 

*Category: Head North*  
*Clue: They're the two states you could be reentering if you're crossing Florida's northern border.*  
*Answer: Georgia and Alabama*
- Decomposition
 

*Category: "Rap" Sheet*  
*Clue: This archaic term for a mischievous or annoying child can also mean a rogue or scamp.*  
*Subclue 1: This archaic term for a mischievous or annoying child.*  
*Subclue 2: This term can also mean a rogue or scamp.*  
*Answer: Rapscallion*
- Puzzles
 

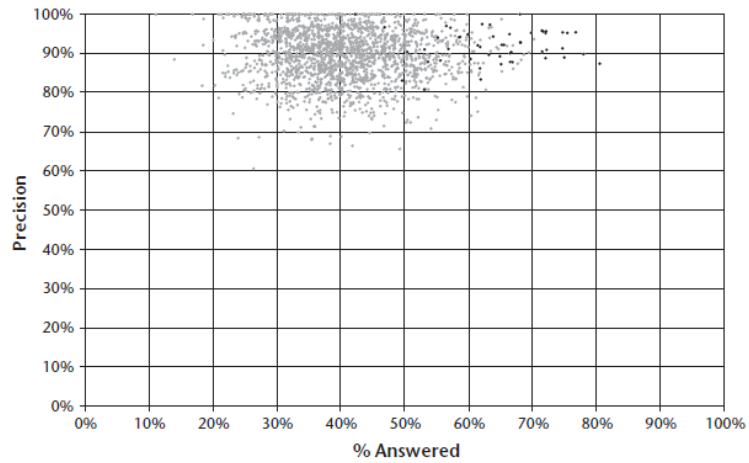
*Category: Rhyme Time*  
*Clue: It's where Pele stores his ball.*  
*Subclue 1: Pele ball (soccer)*  
*Subclue 2: where store (cabinet, drawer, locker, and so on)*  
*Answer: soccer locker*

## Precision vs. Percentage Attempted



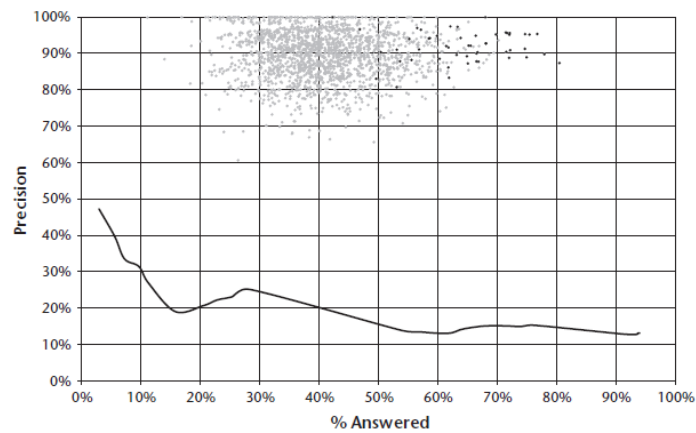
Upper line: perfect confidence estimation

## Champion Human Performance



- Dark dots correspond to Ken Jennings's games

## Baseline Performance



- (IBM) PIQUANT system

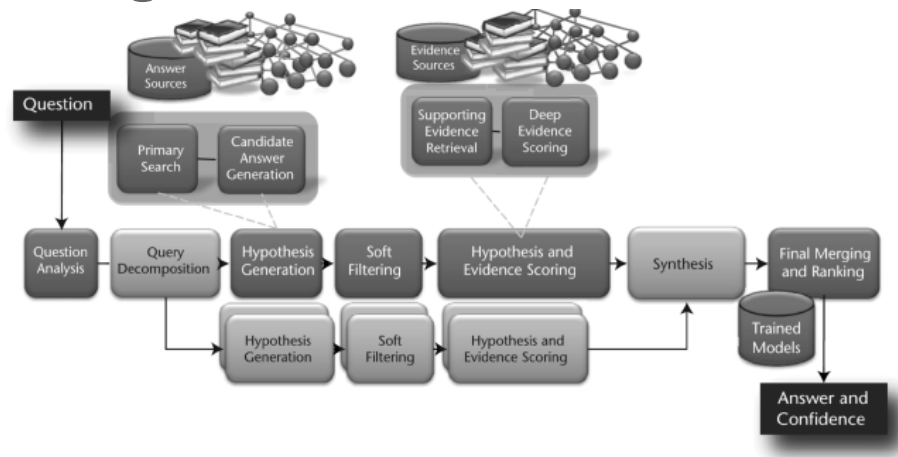
## The DeepQA Approach

- Adapting PIQUANT did not work out
- “The system we have built and are continuing to develop, called DeepQA, is a massively parallel probabilistic evidence-based architecture. For the Jeopardy Challenge, we use more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique we use is how we combine them in DeepQA such that overlapping approaches can bring their strengths to bear and contribute to improvements in accuracy, confidence, or speed.”

## Overarching Principles

- Massive parallelism
- Many experts
  - Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.
- Pervasive confidence estimation
- Integrate shallow and deep knowledge

## High-Level Architecture



## Question Analysis

- “The DeepQA approach encourages a mixture of experts at this stage, and in the Watson system we produce shallow parses, deep parses (McCord 1990), logical forms, semantic role labels, coreference, relations, named entities, and so on, as well as specific kinds of analysis for question answering.”

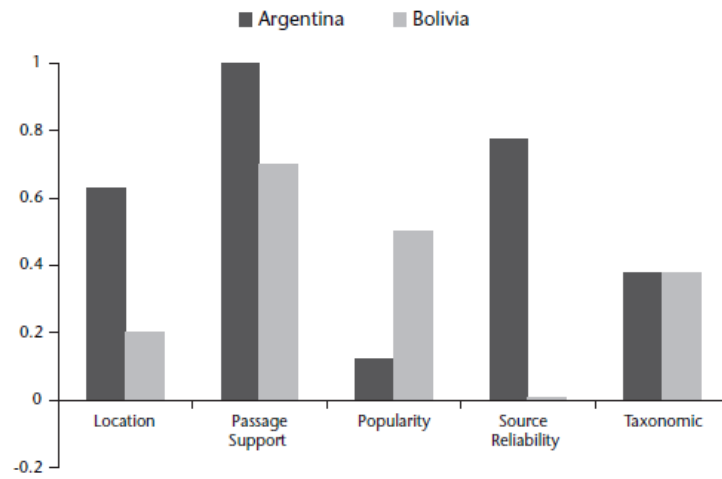
## Hypothesis Generation

- “The operative goal for primary search eventually stabilized at about 85 percent binary recall for the top 250 candidates; that is, the system generates the correct answer as a candidate answer for 85 percent of the questions somewhere within the top 250 ranked candidates.”
- “If the correct answer(s) are not generated at this stage as a candidate, the system has no hope of answering the question. This step therefore significantly favors recall over precision, with the expectation that the rest of the processing pipeline will tease out the correct answer, even if the set of candidates is quite large.”

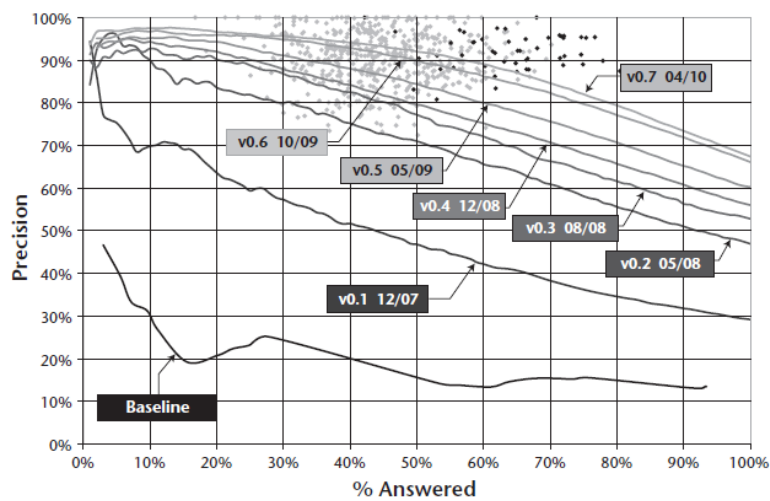
## Hypothesis and Evidence Scoring

- Q: He was presidentially pardoned on September 8, 1974.
- A. Nixon
- Evidence: “Ford pardoned Nixon on Sept. 8, 1974”

## Search Engine Failure (want Argentina)



## Progress



## Results

- Watson wins on Jeopardy!