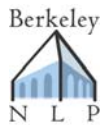


Natural Language Processing



Question Answering

Dan Klein – UC Berkeley

The following slides are largely from Chris Manning, including many slides originally from Sanda Harabagiu, ISI, and Nicholas Kushmerick.

Watson



Large-Scale NLP: Watson

A screenshot of a web browser showing a search interface. The search query is "a camel is a horse designed by?". The results page shows a list of phrases, with "a camel is a horse designed by a committee" selected. Below this, there is a section titled "The Phrase Finder" with a search bar and a "Search" button. The search results show a discussion forum entry for "A camel is a horse designed by committee" posted by Ruben P. Mendez on April 16, 2004. The entry includes a question: "Does anyone know the origin of the maxim? I heard it way back at the United Nations, which is checkfull of committees. It may have originated there, but I'd like an authoritative explanation. Thanks." and two responses: "Re: A camel is a horse designed by committee" by SR 16/Apr/04 and "Re: A camel is a horse designed by committee" by Henry 16/Apr/04.

QA vs Search



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



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–

TREC



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 a notion of confidence has been introduced.



Sample TREC questions

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



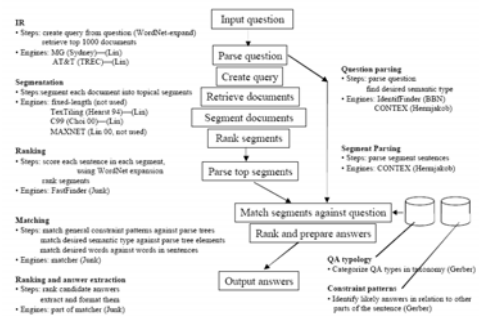
Top Performing Systems

- 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)
 - Emerging standard: analysis, soft-matching, abduction

Pattern Induction: ISI



Webclopedia Architecture



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

Pattern Precision

- WHY-FAMOUS**
 - 1.0 <ANSWER> <NAME> called
 - 1.0 laureate <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>
- 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 of Web data compensates

Aggregation: AskMSR


AskMSR

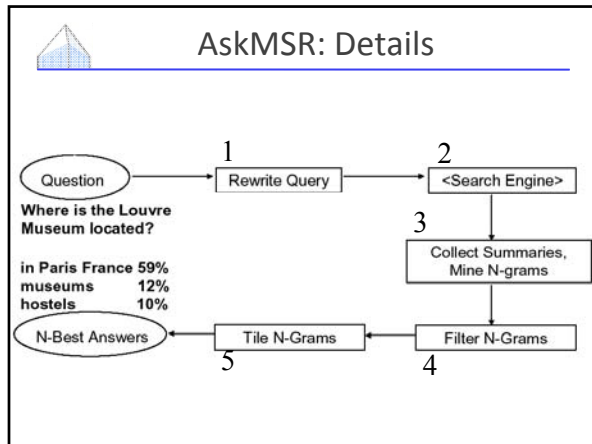
- Web Question Answering: Is More Always Better?**
 - Dumais, Banko, Brill, Lin, Ng (Microsoft, MIT, Berkeley)
- Q: "Where is the Louvre located?"**
- Want "Paris" or "France" or "75058 Paris Cedex 01" or a map**
- Don't just want URLs**



AskMSR: Shallow approach

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

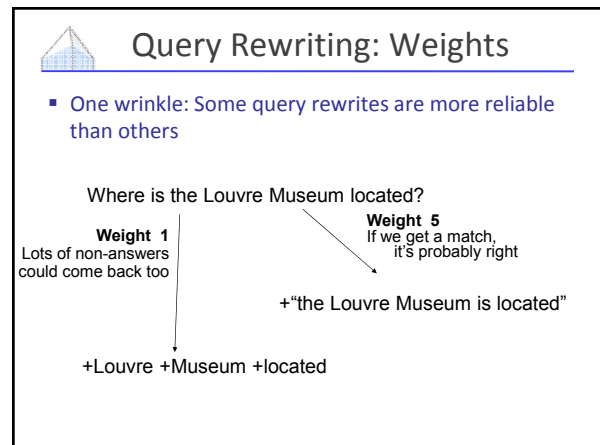




- ## Step 1: Rewrite queries
- Intuition: The user's question is often syntactically quite close to sentences that contain the answer
 - Where is the Louvre Museum located?
 - The Louvre Museum is located in Paris
 - Who created the character of Scrooge?
 - Charles Dickens created the character of Scrooge.

- ## 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"

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



- ## 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

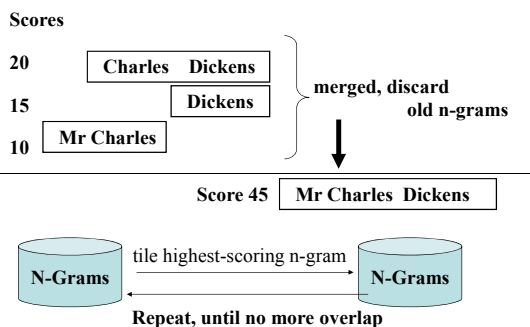


Step 4: Filtering N-Grams

- 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)



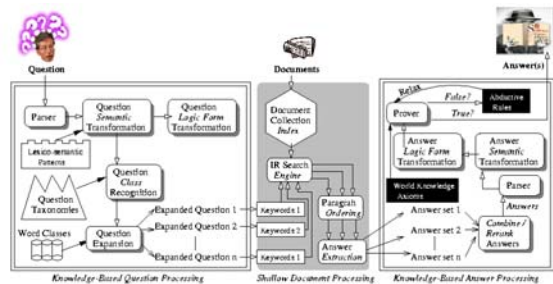
Issues

- In many scenarios (e.g., an individual's email...) we only have a limited 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

Abduction: LCC



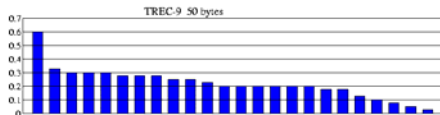
LCC: Harabagiu, Moldovan et al.





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 IN 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'

Watson

Slides from Ferrucci et al, AI Magazine, 2010



Jeopardy...

Category: General Science
Clue: When hit by electrons, a phosphor gives off electromagnetic energy in this form.
Answer: Light (or Photons)

Category: Lincoln Blogs
Clue: Secretary Chase just submitted this to me for the third time; guess what, pal. This time I'm accepting it.
Answer: his resignation

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

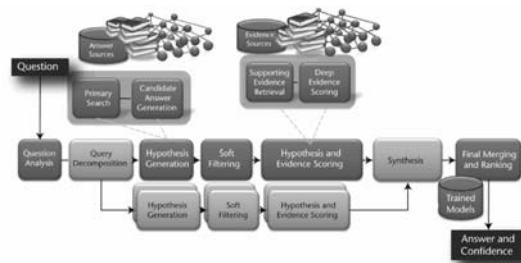
Category: Decorating
Clue: Though it sounds "harsh," it's just embroidery, often in a floral pattern, done with yarn on cotton cloth.
Answer: crewel

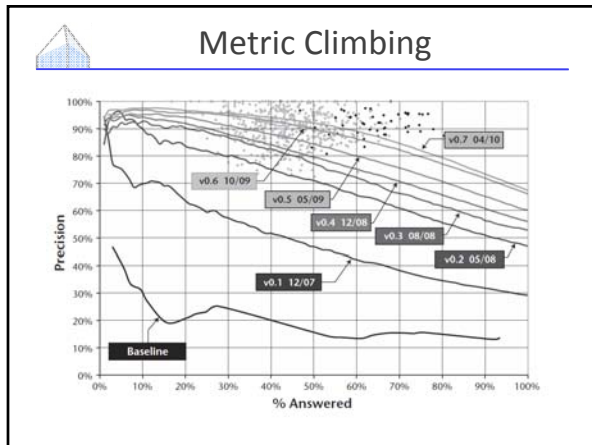
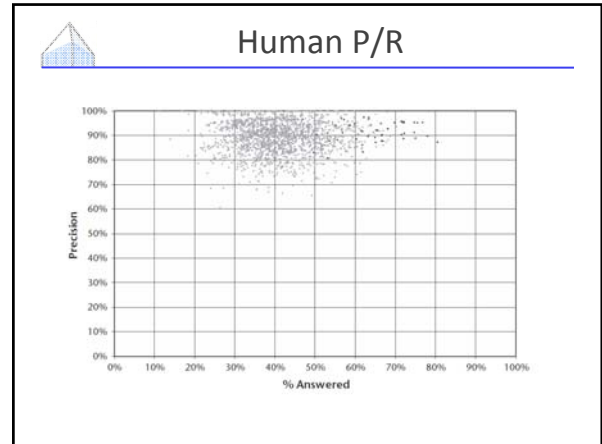
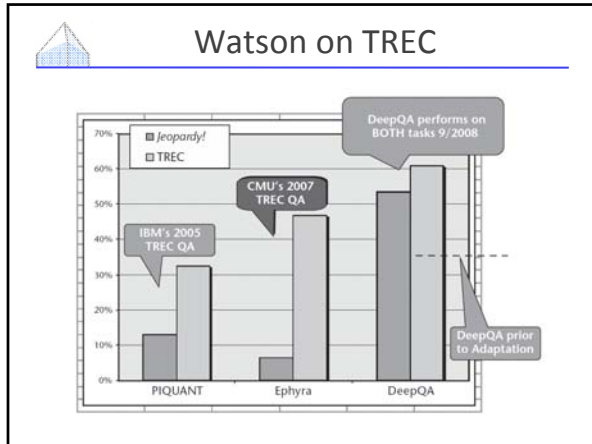
Category: "Rap" Sheet
Clue: This archaic term for a mischievous or annoying child can also mean a rogue or scamp.
Subclue 1: This archaic term for a mischievous or annoying child.
Subclue 2: This term can also mean a rogue or scamp.
Answer: Rapsallion

Category: Before and After Goes to the Movies
Clue: Film of a typical day in the life of the Beatles, which includes running from bloodthirsty zombie fans in a Romero classic.
Subclue 2: Film of a typical day in the life of the Beatles.
Answer 1: (A Hard Day's Night)
Subclue 2: Running from bloodthirsty zombie fans in a Romero classic.
Answer 2: (Night of the Living Dead)
Answer: A Hard Day's Night of the Living Dead



Architecture





Complex QA

Example of Complex Questions

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?