(TREC-style) Text-base Question Answering systems

Christopher Manning
CS224N/Ling 237 2005

(includes slides borrowed from Sanda Harabagiu, ISI, Nicholas Kushmerick, Jim Martin, Roxana Girju)

Question Answering from text

- An idea originating from the IR community
- With massive collections of full-text documents, simply finding *relevant documents* is of limited use: we want *answers* from textbases
- QA: give the user a (short) answer to their question, perhaps supported by evidence.
- The common person's view? [From a novel]
  - "I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota ... I'm the first guy to get the modem humming. But as a source of information, it sucks. You got a billion pieces of data, struggling to be heard and seen and downloaded, and anything I want to know seems to get trampled underfoot in the crowd."
    - M. Marshall. *The Straw Men*. HarperCollins Publishers, 2002.

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.)?

People *want* to ask questions...

Examples from AltaVista query log (late 1990s)
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?

Examples from Excite query log (12/1999)
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% of query logs
Seems to be a focus of new MSN Search Product

AskJeeves

- **AskJeeves** is probably most hyped example of "Question answering"
- It largely does 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
- If that fails, it falls back to regular web search
- A potentially interested middle ground, but a fairly weak shadow of real 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: e.g., LUNAR system
    - There's still Microsoft English Query
  - Spoken dialog systems
    - Currently very active and commercially relevant

A Brief (Academic) History

- Focusing on open-domain QA is new focus
  - MURAX (Kupiec 1993): Encyclopedia answers
  - Hirschman: Reading comprehension tests
  - TREC QA competition: 1999–
- But not really new either: Simmons et al. 1965
  - Take an encyclopedia and load it onto a computer.
  - Take a question and parse it into a logical form
  - Perform simple information retrieval to get relevant texts, parse those into a logical form, match and rank
    - What do worms eat? Worms eat ???
      - Candidates
        - Worms eat grass
        - Grass is eaten by worms
        - Birds eat worms

Online QA Examples

- Examples
  - **LCC:**
    http://www.languagecomputer.com/demos/question_answering/index.html
  - **AnswerBus** is an open-domain question answering system: www.answerbus.com
  - **Ionaut:** http://www.ionaut.com:8400/
  - **EasyAsk, AnswerLogic, AnswerFriend, Start, Quasm, Mulder, Webclopedia, ISI TextMap, etc.**

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

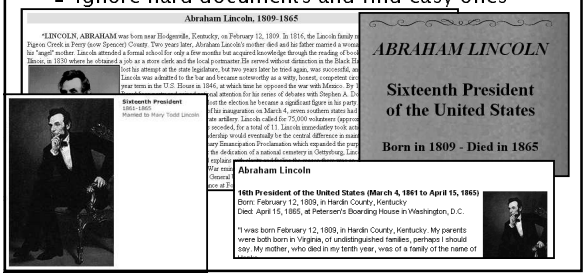
- The current collection uses news articles from the following sources:
  - AP newswire, 1998-2000
  - New York Times newswire, 1998-2000
  - Xinhua News Agency newswire, 1996-2000
- In total there are 1,033,461 documents in the collection. 3GB of text.
- This is a lot of text to process entirely using advanced NLP techniques so the systems usually consist of an initial information retrieval phase followed by more advanced processing.
- Many supplement this text with use of the web, and other knowledge bases

Top Performing Systems

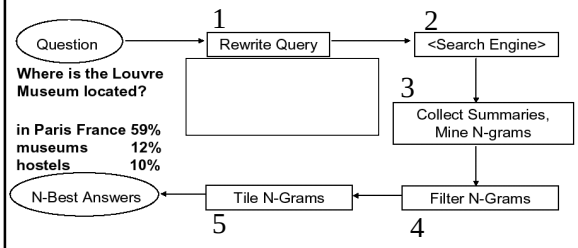
- Currently the best performing systems at TREC can answer approximately 70% of the questions !!!
- Approaches and successes have varied a fair deal
  - Knowledge-rich approaches, using a vast array of NLP techniques stole the show in 2000, 2001
    - Notably Harabagiu, Moldovan et al. – SMU/UTD/LCC
  - 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)

AskMSR: Shallow approach

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



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

- 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"

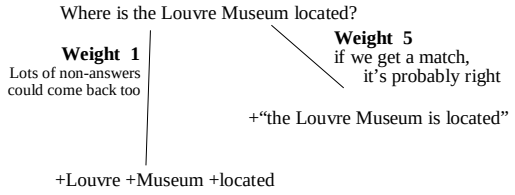
Nonsense, but who cares? It's only a few more queries to Google.

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?)

Query Rewriting – weights

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



Step 2: Query search engine

- Send all rewrites to a Web 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

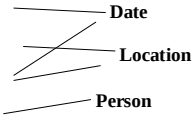
- Unigram, bigram, trigram, … N-gram: list of N adjacent terms in a sequence
- Eg, "Web Question Answering: Is More Always Better"
  - Unigrams: Web, Question, Answering, Is, More, Always, Better
  - Bigrams: Web Question, Question Answering, Answering Is, Is More, More Always, Always Better
  - Trigrams: Web Question Answering, Question Answering Is, Answering Is More, Is More Always, More Always Betters

Mining N-Grams

- Simple: Enumerate all N-grams (N=1,2,3 say) in all retrieved snippets
  - Use hash table and other fancy footwork to make this efficient
- 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

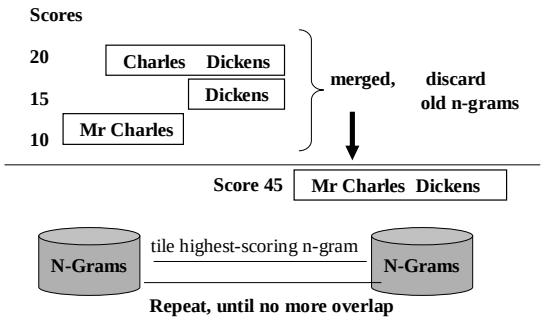
Step 4: Filtering N-Grams

- Each question type is associated with one or more "**data-type filters**" = regular expression
- When…
- Where…
- What …
- Who …



- 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



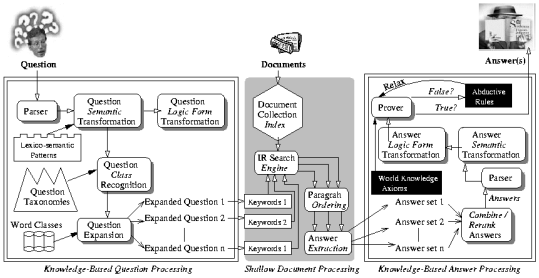
Results

- 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 enormity 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)

Limitations

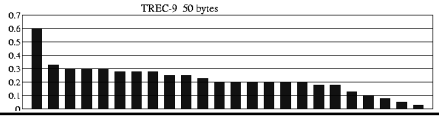
- In many scenarios (e.g., monitoring an individuals email...) we only have a small set of documents
- Works best/only for "Trivial Pursuit"-style fact-based questions
- Limited/brittle repertoire of
  - question categories
  - answer data types/filters
  - query rewriting rules

Full NLP QA
LCC: Harabagiu, Moldovan et al.

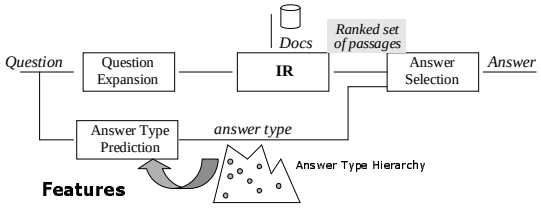


Value from sophisticated NLP - Pasca and Harabagiu (2001)

- 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



Answer types in State-of-the-art QA systems



Features

- Answer type
  - Labels questions with answer type based on a taxonomy
  - Classifies questions (e.g. by using a maximum entropy model)

Answer Types

- Of course, determining the answer type isn't that easy...
  - Who questions can have organizations as answers
    - Who sells the most hybrid cars?
  - Which questions can have people as answers
    - Which president went to war with Mexico?

QA Typology (from ISI USC)

- Typology of typical Q forms—94 nodes (47 leaf nodes)
- Analyzed 17,384 questions (from answers.com)

```
(THING
 ((AGENT
   (NAME (FEMALE-FIRST-NAME (EVE MARY ...))
         (MALE-FIRST-NAME (LAWRENCE SAM ...))))
         (COMPANY-NAME (BOEING AMERICAN-EXPRESS))
         JESUS ROMANOFF ...)
   (ANIMAL-HUMAN (ANIMAL (WOODCHUCK YAK ...))
                 PERSON)
   (ORGANIZATION (SQUADRON DICTATORSHIP ...))
   (GROUP-OF-PEOPLE (POSSE CHOIR ...))
   (STATE-DISTRICT (TIROL MISSISSIPPI ...))
   (CITY (ULAN-BATOR VIENNA ...))
   (COUNTRY (SULTANATE ZIMBABWE ...))))
 (PLACE
   (STATE-DISTRICT (CITY COUNTRY...))
   (GEOLOGICAL-FORMATION (STAR CANYON...))
   AIRPORT COLLEGE CAPITOL ...)
 (ABSTRACT
  (LANGUAGE (LETTER-CHARACTER (A B ...)))
  (QUANTITY
   (NUMERICAL-QUANTITY INFORMATION-QUANTITY
   MASS-QUANTITY MONETARY-QUANTITY
   TEMPORAL-QUANTITY ENERGY-QUANTITY
   TEMPERATURE-QUANTITY ILLUMINATION-QUANTITY
   (SPATIAL-QUANTITY
    (VOLUME-QUANTITY AREA-QUANTITY DISTANCE-QUANTITY)) ...
    PERCENTAGE)))
 (UNIT
  ((INFORMATION-UNIT (BIT BYTE ... EXABYTE))
   (MASS-UNIT (OUNCE ...)) (ENERGY-UNIT (BTU ...))
   (CURRENCY-UNIT (ZLOTY PESO ...))
   (TEMPORAL-UNIT (ATTOSECOND ... MILLENIUM))
   (TEMPERATURE-UNIT (FAHRENHEIT KELVIN CELCIUS))
   (ILLUMINATION-UNIT (LUX CANDELA))
   (SPATIAL-UNIT
    ((VOLUME-UNIT (DECILITER ...))
     (DISTANCE-UNIT (NANOMETER ...))))
     (AREA-UNIT (ACRE)) ... PERCENT))
 (TANGIBLE-OBJECT
  ((FOOD (HUMAN-FOOD (FISH CHEESE ...)))
   (SUBSTANCE
    ((LIQUID (LEMONADE GASOLINE BLOOD ...))
     (SOLID-SUBSTANCE (MARBLE PAPER ...))
     (GAS-FORM-SUBSTANCE (GAS AIR)) ...))
   (INSTRUMENT (DRUM DRILL (WEAPON (ARM GUN)) ...)
   (BODY-PART (ARM HEART ...))
   (MUSICAL-INSTRUMENT (PIANO)))
   ... *GARMENT *PLANT DISEASE)
```

Lexical Terms Extraction as input to Information Retrieval

- Questions approximated by sets of unrelated words (lexical terms)
- Similar to bag-of-word IR models: but choose nominal non-stop words and verbs

Question (from TREC QA track)	Lexical terms
Q002: What was the monetary value of the Nobel Peace Prize in 1989?	monetary, value, Nobel, Peace, Prize
Q003: What does the Peugeot company manufacture?	Peugeot, company, manufacture
Q004: How much did Mercury spend on advertising in 1993?	Mercury, spend, advertising, 1993

Keyword Selection Algorithm

1. Select all non-stopwords 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 adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select the answer type word

Passage Extraction Loop

- Passage Extraction Component
  - Extracts passages that contain all selected keywords
  - Passage size dynamic
  - Start position dynamic
- Passage quality and keyword adjustment
  - In the first iteration use the first 6 keyword selection heuristics
  - If the number of passages is lower than a threshold ⇒ query is too strict ⇒ drop a keyword
  - If the number of passages is higher than a threshold ⇒ query is too relaxed ⇒ add a keyword

Passage Scoring

- Passage ordering is performed using a sort that involves three scores:
  - The number of words from the question that are recognized in the same sequence in the window
  - The number of words that separate the most distant keywords in the window
  - The number of unmatched keywords in the window

Rank candidate answers in retrieved passages

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

Extracting Answers for Factoid Questions: NER!

- In TREC 2003 the LCC QA system extracted 289 correct answers for factoid questions
- The Name Entity Recognizer was responsible for 234 of them

QUANTITY	55	ORGANIZATION	15	PRICE	3
NUMBER	45	AUTHORED WORK	11	SCIENCE NAME	2
DATE	35	PRODUCT	11	ACRONYM	1
PERSON	31	CONTINENT	5	ADDRESS	1
COUNTRY	21	PROVINCE	5	ALPHABET	1
OTHER LOCATIONS	19	QUOTE	5	URL	1
CITY	19	UNIVERSITY	3		

Special Case of Names

Questions asking for names of authored works

1934: What is the play "West Side Story" based on? Answer: Romeo and Juliet
1976: What is the motto for the Boy Scouts? Answer: Be Prepared
1982: What movie won the Academy Award for best picture in 1989? Answer: Driving Miss Daisy
2080: What peace treaty ended WWI? Answer: Versailles
2102: What American landmark stands on Liberty Island? Answer: Statue of Liberty

NE-driven QA

- The results of the past 5 TREC evaluations of QA systems indicate that current state-of-the-art QA is determined by the recognition of Named Entities:
  - *Precision of recognition*
  - *Coverage of name classes*
  - *Mapping into concept hierarchies*
  - *Participation into semantic relations (e.g. predicate-argument structures or frame semantics)*

Concept Taxonomies

- For 29% of questions the LCC QA system relied on an off-line taxonomy with semantic classes such as:
  - Disease
  - Drugs
  - Colors
  - Insects
  - Games
- ***The majority of these semantic classes are also associated with patterns that enable their identification***

Semantics and Reasoning for QA: Predicate-argument structure

- **Q336:** *When was Microsoft established?*
- **This question is difficult because Microsoft tends to establish lots of things…**

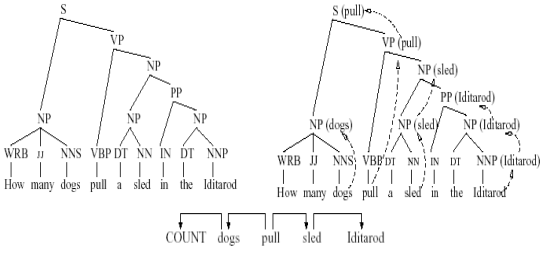
  Microsoft plans to establish manufacturing partnerships in Brazil and Mexico in May.

- **Need to be able to detect sentences in which `Microsoft' is object of `establish' or close synonym.**
- **Matching sentence:**

  Microsoft Corp was founded in the US in 1975, incorporated in 1981, and established in the UK in 1982.

- Requires analysis of sentence syntax/semantics!

Semantics and Reasoning for QA: Syntax to Logical Forms



- Syntactic analysis plus semantic –> logical form
- Mapping of question and potential answer LFs to find the best match

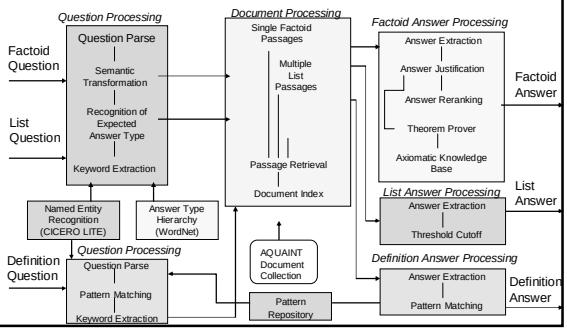
Abductive inference

- 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 very 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?
- get(TEMPERATURE, inside(volcano(active)))
- "lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit"
- fragments(lava, TEMPERATURE(degrees(300)), belched(out, mountain))
  - 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'

The Architecture of LCC's QA System circa 2003



Not all problems are solved yet!

- Where do lobsters like to live?
  - on a Canadian airline
- Where are zebras most likely found?
  - near dumps
  - in the dictionary
- Why can't ostriches fly?
  - Because of American economic sanctions
- What's the population of Mexico?
  - Three
- What can trigger an allergic reaction?
  - ..something that can *trigger* an allergic reaction

References

- R. F. Simmons, Natural language question-answering systems: 1969. Communications of the ACM. Volume 13, 1970.
- **AskMSR: Question Answering Using the Worldwide Web**
  - Michele Banko, Eric Brill, Susan Dumais, Jimmy Lin
  - http://www.ai.mit.edu/people/jimmylin/publications/Banko-etal-AAAI02.pdf
  - In Proceedings of 2002 AAAI SYMPOSIUM on Mining Answers from Text and Knowledge Bases, March 2002
- **Web Question Answering: Is More Always Better?**
  - Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, Andrew Ng
  - http://research.microsoft.com/~sdumais/SIGIR2002-QA-Submit-Conf.pdf
- D. Ravichandran and E.H. Hovy. 2002. **Learning Surface Patterns for a Question Answering System**. ACL conference, July 2002.

References

S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Gîrju, V.Rus and P. Morărescu. *FALCON: Boosting Knowledge for Answer Engines*. The Ninth Text REtrieval Conference (TREC 9), 2000.

Marius Pasca and Sanda Harabagiu, High Performance Question/Answering, in *Proceedings of the 24th Annual International ACL SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2001)*, September 2001, New Orleans LA, pages 366-374.

L. Hirschman, M. Light, E. Breck and J. Burger, *Deep Read: A Reading Comprehension System*. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.

C. Kwok, O. Etzioni and D. Weld. *Scaling Question Answering to the Web*. ACM Transactions in Information Systems, Vol 19, No. 3, July 2001, pages 242-262.

M. Light, G. Mann, E. Riloff and E. Breck, *Analyses for Elucidating Current Question Answering Technology*. Journal of Natural Language Engineering, Vol. 7, No. 4 (2001).

M. M. Soubbotin. *Patterns of Potential Answer Expressions as Clues to the Right Answers*. Proceedings of the Tenth Text REtrieval Conference (TREC 2001).