Open Domain Question Answering: Techniques, Resources and Systems

Adapted from a tutorial by Bernardo Magnini, ITC-Ist, Trento, Italy, 2003
Outline of the Tutorial

I. Introduction to QA
II. QA at TREC
III. System Architecture
    - Question Processing
    - Answer Extraction
IV. Answer Validation on the Web
I. Introduction to Question Answering

- What is Question Answering
- Applications
- Users
- Question Types
- Answer Types
- Evaluation
- Presentation
- Brief history
Query Driven vs Answer Driven Information Access

- What does LASER stand for?
- When did Hitler attack Soviet Union?
  - Using Google we find documents containing the question itself, no matter whether or not the answer is actually provided.
- Current information access is query driven.
- Question Answering proposes an answer driven approach to information access.
Question Answering

- Find the answer to a question in a large collection of documents
  - questions (in place of keyword-based query)
  - answers (in place of documents)
Alternatives to Information Retrieval

- **Document Retrieval**
  - users submit queries corresponding to their information need
  - system returns (voluminous) list of full-length documents
  - it is the responsibility of the users to find their original information need, within the returned documents

- **Open-Domain Question Answering (QA)**
  - users ask fact-based, natural language questions
    - What is the highest volcano in Europe?
  - system returns list of short answers
    - Under Mount Etna, the highest volcano in Europe, perches the fabulous town ...
  - more appropriate for specific information needs
What is QA?

- Find the answer to a question in a large collection of documents

**What is the brightest star visible from Earth?**

1. *Sirio A* is the **brightest star visible from Earth** even if it is...

2. the planet is 12-times brighter than *Sirio, the brightest star* in the sky...
Problem: discovery implicit relations among question and answers

Who is the author of the "Star Spangled Banner"?

...Francis Scott Key wrote the "Star Spangled Banner" in 1814.

...comedian-actress Roseanne Barr sang her famous rendition of the "Star Spangled Banner" before ...
Problem: discovery implicit relations among question and answers

Which is the Mozart *birth date*?

.... Mozart *(1751 – 1791) ....*
Problem: discovery implicit relations among question and answers

**Which is the distance between Naples and Ravello?**

"From the **Naples** Airport follow the sign to Autostrade (green road sign). Follow the directions to Salerno (A3). Drive for about 6 Km. Pay toll (Euros 1.20). Drive appx. 25 Km. Leave the Autostrade at Angri (Uscita Angri). Turn left, follow the sign to Ravello through Angri. Drive for about 2 Km. Turn right following the road sign "Costiera Amalfitana". Within 100m you come to traffic lights prior to narrow bridge. Watch not to miss the next Ravello sign, at appx. 1 Km from the traffic lights. Now relax and enjoy the views (follow this road for 22 Km). Once in **Ravello** ..."
QA: Applications (1)

Information access:

- Structured data (databases)
- Semi-structured data (e.g. comment field in databases, XML)
- Free text

To search over:

- The Web
- Fixed set of text collection (e.g. TREC)
- A single text (reading comprehension evaluation)
QA: Applications (2)

- Domain independent QA
- Domain specific (e.g. help systems)
- Multi-modal QA
  - Annotated images
  - Speech data
QA: Users

- Casual users, first time users
  - Understand the limitations of the system
  - Interpretation of the answer returned

- Expert users
  - Difference between novel and already provided information
  - User Model
QA: Questions (1)

Classification according to the **answer type**
- **Factual questions** (*What is the larger city...*)
- **Opinions** (*What is the author’s attitude...*)
- **Summaries** (*What are the arguments for and against...*)

Classification according to the **question speech act**:
- **Yes/NO questions** (*Is it true that...*)
- **WH questions** (*Who was the first president...*)
- **Indirect Requests** (*I would like you to list...*)
- **Commands** (*Name all the presidents...*)
Difficult questions

- **Why, How questions** require understanding causality or instrumental relations

- **What questions** have little constraint on the answer type (e.g. *What did they do?*)
QA: Answers

- Long answers, with justification
- Short answers (e.g. phrases)
- Exact answers (named entities)
- Answer construction:
  - Extraction: cut and paste of snippets from the original document(s)
  - Generation: from multiple sentences or documents
  - QA and summarization (e.g. What is this story about?)
QA: Information Presentation

- Interfaces for QA
  - Not just isolated questions, but a dialogue
  - Usability and user satisfaction
- Critical situations
  - Real time, single answer
- Dialog-based interaction
  - Speech input
  - Conversational access to the Web
QA: Brief History (1)

- **NLP interfaces to databases:**
  - Limitations: structured knowledge and limited domain

- **Story comprehension:** Shank (1977), Kintsch (1998), Hirschman (1999)
QA: Brief History (2)

- **Information retrieval (IR)**
  - Queries are questions
  - List of documents are answers
  - QA is close to passage retrieval
  - Well established methodologies (i.e. Text Retrieval Conferences TREC)

- **Information extraction (IE):**
  - Pre-defined templates are questions
  - Filled template are answers
Growing interest in QA (TREC and CLEF evaluation campaign).

Recent focus on **multilinguality** and **context aware QA**
Research Context (2)

Automatic Summarization

Compactness

Automatic Question Answering

Answers must be faithful w.r.t. questions (correctness) and compact (exactness)

Machine Translation

Faithfulness

As compact as possible

As faithful as possible
II. Question Answering at TREC

- The problem simplified
- Questions and answers
- Evaluation metrics
- Approaches
The problem simplified: The Text Retrieval Conference

- **Goal**
  - Encourage research in information retrieval based on large-scale collections

- **Sponsors**
  - NIST: National Institute of Standards and Technology
  - ARDA: Advanced Research and Development Activity
  - DARPA: Defense Advanced Research Projects Agency

- Since 1999

- Participants are research institutes, universities, industries
TREC Questions

Q-1391: How many feet in a mile?
Q-1057: Where is the volcano Mauna Loa?
Q-1071: When was the first stamp issued?
Q-1079: Who is the Prime Minister of Canada?
Q-1268: Name a food high in zinc.

Q-896: Who was Galileo?
Q-897: What is an atom?

Q-711: What tourist attractions are there in Reims?
Q-712: What do most tourists visit in Reims?
Q-713: What attracts tourists in Reims?
Q-714: What are tourist attractions in Reims?

Fact-based, short answer questions
Definition questions
Reformulation questions
Criteria for judging an answer

- **Relevance**: it should be responsive to the question
- **Correctness**: it should be factually correct
- **Conciseness**: it should not contain extraneous or irrelevant information
- **Completeness**: it should be complete, i.e. partial answer should not get full credit
- **Simplicity**: it should be simple, so that the questioner can read it easily
- **Justification**: it should be supplied with sufficient context to allow a reader to determine why this was chosen as an answer to the question
Exact Answers

- Basic unit of a response: [answer-string, docid] pair
- An answer string must contain a complete, exact answer and nothing else.

What is the longest river in the United States?
The following are **correct, exact answers:**

Mississippi,
the Mississippi,
the Mississippi River,
Mississippi River
mississippi

while none of the following are correct exact answers:
At 2,348 miles the Mississippi River is the longest river in the US.
2,348 miles; Mississippi
Missipp
Assessments

- Four possible judgments for a triple
  [ Question, document, answer ]

- **Rigth**: the answer is appropriate for the question
- **Inexact**: used for non complete answers
- **Unsupported**: answers without justification
- **Wrong**: the answer is not appropriate for the question
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the capital city of New Zealand?</td>
<td>R 1530 XIE19990325.0298 Wellington</td>
</tr>
<tr>
<td>What is the Boston Strangler's name?</td>
<td>R 1490 NYT20000913.0267 Albert DeSalvo</td>
</tr>
<tr>
<td>What is the world's second largest island?</td>
<td>R 1503 XIE19991018.0249 New Guinea</td>
</tr>
<tr>
<td>What year did Wilt Chamberlain score 100 points?</td>
<td>U 1402 NYT19981017.0283 1962</td>
</tr>
<tr>
<td>Who is the governor of Tennessee?</td>
<td>R 1426 NYT19981030.0149 Sundquist</td>
</tr>
<tr>
<td>What's the name of King Arthur's sword?</td>
<td>U 1506 NYT19980618.0245 Excalibur</td>
</tr>
<tr>
<td>When did Einstein die?</td>
<td>R 1601 NYT19990315.0374 April 18, 1955</td>
</tr>
<tr>
<td>What was the name of the plane that dropped the Atomic Bomb on Hiroshima?</td>
<td>X 1848 NYT19991001.0143 Enola</td>
</tr>
<tr>
<td>What was the name of FDR's dog?</td>
<td>R 1838 NYT20000412.0164 Fala</td>
</tr>
<tr>
<td>What day did Neil Armstrong land on the moon?</td>
<td>R 1674 APW19990717.0042 July 20, 1969</td>
</tr>
<tr>
<td>Who was the first Triple Crown Winner?</td>
<td>X 1716 NYT19980605.0423 Barton</td>
</tr>
<tr>
<td>When was Lyndon B. Johnson born?</td>
<td>R 1473 APW19990826.0055 1908</td>
</tr>
<tr>
<td>Who was Woodrow Wilson's First Lady?</td>
<td>R 1622 NYT19980903.0086 Ellen</td>
</tr>
<tr>
<td>Where is Anne Frank's diary?</td>
<td>W 1510 NYT19980909.0338 Young Girl</td>
</tr>
</tbody>
</table>
1402: What year did Wilt Chamberlain score 100 points?

DIOGENE: 1962

ASSESSMENT: UNSUPPORTED

PARAGRAPH: NYT19981017.0283

Petty's 200 victories, 172 of which came during a 13-year span between 1962-75, may be as unapproachable as Joe DiMaggio's 56-game hitting streak or Wilt Chamberlain's 100-point game.
'QUEST FOR CAMELOT,' with the voices of Andrea Carr, Gabriel Byrne, Cary Elwes, John Gielgud, Jessalyn Gilsig, Eric Idle, Gary Oldman, Bronson Pinchot, Don Rickles and Bryan White. Directed by Frederik Du Chau (G, 100 minutes). Warner Brothers' shaky entrance into the Disney-dominated sweepstakes of the musicalized animated feature wants to be a juvenile feminist "Lion King" with a musical heart that fuses "Riverdance" with formulaic Hollywood gush. But its characters are too wishy-washy and visually unfocused to be compelling, and the songs (by David Foster and Carole Bayer Sager) so forgettable as to be extraneous. In this variation on the Arthurian legend, a nondescript Celtic farm girl named Kayley with aspirations to be a knight wrests the magic sword Excalibur from the evil would-be emperor Ruber (a Hulk Hogan look-alike) and saves the kingdom (Holden).
1848: What was the name of the plane that dropped the Atomic Bomb on Hiroshima?

DIOGENE: Enola
PARAGRAPH: NYT19991001.0143

ASSESSMENT: INEXACT

Tibbets piloted the Boeing B-29 Superfortress Enola Gay, which dropped the atomic bomb on Hiroshima on Aug. 6, 1945, causing an estimated 66,000 to 240,000 deaths. He named the plane after his mother, Enola Gay Tibbets.
1716: Who was the first Triple Crown Winner?

DIOGENE: Barton
PARAGRAPH: NYT19980605.0423

ASSESSMENT: INEXACT

Not all of the Triple Crown winners were immortals. The first, Sir Barton, lost six races in 1918 before his first victory, just as Real Quiet lost six in a row last year. Try to find Omaha and Whirlaway on anybody's list of all-time greats.
Where is Anne Frank's diary?

DIOGENE: Young Girl
PARAGRAPH: NYT19980909.0338

ASSESSMENT: WRONG

Otto Frank released a heavily edited version of “B” for its first publication as “Anne Frank: Diary of a Young Girl” in 1947.
TREC Evaluation Metric: Mean Reciprocal Rank (MRR)

- **Reciprocal Rank** = inverse of rank at which first correct answer was found (5 answers allowed per question):
  
  \[ [1, 0.5, 0.33, 0.25, 0.2, 0] \]

- **MRR**: average over all questions
- **Strict score**: unsupported count as incorrect
- **Lenient score**: unsupported count as correct
TREC Evaluation Metrics: Confidence-Weighted Score (CWS)
(if only one answer per question, in order of confidence)

Sum for $i = 1$ to $\text{No\_questions}$ (#-correct-up-to-question $i$ / $i$)

System A:
1 $\rightarrow$ C
2 $\rightarrow$ W
3 $\rightarrow$ C
4 $\rightarrow$ C
5 $\rightarrow$ W

$\frac{1}{1} + \frac{1+0}{2} + \frac{1+0+1}{3} + \frac{(1+0+1+1)}{4} + \frac{(1+0+1+1+0)}{5}$

Total: 0.7

System B:
1 $\rightarrow$ W
2 $\rightarrow$ W
3 $\rightarrow$ C
4 $\rightarrow$ C
5 $\rightarrow$ C

$\frac{0 + (0+0)}{2} + \frac{0+0+1}{3} + \frac{(0+0+1+1)}{4} + \frac{(0+0+1+1+1)}{5}$

Total: 0.29
Main Approaches at TREC

- Knowledge-Based
- Web-based
- Pattern-based
Knowledge-Based Approach

**Linguistic-oriented methodology**

- Determine the answer type from question form
- Retrieve small portions of documents
- Find entities matching the answer type category in text snippets

- Majority of systems use a lexicon (usually WordNet)
  - To find answer type
  - To verify that a candidate answer is of the correct type
  - To get definitions

- Complex architecture...
Web-Based Approach

- Question Processing Component
- Search Component
- Answer Extraction Component

WEB

Auxiliary Corpus
TREC Corpus

QUESTION

ANSWER
Pattern-Based Approach (1/3)

- **Knowledge poor**
- **Strategy**
  - Search for predefined patterns of textual expressions that may be interpreted as answers to certain question types.
  - The presence of such patterns in answer string candidates may provide evidence of the right answer.
Pattern-Based Approach (2/3)

Conditions

- Detailed categorization of question types
  - Up to 9 types of the “Who” question; 35 categories in total
- Significant number of patterns corresponding to each question type
  - Up to 23 patterns for the “Who-Author” type, average of 15
- Find multiple candidate snippets and check for the presence of patterns (emphasis on recall)
Pattern-based approach (3/3)

- Example: patterns for definition questions
- Question: What is A?
  1. <A; is/are; [a/an/the]; X> ...23 correct answers
  2. <A; comma; [a/an/the]; X; [comma/period]> ...26 correct answers
  3. <A; [comma]; or; X; [comma]> ...12 correct answers
  4. <A; dash; X; [dash]> ...9 correct answers
  5. <A; parenthesis; X; parenthesis> ...8 correct answers
  6. <A; comma; [also] called; X [comma]> ...7 correct answers
  7. <A; is called; X> ...3 correct answers

**total: 88 correct answers**
Use of answer patterns

1. For generating queries to the search engine.
   How did Mahatma Gandhi die?
   - Mahatma Gandhi die <HOW>
   - Mahatma Gandhi die of <HOW>
   - Mahatma Gandhi lost his life in <WHAT>

The TEXTMAP system (ISI) uses 550 patterns, grouped in 105 equivalence blocks. On TREC-2003 questions, the system produced, on average, 5 reformulations for each question.

2. For answer extraction
   When was Mozart born?
   - P=1 <PERSON> (<BIRTHDATE> - DATE)
   - P=.69 <PERSON> was born on <BIRTHDATE>
Acquisition of Answer Patterns

Relevant approaches:
- Manually developed surface pattern library (Soubbotin, Soubbotin, 2001)
- Automatically extracted surface patterns (Ravichandran, Hovy 2002)

Pattern learning:
1. Start with a seed, e.g. (Mozart, 1756)
2. Download Web documents using a search engine
3. Retain sentences that contain both question and answer terms
4. Construct a suffix tree for extracting the longest matching substring that spans <Question> and <Answer>
5. Calculate precision of patterns
   Precision = # of correct patterns with correct answer / # of total patterns
Capturing variability with patterns

- Pattern based QA is more effective when supported by variable typing obtained using NLP techniques and resources.
  
  When was <A> born?
  
  <A:PERSON> (<ANSWER:DATE> - <A:PERSON> was born in <ANSWER:DATE>

- Surface patterns can not deal with word reordering and apposition phrases:
  
  *Galileo, the famous astronomer, was born in* ...

- The fact that most of the QA systems use syntactic parsing demonstrates that the successful solution of the answer extraction problem goes beyond the surface form analysis
Answer patterns that capture the syntactic relations of a sentence.

When was <A> invented?
The matching phase turns out to be a problem of partial match among syntactic trees.

Syntactic answer patterns (2)

The first phonograph was invented in 1877.
III. System Architecture

- **Knowledge Based** approach
  - Question Processing
  - Search component
  - Answer Extraction
Knowledge based QA

Question Processing Component
- Tokenization & POS Tagging
  - Multiwords Recognition
    - Question Parsing
      - Word Sense Disambiguation
        - Answer Type Identification
          - Keywords Expansion

Search Component
- Document Collection
  - Search Engine
    - Query Composition

Answer Extraction Component
- Answer
  - Answer Validation
    - Answer Identification
      - Named Entities Recognition
        - Paragraph Filtering

Component flows from Question Processing to Search to Answer Extraction.
Question Analysis (1)

- **Input:** NLP question
- **Output:**
  - query for the search engine (i.e. a boolean composition of weighted keywords)
  - Answer type
  - Additional constraints: question focus, syntactic or semantic relations that should hold for a candidate answer entity and other entities
Question Analysis (2)

Steps:

1. Tokenization
2. POS-tagging
3. Multi-words recognition
4. Parsing
5. Answer type and focus identification
6. Keyword extraction
7. Word Sense Disambiguation
8. Expansions
Who was the inventor of the electric light?
Who was the inventor of the electric light?
Syntactic Parsing

- Identify syntactic structure of a sentence
  - noun phrases (NP), verb phrases (VP), prepositional phrases (PP) etc.

Why did David Koresh ask the FBI for a word processor?
Answer Type and Focus

- **Focus** is the word that expresses the relevant entity in the question
  - Used to select a set of relevant documents
  - ES: Where was Mozart born?

- **Answer Type** is the category of the entity to be searched as answer
  - PERSON, MEASURE, TIME PERIOD, DATE, ORGANIZATION, DEFINITION
  - ES: Where was Mozart born?
    - LOCATION
What famous communist leader died in Mexico City?

RULENAME: WHAT-WHO

TEST: [“what” [¬ NOUN]* [NOUN:person-p]j +]

OUTPUT: [“PERSON” J]

Answer type: PERSON

Focus: leader

This rule matches any question starting with what, whose first noun, if any, is a person (i.e. satisfies the person-p predicate)
Who was the inventor of the electric light?
**Word Sense Disambiguation**

*What is the brightest *star* visible from Earth?*

<table>
<thead>
<tr>
<th>Word</th>
<th>Definition 1</th>
<th>Definition 2</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAR</td>
<td>star#1: celestial body</td>
<td>star#2: an actor who plays ...</td>
<td>ASTRONOMY</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ART</td>
</tr>
<tr>
<td>BRIGHT</td>
<td>bright #1: bright, brilliant, shining</td>
<td></td>
<td>PHYSICS</td>
</tr>
<tr>
<td></td>
<td>bright #2: popular, glorious</td>
<td></td>
<td>GENERIC</td>
</tr>
<tr>
<td></td>
<td>bright #3: promising, auspicious</td>
<td></td>
<td>GENERIC</td>
</tr>
<tr>
<td>VISIBLE</td>
<td>visible#1: conspicuous, obvious</td>
<td></td>
<td>PHYSICS</td>
</tr>
<tr>
<td></td>
<td>visible#2: visible, seeable</td>
<td></td>
<td>ASTRONOMY</td>
</tr>
<tr>
<td>EARTH</td>
<td>earth#1: Earth, world, globe</td>
<td></td>
<td>ASTRONOMY</td>
</tr>
<tr>
<td></td>
<td>earth#2: estate, land, landed_estate, acres</td>
<td></td>
<td>ECONOMY</td>
</tr>
<tr>
<td></td>
<td>earth#3: clay</td>
<td></td>
<td>GEOLOGY</td>
</tr>
<tr>
<td></td>
<td>earth#4: dry_land, earth, solid, ground</td>
<td></td>
<td>GEOGRAPHY</td>
</tr>
<tr>
<td></td>
<td>earth#5: land, ground, soil</td>
<td></td>
<td>GEOGRAPHY</td>
</tr>
<tr>
<td></td>
<td>earth#6: earth, ground</td>
<td></td>
<td>GEOLOGY</td>
</tr>
</tbody>
</table>
Keyword Composition

- Keywords and expansions are composed in a boolean expression with AND/OR operators

- Several possibilities:
  - AND composition
    - (OR (inventor AND electric_light)
      OR (inventor AND incandescent_lamp)
      OR (discoverer AND electric_light)
      .........................
      OR inventor OR electric_light))
For real time QA applications off-line pre-processing of the text is necessary

- Term indexing
- POS-tagging
- Named Entities Recognition
Candidate Answer Document Selection

- **Passage Selection**: Individuate relevant, small, text portions

- Given a document and a list of keywords:
  - Paragraph length (e.g. 200 words)
  - Consider the percentage of keywords present in the passage
  - Consider if some keyword is obligatory (e.g. the focus of the question).
Candidate Answer Document Analysis

- Passage text tagging
- **Named Entity Recognition**

Who is the author of the “Star Spangled Banner”?

...<PERSON>Francis Scott Key</PERSON> wrote the “Star Spangled Banner” in <DATE>1814</DATE>

Some systems:
- passages parsing (Harabagiu, 2001)
- logical form (Zajac, 2001)
Who is the author of the "Star Spangled Banner"?

...<PERSON>Francis Scott Key</PERSON> wrote the "Star Spangled Banner" in <DATE>1814</DATE>

Answer Type = PERSON
Candidate Answer = Francis Scott Key

Ranking candidate answers: keyword density in the passage, apply additional constraints (e.g. syntax, semantics), rank candidates using the Web
IV. Answer Validation

- Automatic answer validation
- Approach:
  - web-based
  - use of patterns
  - combine statistics and linguistic information
- Discussion
- Conclusions
QA Architecture
The problem: Answer Validation

Given a question $q$ and a candidate answer $a$, decide if $a$ is a correct answer for $q$

What is the capital of the USA?

- Washington D.C.
- San Francisco
- Rome
The problem: Answer Validation

*Given a question* $q$ *and a candidate answer* $a$, *decide if* $a$ *is a correct answer for* $q$.

What is the capital of the USA?

- Washington D.C.  
  correct
- San Francisco  
  wrong
- Rome  
  wrong
Requirements for Automatic AV

- **Accuracy**: it has to compare well with respect to human judgments
- **Efficiency**: large scale (Web), real time scenarios
- **Simplicity**: avoid the complexity of QA systems
Approach

- **Web-based**
  - take advantage of Web redundancy

- **Pattern-based**
  - the Web is mined using patterns (i.e. *validation patterns*) extracted from the question and the candidate answer

- **Quantitative (as opposed to content-based)**
  - check if the question and the answer tend to appear together in the Web considering the number of documents returned (i.e. documents are not downloaded)
Web Redundancy

What is the capital of the USA? Washington

<table>
<thead>
<tr>
<th>Capital Region USA: Fly-Drive Holidays in and Around Washington D.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>the Insider’s Guide to the Capital Area Music Scene (Washington D.C., USA).</td>
</tr>
<tr>
<td>The Capital Tangueros (Washington DC Area, USA)</td>
</tr>
<tr>
<td>I live in the Nation’s Capital, Washington Metropolitan Area (USA)</td>
</tr>
<tr>
<td>In 1790 Capital (also USA’s capital): Washington D.C. Area: 179 square km</td>
</tr>
</tbody>
</table>
**Capital Region USA**: Fly-Drive Holidays in and Around Washington D.C.

- The Insider’s Guide to the Capital Area Music Scene (Washington D.C., USA).
- The **Capital Tangueros** (Washington DC Area, USA)
- I live in the Nation’s **Capital, Washington Metropolitan Area (USA)**
- In 1790 **Capital** (also USA’s capital): **Washington D.C. Area: 179 square km**

[**Capital NEAR USA NEAR Washington**]
Related Work

- **Pattern-based QA**
  - Brill, 2001 – TREC-10
  - Subboting, 2001 – TREC-10
  - Ravichandran and Hovy, ACL-02

- **Use of the Web for QA**
  - Clarke et al. 2001 – TREC-10
  - Radev, et al. 2001 - CIKM

- **Statistical approach on the Web**
  - PMI-IR: Turney, 2001 and ACL-02
Architecture

- question
- candidate answer

validation pattern

- filtering
- \#doc < k

answer validity score

- \#doc

- > t
- < t

- correct answer
- wrong answer
Extracting Validation Patterns

question → candidate answer

stop-word filter

term expansion

question pattern (Qp)

answer type

named entity recognition

stop-word filter

answer pattern (Ap)

validation pattern
Answer Validity Score

- **PMI-IR** algorithm (Turney, 2001)

\[
\text{PMI} (Q_p, A_p) = \frac{P(Q_p, A_p)}{P(Q_p) \times P(A_p)}
\]

- The result is interpreted as evidence that the validation pattern is consistent, which imply answer accuracy
Answer Validity Score

PMI \((Q_p, A_p)\) = \frac{\text{hits}(Q_p \text{ NEAR } A_p)}{\text{hits}(Q_p) \times \text{hits}(A_p)}

- Three searches are submitted to the Web:
  - \text{hits}(Q_p)
  - \text{hits}(A_p)
  - \text{hits}(Q_p \text{ NEAR } A_p)
Example

A1 = The Stanislaus County district attorney’s
A2 = In Modesto, San Francisco, and

What county is Modesto, California in?

Stop-word filter

Answer type: Location

$Q_p = \text{[county NEAR Modesto NEAR California]}$

$P(Q_p) = P(\text{county, Modesto, California}) = \frac{909}{3 \times 10^8}$
Example (cont.)

The Stanislaus County district attorney’s

In Modesto, San Francisco, and

NER(location)

\[ A1p = [\text{Stanislaus}] \]
\[ A2p = [\text{San Francisco}] \]

\[ P(\text{Stanislaus}) = \frac{73641}{3 \times 10^8} \]
\[ P(\text{San Francisco}) = \frac{4072519}{3 \times 10^8} \]
The **Stanislaus** County district attorney’s

In Modesto, **San Francisco**, and

\[
P(Q_p, A_{1p}) = \frac{552}{3 \times 10^8}
\]

\[
P(Q_p, A_{2p}) = \frac{11}{3 \times 10^8}
\]

\[
PMI(Q_p, A_{1p}) = 2473
\]

\[
PMI(Q_p, A_{2p}) = 0.89
\]

\[
t = 0.2 \times \text{MAX}(AVS)
\]

> t  \quad \text{correct answer}

< t  \quad \text{wrong answer}
Experiments

- **Data set:**
  - 492 TREC-2001 questions
  - 2726 answers: 3 correct answers and 3 wrong answers for each question, randomly selected from TREC-10 participants human-judged corpus

- **Search engine:** Altavista
  - used to allow the NEAR operator
Q 916: What river in the US is known as the Big Muddy?

- The Mississippi
- Known as Big Muddy, the Mississippi is the longest
- as Big Muddy, the Mississippi is the longest
- messed with. Known as Big Muddy, the Mississippi
- Mississippi is the longest river in the US
- the Mississippi is the longest river (Mississippi)
- has brought the Mississippi to its lowest
- ipes. In Life on the Mississippi, Mark Twain wrote t
- Southeast; Mississippi; Mark Twain; officials began
- Known; Mississippi; US; Minnesota; Gulf Mexico
- Mud Island,; Mississippi; ”The;--history,;Memphis
Baseline

- Consider the documents provided by NIST to TREC-10 participants (1000 documents for each question)
- If the candidate answer occurs (i.e. string match) at least one time in the top 10 documents it is judged correct, otherwise it is considered wrong
- Baseline (~58% correct answers), validation with PMI (~78% correct answers)
Discussion (1)

- Definition questions are the more problematic
  - on the subset of 249 named-entities questions success rate is higher (i.e. 86.3)
- Relative threshold improve performance (+ 2%) over fixed threshold
- Non symmetric measures of co-occurrence work better for answer validation (+ 2%)
- Source of errors:
  - Answer type recognition
  - Named-entities recognition
  - TREC answer set (e.g. tokenization)
Discussion (2)

- Automatic answer validation is a key challenge for Web-based question/answering systems

- Requirements:
  - accuracy with respect to human judgments: 80% success rate is a good starting point
  - efficiency: documents are not downloaded
  - simplicity: based on patterns

- At present, it is suitable for a generate&test component integrated in a QA system