

Question Answering

Lê Thanh Hương
Bộ môn Hệ thống Thông tin
Viện CNTT & TT – Trường ĐHBKHN
Email: huonglt@soict.hust.edu.vn

Question Answering

- An idea originating from the IR community
- IR: find *relevant documents*, but we want *answers* from textbases
- QA: give short answer, perhaps supported by evidence

Sample TREC questions

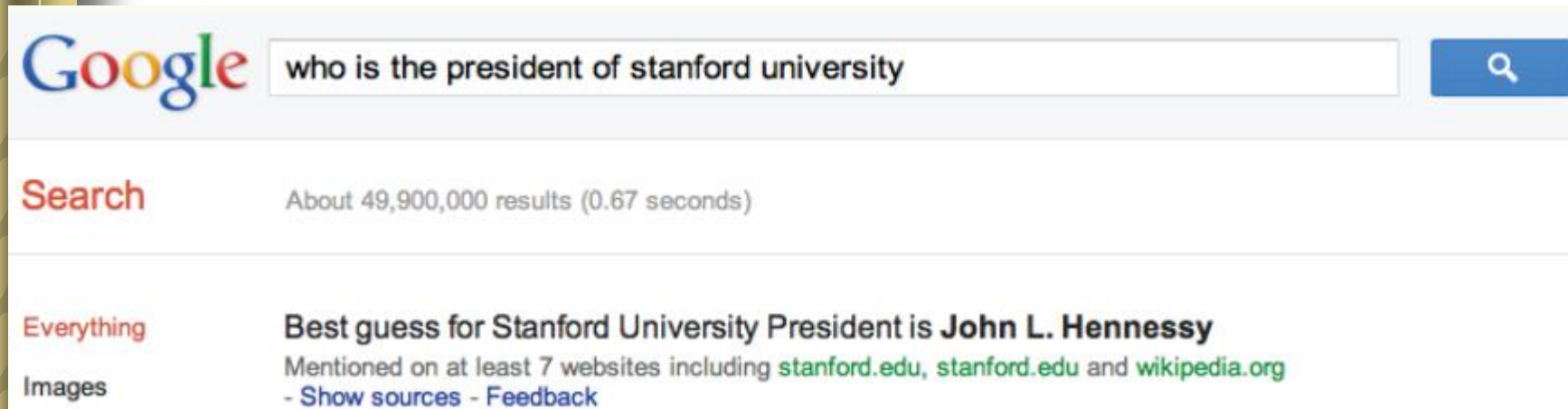
- Who is the author of the book “The Iron Lady: A Biography of Margaret Thatcher”?
- What was the monetary value of the Nobel Peace Prize in 1989?
- What does the Peugeot company manufacture?
- How much did Mercury spend on advertising in 1993?
- Why did David Koresh ask the FBI for a word processor?

People want to ask questions

- Examples from AltaVista query log (late 1990s)
 - Who invented surf music?
 - How to make stink bombs
 - Which english translation of the bible is used in official catholic liturgies?
- Examples from Excite query log (12/1999)
 - How can i find someone in Texas
 - Where can i find information on puritan religion?
 - What vacuum cleaner does Consumers Guide recommend

Online QA Examples

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



The image shows a screenshot of a Google search interface. The search bar contains the text "who is the president of stanford university". To the right of the search bar is a blue search button with a magnifying glass icon. Below the search bar, the word "Search" is displayed in red. To the right of "Search" is the text "About 49,900,000 results (0.67 seconds)". Below this, the word "Everything" is displayed in red. To the right of "Everything" is the text "Best guess for Stanford University President is **John L. Hennessy**". Below this, the text "Mentioned on at least 7 websites including stanford.edu, stanford.edu and wikipedia.org" is displayed. At the bottom, the text "- Show sources - Feedback" is displayed.

AskJeeves

- ...is most hyped example of QA
- ...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, return regular web search
- A potentially interested middle ground, but a weak shadow of real QA

The screenshot shows the AskJeeves website interface. At the top, the URL is `uk.ask.com/web?qsrc=1&o=0&l=dir&q=who+is+the+president+of+The+United+States++2012&dm=all`. The search bar contains the query "who is the president of The United States 2012". To the right of the search bar is a blue "Find Answers" button. Below the search bar, there are two columns of results. The left column is titled "Explore Answers About" and lists several links related to the United States, including "United States History Timeline", "United States Atlas", "United States Road Map", "United States Facts", "Visitors Visa Requirements United States America", and "A List of Presidents in Order". The right column is titled "Popular Q&A" and shows two questions and answers. The first question is "Q: Who the president of the united states 2012?" with the answer "A: President Barack Obama has won re-election and will serve 4 more years a..." and a source link to www.chacha.com. The second question is "Q: Who is vice president of united states 2012?" with the answer "A: The vice president of The United States of America in 2012 is Joe Biden." and a source link to wiki.answers.com. The top navigation bar includes "Answers", "Advanced Search", "Settings", and "Your Cookie Cho". The AskJeeves logo is visible in the top left corner.

Answers Advanced Search Settings Your Cookie Cho

Ask Jeeves The We UK Onl

Explore Answers About

- Everything ▶ **United States** History Timeline
- Images **United States** Atlas
- Video **United States** Road Map
- Reference **United States** Facts
- Q&A Visitors Visa Requirements **United States** America
- A List of **Presidents** in Order

Popular Q&A

Q: Who the president of the united states 2012?
A: President Barack Obama has won re-election and will serve 4 more years a... [Read More »](#)
Source: www.chacha.com

Q: Who is vice president of united states 2012?
A: The vice president of The United States of America in 2012 is Joe Biden. [Read More »](#)
Source: wiki.answers.com



TEXTMAP
THE ENTITY SEARCH ENGINE

Monitoring the World So You Don't Have To ...

ENTITIES

SOL

Search!

[TextMap](#) : [TextMed](#) : [Textblq](#) : [TextBiz](#) : [Make homepage!](#) : [Link to us](#) : [Help?](#)

Question Answering

Wednesda

in what year did John Lennon die?

Answer: 1980

[[The Beatles Anthology](#) 02/28/2006 [wiki](#)]



TEXTMAP

THE ENTITY SEARCH ENGINE

Monitoring the World So You Don't Have To ...



who is the Prime Minister of vietnam

Search!

TextMap All Sources

[TextMap](#) : [TextMed](#) : [TextBlg](#) : [TextBiz](#) : [Make homepage!](#) : [Link to us](#) : [Help?](#)

Search Results 1-25 of about 330,000

[Next >>](#)

Rank	Entity	Score	Type	Popularity	Top Month for Query
1	Vietnam	<div style="width: 100%; height: 10px; background-color: blue;"></div>	COUNTRY	<div style="width: 80%; height: 10px; background-color: blue; border: 1px solid orange;"></div>	November 2006
2	Iraq	<div style="width: 80%; height: 10px; background-color: blue; border: 1px solid orange;"></div>	COUNTRY	<div style="width: 90%; height: 10px; background-color: blue; border: 1px solid orange;"></div>	November 2006
3	Tony Blair	<div style="width: 80%; height: 10px; background-color: blue; border: 1px solid orange;"></div>	PERSON	<div style="width: 80%; height: 10px; background-color: blue; border: 1px solid orange;"></div>	May 2007



Search! TextMap All Sources

[TextMap](#) : [TextMed](#) : [Textblq](#) : [TextBiz](#) : [Make homepage!](#) : [Link to us](#) : [Help?](#)

Vietnam COUNTRY

Sentiment Score: 67.3 + 21.9

Articles Referencing Vietnam [\[More Articles\]](#) [\(What is this?\)](#)

Title

- [VA hospital honors veterans with carnival](#)
- [Lead-tainted toys recalled](#)
- [Homemade explosives found in Fife](#)
- [Thompson is ho-hum in debate debut](#)
- [Central America faces new test in Asia](#)
- [Bush's fear factor](#)
- [Two doctors blame boot camp death on sickle cell](#)

Relational Network: [\(What is this?\)](#)

Referen

News S

Sentim

Top Performing Systems

- ...can answer ~70% of the questions
- Approaches:
 - Knowledge-rich approaches, using many NLP techniques (Harabagiu, Moldovan et al.-SMU/UTD/LCC)
 - AskMRS: shallow approach
 - Middle ground use large collection of surface matching patterns (ISI)

AskMRS: shallow approach

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

Abraham Lincoln, 1809-1865

"LINCOLN, ABRAHAM was born near Hodgenville, Kentucky, on February 12, 1809. In 1816, the Lincoln family moved to Pigeon Creek in Perry (now Spencer) County. Two years later, Abraham Lincoln's mother died and his father married a woman his 'angel' mother. Lincoln attended a formal school for only a few months but acquired knowledge through the reading of books. He moved to Illinois, in 1830 where he obtained a job as a store clerk and the local postmaster. He served without distinction in the Black Hawk War. He lost his attempt at the state legislature, but two years later he tried again, was successful, and Lincoln was admitted to the bar and became noteworthy as a witty, honest, competent circuit lawyer. He served a one-year term in the U.S. House in 1846, at which time he opposed the war with Mexico. By 1858 he had gained national attention for his series of debates with Stephen A. Douglas. He lost the election but became a significant figure in his party. At his inauguration on March 4, seven southern states had seceded from the Union. Lincoln called for 75,000 volunteers (approximately 43,000 responded), for a total of 11. Lincoln immediately took action. His leadership would eventually be the central difference in many of the battles. The Emancipation Proclamation which expanded the purpose of the war to the abolition of slavery. The dedication of a national cemetery in Gettysburg, Lincoln explained the meaning of the war. He was elected President in 1860. He was assassinated in 1865. He was General of the Army and Vice President at the time of his death.

ABRAHAM LINCOLN

Sixteenth President of the United States



Born in 1809 - Died in 1865

Sixteenth President
1861-1865
Married to Mary Todd Lincoln

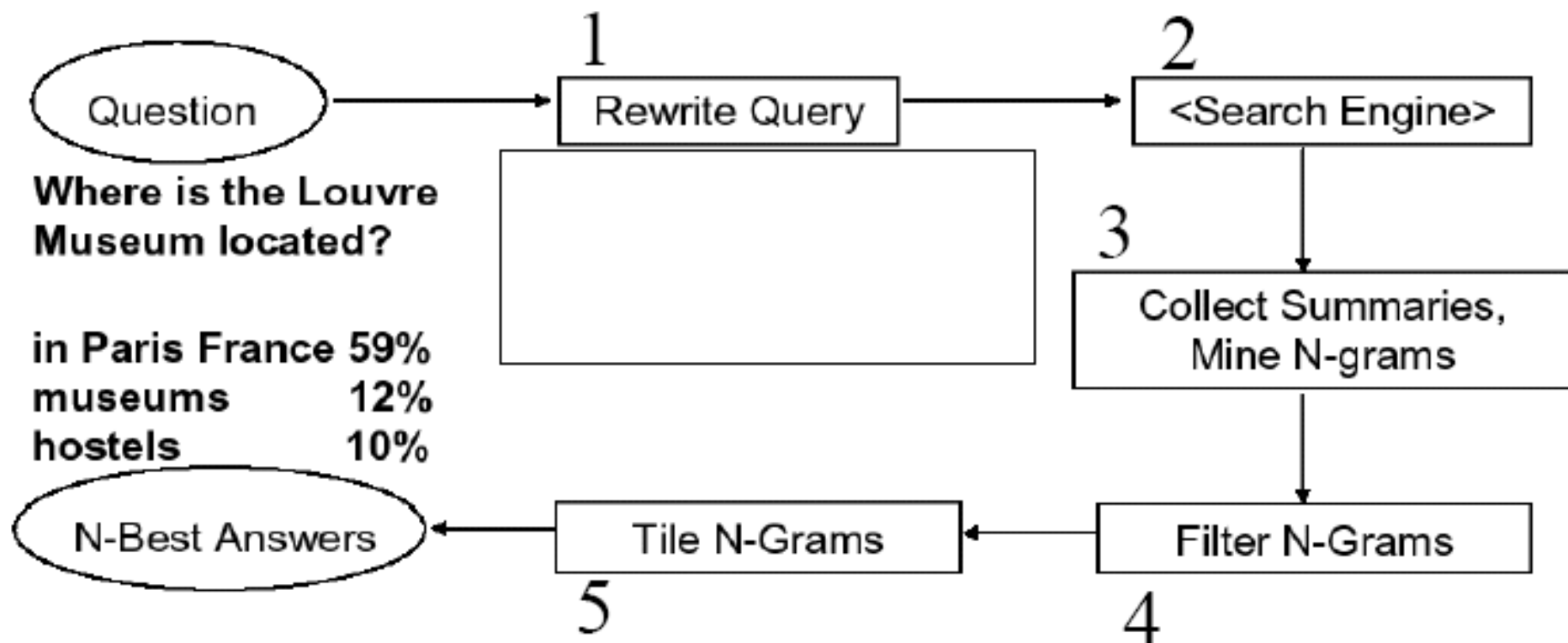
Abraham Lincoln

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

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



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 Scroogle?
 - ***Charles Dickens*** created the character of Scrooge.

Query rewriting

- Classify question into 7 categories
 - Who is/was/are/were...?
 - When is/did/will/are/were...?
 - Where is/are/were...?
- a) Category-specific transformation rules
 - E.g., For Where question, move “is” to all possible locations
 - Where **is** the Louvre Museum located?
 - **is** the Louvre Museum located?
 - the **is** Louvre Museum located?
 - the Louvre **is** Museum located?
 - the Louvre Museum **is** located?
 - the Louvre Museum located **is**?
- b) Expected answer “Datatype” (eg, Date, Person, Location,...)
 - When was the French Revolution? → DATE
- Hand-crafted classification/rewrite/datatype rules (Could they be automatically learned?)

Query Rewriting - weights

- Some query rewrites are more reliable than others

Where is the Louvre Museum located?

Weight 1

Lots of non-answers
could come back too

Weight 5

if we get a match,
it's probably right

+“the Louvre Museum is located”

+Louvre +Museum +located

Step 2: Query search engine

- Send all rewrites to a Web search engine
- Retrieve top N answers (100?)
- Rely just on search engine's words/phrases, not the full text of the actual document





Step 3: Mining N-Grams

- Unigram, bigram, trigram, ..., N-gram: list of N adjacent term in a sequence
 - Eg. “Web Question Answering: Is More Always Better”
 - Unigram: Web, Question, Answering, Is, More, Always, Better
 - Bigram: Web Question, Question Answering, Answering Is, Is More, More Always, Always Better
 - Trigram: ...

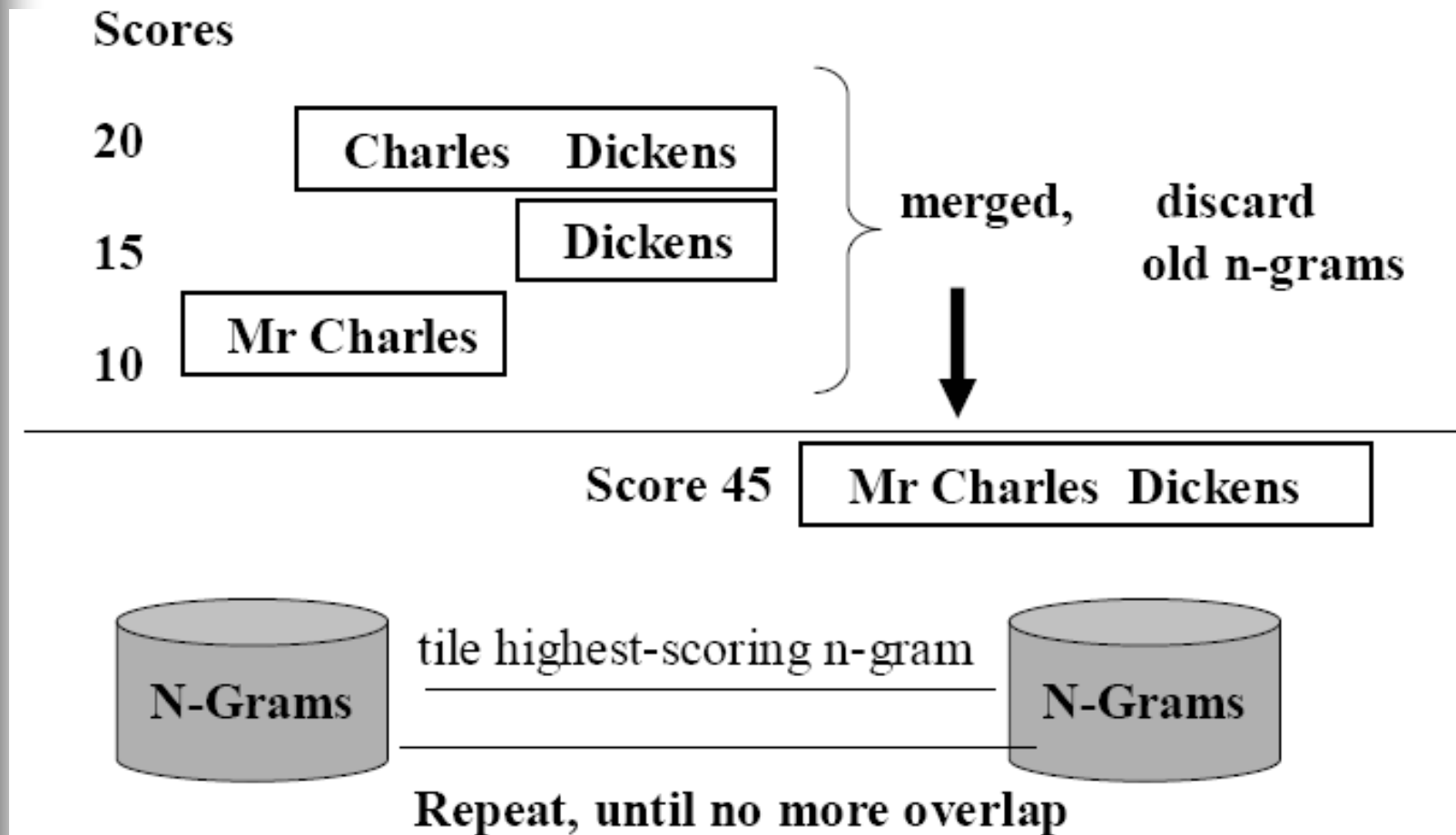
Mining N-grams

- Simple: Enumerate all N-grams (N=1,2,3...) in all retrieved phrases
 - Use hash table and other tools to make this efficient
- Weight of an n-gram: occurrence count
 - Eg, “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

Step 4: Filtering N-Grams

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

Step 5: Tiling the Answers



Results

- Standard TREC contest test-bed:
 - ~1M documents; 900 questions
 - Technique doesn't do well (but rank in top 9/30 participants!)
- Limitation:
 - Works best only for fact-based questions
 - Limited range of
 - Question categories
 - Answer data types
 - Query rewriting rules

Surface matching patterns (Ravichandran and Hovy, ISI)

- When was X born?
 - Mozart was born in 1756
 - Gandhi (1869—1948)
- <NAME> was born in <BIRTHDATE>
- <NAME> (<BIRTHDATE>-
- Use a Q-A pair to query a search engine
- Extract patterns and compute their accuracy

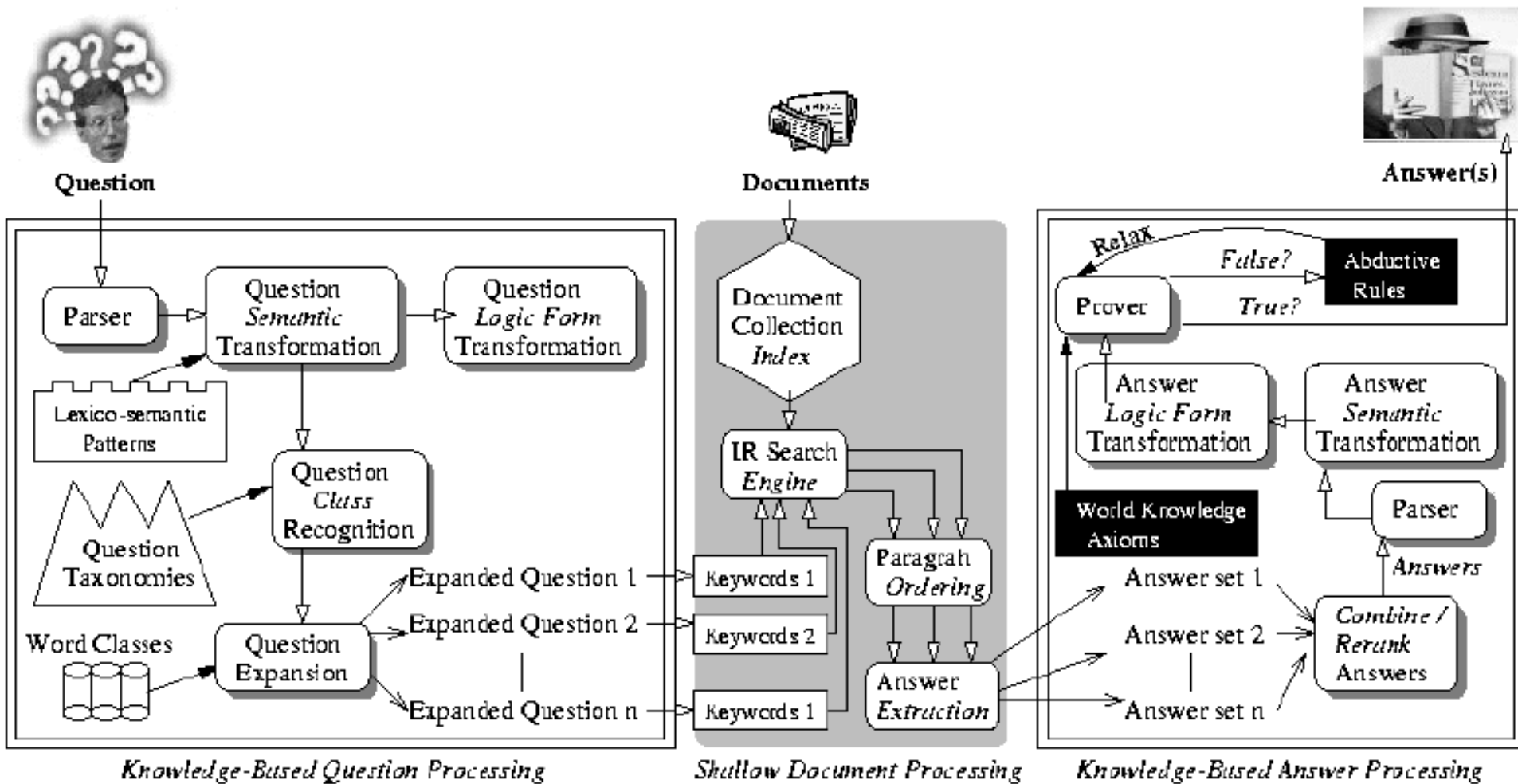
Example: INVENTOR

- <ANSWER> invents <NAME>
- the <NAME> was invented by <ANSWER>
- <ANSWER> invented the <NAME> in
- <ANSWER>'s invention of the <NAME>
- ...

- Many of these patterns have high accuracy
 - But still some mistakes

Full NLP QA

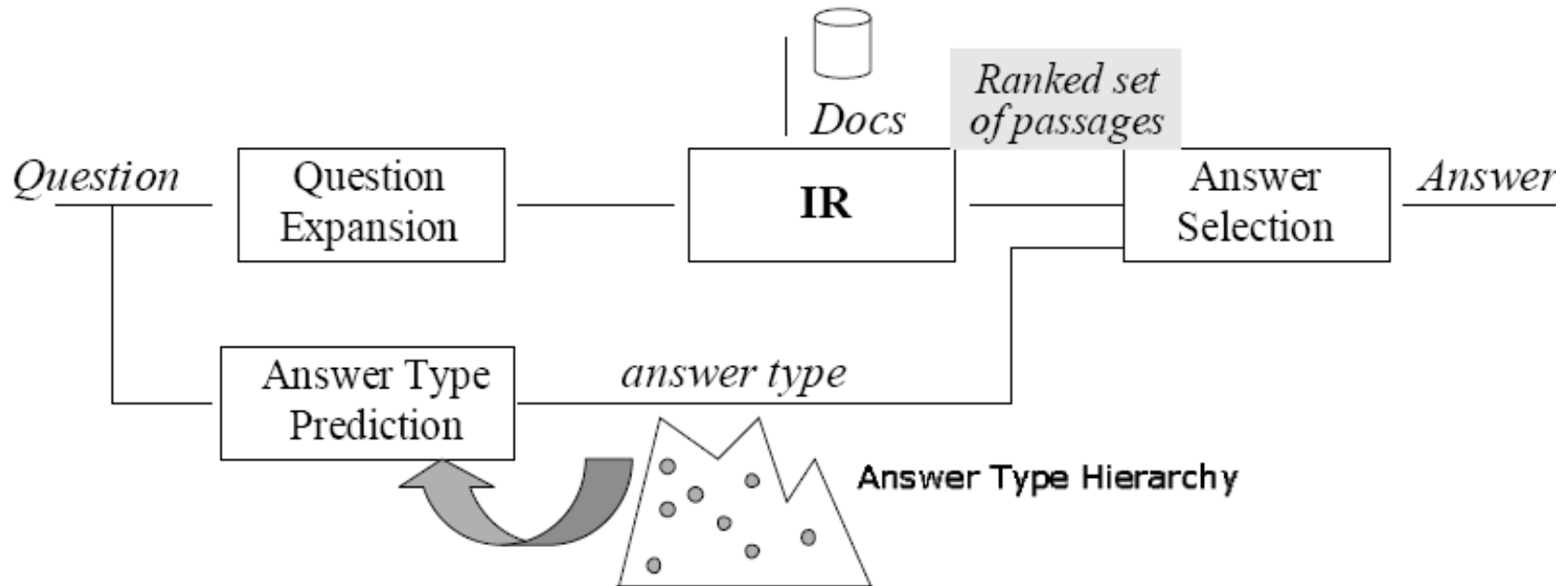
LCC: Harabagiu, Moldovan et al.



Value from sophisticated NLP – Pasca & Harabagiu (2001)

- Good IR is needed: SMART paragraph retrieval
- Large taxonomy of question types and expected answer types is crucial
- Statistical parser is 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 simply 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 (eg., by using a maximum entropy model)

Answer Types

- “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?

Keyword Selection Algorithm

Select all...

- Non-stopwords in quotations
- NNP words in recognized named entities
- Complex nominals with their adjectival modifiers
- Other complex nominals
- Nouns with adjectival modifiers
- Other nouns
- Verbs
- The answer type word

Passage Extraction Loop

- **Passage Extraction Component**
 - **Extracts passages that contain all selected keywords**
 - **Passage size/start position dynamic**
- **Passage quality and keyword adjustment**
 - 1st iteration: use the first 6 keyword selection heuristics
 - If #passages $< \theta$ \rightarrow query is too strict \rightarrow drop a keyword
 - If #passages $> \theta$ \rightarrow query is too relaxed \rightarrow add a keyword

Passage Scoring

Involve 3 scores:

- #words from the question that are recognized in the same sequence in the window
- #words that separate the most distant keywords in the window
- #unmatched keywords in the window

Rank candidate answers in the retrieved passages

- 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 Ailen, 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

Name Entity Recognition

- Current QA is determined by the recognition of name entities

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	URI	1
CITY	19	UNIVERSITY	3		

- Precision of recognition
- Coverage of name classes
- Mapping into concept hierarchies
- Participation into semantic relations (eg, predicate-argument structures or frame semantics)₃₂

Semantics and Reasoning for QA: Predicate-argument structure

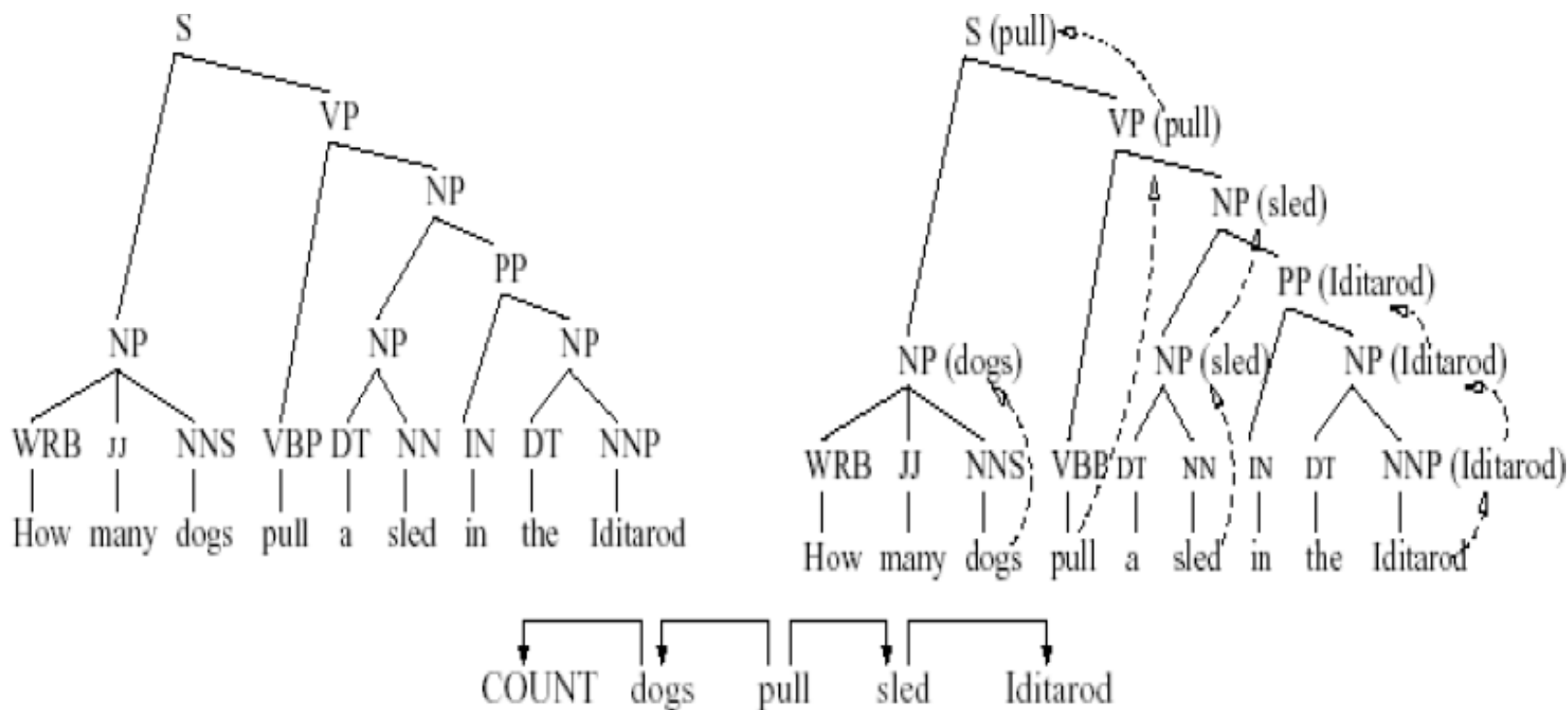
- *When was Microsoft established?*

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.
- Require 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

Inference

- System attempts inference to justify an answer (often following lexical chains)
- Their inference is a 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

Not all problems are solved yet!

- Where do lobsters like to live?
 - On a Canadian airline
- Where are zebras most likely found?
 - Nearly dumps
 - In the dictionary
- Why can't ostriches fly?
 - Because of American economic sanctions
- What's the population of Mexico?
 - Three