



ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Hệ hỏi đáp Question Answering

Viện CNTT & TT – Trường ĐHBKHN

Hệ hỏi đáp

- Lấy ý tưởng từ hệ tìm kiếm
- IR: find *relevant documents*, but we want *answers* from textbases
- QA: đưa ra câu hỏi ngắn, có thể kèm theo bằng chứng

Một số câu hỏi đáp từ tập TREC

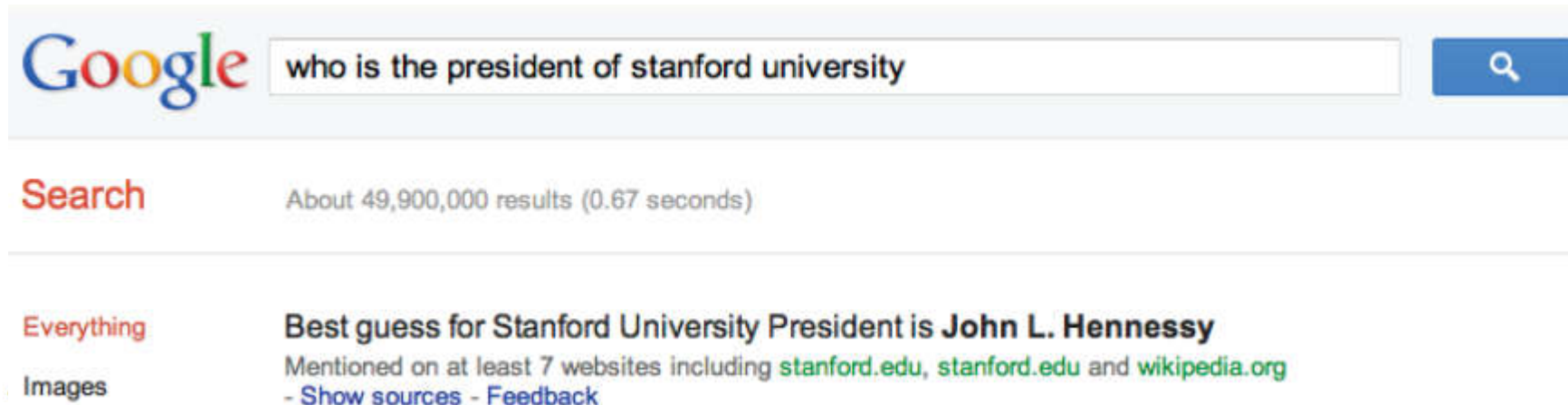
- Who is the author of the book “The Iron Lady: A Biography of Margaret Thatcher”?
- What was the monetary value of the Nobel Peace Prize in 1989?
- What does the Peugeot company manufacture?
- How much did Mercury spend on advertising in 1993?
- Why did David Koresh ask the FBI for a word processor?

Một số câu hỏi của con người

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

Một số mẫu từ web

- 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



Các cách tiếp cận

- Có bộ dữ liệu QA cho trước
 - Đo độ tương đồng câu
 - Lấy câu trả lời của câu hỏi tương đồng nhất
 - VD: AskJeeves
 - Huấn luyện sử dụng học sâu để dự đoán câu trả lời
- Không có bộ dữ liệu QA, có CSDL hoặc CSTT
 - Phân tích câu hỏi (sâu, so khớp mẫu,...)
 - Tìm câu trả lời (tra cứu CSDL, so khớp mẫu, suy diễn, ...)
 - VD: TextMap, AskMSR, LCC, ...

AskJeeves

- ... một ví dụ nhân tạo về hệ thống QA
- ... thực hiện so khớp mẫu để khớp câu hỏi với câu trả lời từ tập các câu QA có sẵn
 - Nếu có, đưa ra câu trả lời do con người tạo ra
 - Nếu không, trả về kết quả giống hệ thống tìm kiếm
- 1 hệ thống tầm trung tiềm năng, nhưng sử dụng ít kỹ thuật trong NLP

The screenshot shows the AskJeeves website interface. At the top, there's a navigation bar with the AskJeeves logo, a search bar containing the query "uk.ask.com/web?qsrc=1&o=0&l=dir&q=who+is+the+president+of+The+United+States++2012&dm=all", and links for "Advanced Search", "Settings", and "Your Cookie Choices". Below the navigation bar, the main content area is divided into two columns. The left column features a sidebar with categories like "Everything", "Images", "Video", "Reference", and "Q&A". The main content area on the left displays "Explore Answers About" with links to "United States History Timeline", "United States Atlas", "United States Road Map", "United States Facts", "Visitors Visa Requirements", and "A List of Presidents in Order". The right column displays "Popular Q&A" with two questions: "Q: Who the president of the united states 2012?" and "Q: Who is vice president of united states 2012?". Each question has an answer snippet and a "Read More" link. The "Find Answers" button is prominently displayed in the center of the page.

Answers

Advanced Search Settings Your Cookie Choices

AskJeeves

who is the president of The United States 2012

Find Answers

The Web UK Only

Explore Answers About

Everything Images Video Reference Q&A

United States History Timeline

United States Atlas

United States Road Map

United States Facts

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](#) ; [Textblg](#) ; [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 ...

SOURCES

CONTACT

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><div></div></div>	COUNTRY	<div><div></div></div>	November 2006
2	Iraq	<div><div></div></div>	COUNTRY	<div><div></div></div>	November 2006
3	Tony Blair	<div><div></div></div>	PERSON	<div><div></div></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

Các hệ thống đạt kết quả cao nhất


- ...có thể trả lời ~70% các câu hỏi
- Cách tiếp cận:
 - Sử dụng nguồn tri thức, các kỹ thuật NLP (Harabagiu, Moldovan et al.-SMU/UTD/LCC)
 - AskMRS: tiếp cận nông
 - Hệ thống tầm trung: sử dụng tập lớn các mẫu (ISI)

AskMSR: 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. In 1830, 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 year term in the U.S. House in 1846, at which time he opposed the war with Mexico. By 1854, 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. On his inauguration on March 4, seven southern states had seceded. Lincoln called for 75,000 volunteers (approximately 11,000 were accepted, for a total of 11). Lincoln immediately took action. The Emancipation Proclamation which expanded the purpose of the war to the abolition of slavery. The dedication of a national cemetery in Gettysburg, Lincoln's explanation of the war, and his final speech at Ford's Theatre.




ABRAHAM LINCOLN

Sixteenth President of the United States

Born in 1809 - Died in 1865

Sixteenth President


1861-1865
Married to Mary Todd Lincoln



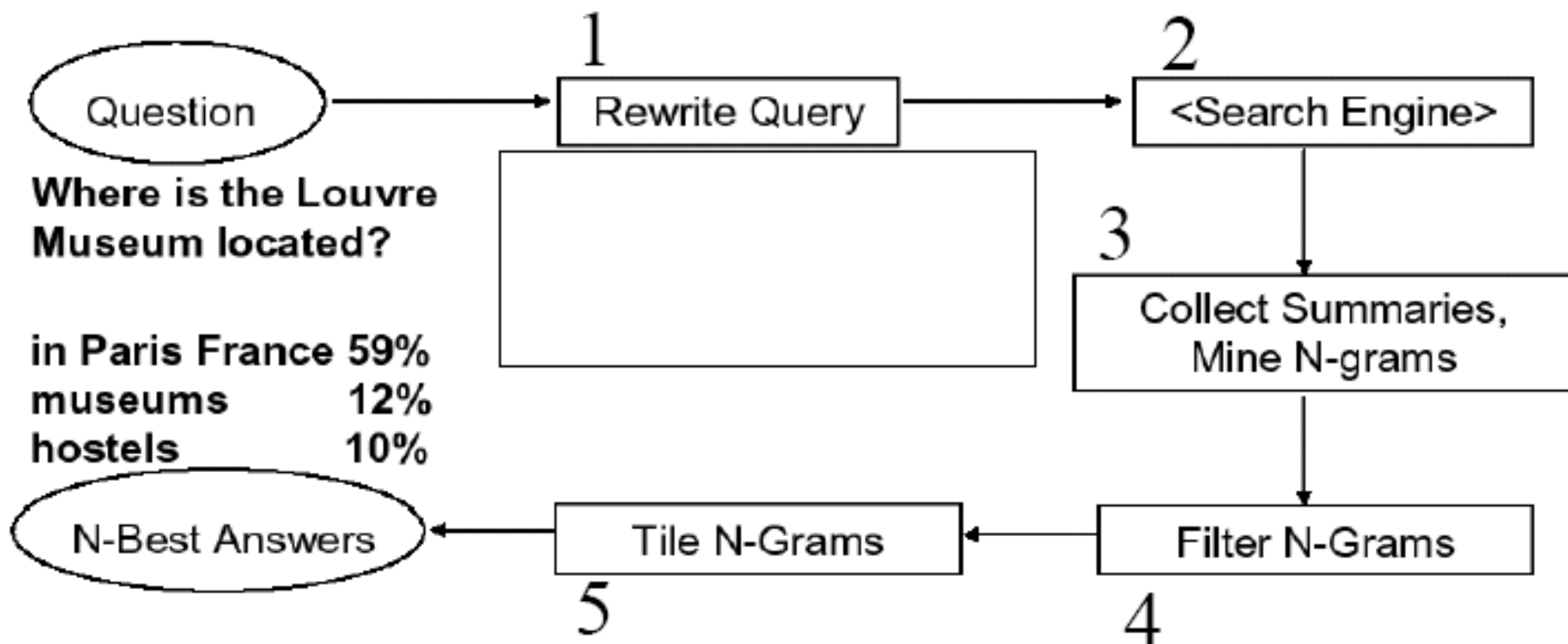
Abraham Lincoln

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 Lincoln."



AskMSR



Bước 1: Viết lại câu hỏi

- Ý tưởng: câu hỏi thường có ngữ pháp gần với câu trả lời
 - 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.

Viết lại câu hỏi

- 7 loại câu hỏi:
 - Who is/was/are/were...?
 - When is/did/will/are/were...?
 - Where is/are/were...?
- a) Luật biến đổi câu hỏi:
 - 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) Chờ câu trả lời dạng “Datatype” (eg, Date, Person, Location,...)
→ When was the French Revolution? → DATE
- Tạo luật thủ công để phân loại/viết lại

Viết lại câu hỏi – trọng số

- Một số câu hỏi đáng tin cậy hơn câu khác

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

Bước 2: Tìm kiếm

- Đưa tất cả mẫu tìm kiếm lên Web search engine
- Lấy top N câu trả lời (100?)
- Chỉ dựa trên từ/cụm từ của công cụ tìm kiếm, không dựa vào toàn bộ nội dung của tài liệu thực tế





Bước 3: Khai thác N-Grams

- Unigram, bigram, trigram, ..., N-gram: danh sách chuỗi N term
 - VD. “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: ...

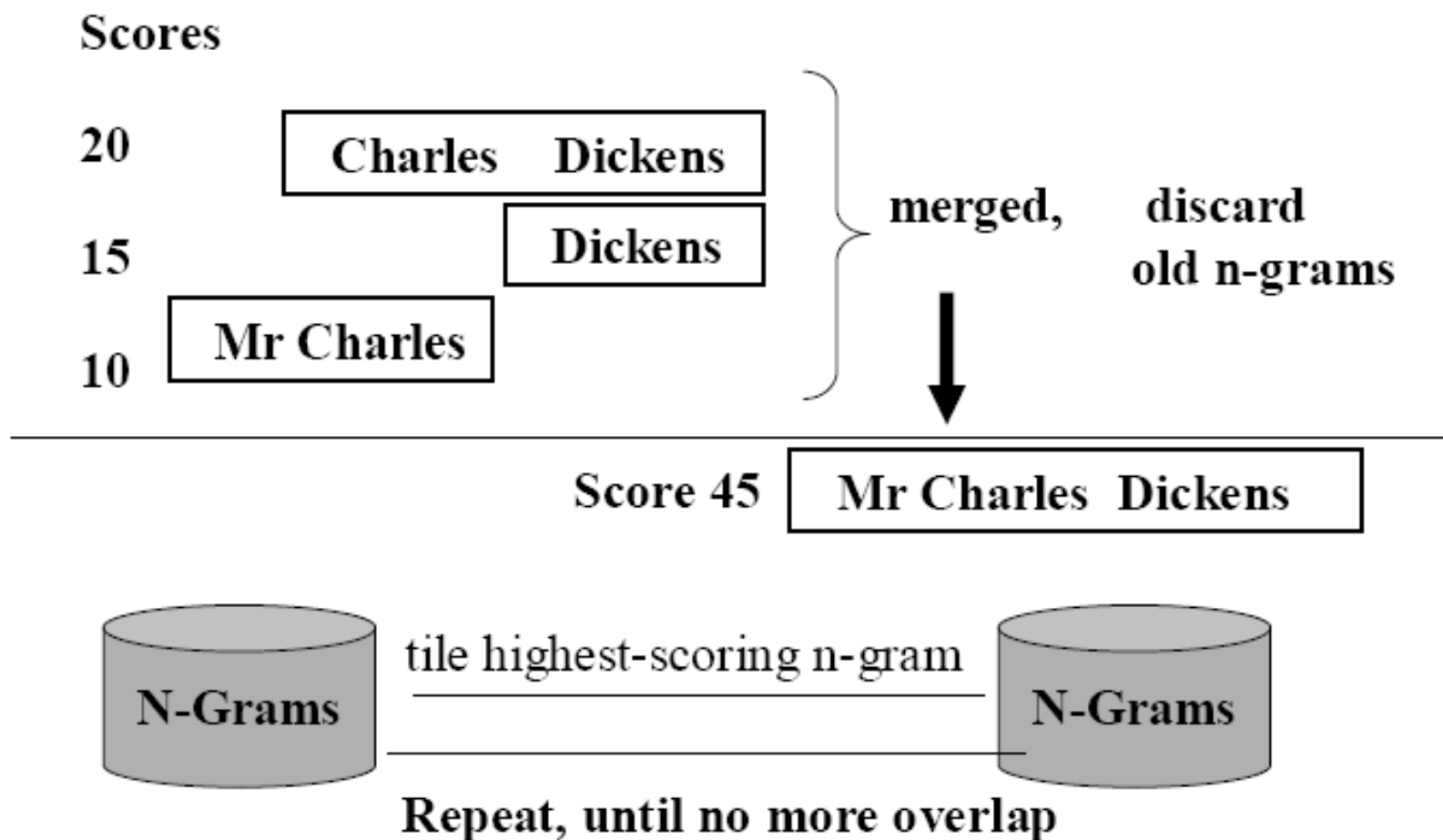
Mining N-grams

- Đơn giản: Liệt kê tất cả N-grams ($N=1,2,3\dots$) trong tất cả các đoạn trả về
 - Sử dụng bảng băm và một số tool khác để tìm kiếm nhanh
- Trọng số của n-gram: đến số lần xuất hiện
 - VD, “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

Bước 4: Lọc ngrams

- Mỗi câu hỏi đi kèm với 1 hoặc nhiều bộ lọc kiểu dữ liệu = regular expression
- When...  Date
- Where...  Location
- What...  Location
- Who...  Person
- Tăng điểm của ngrams khớp với regexp
- Giảm điểm ngrams khớp

Bước 5: Trộn câu trả lời



Kết quả

- Cuộc thi TREC :
 - ~1M tài liệu; 900 câu hỏi
 - Về kỹ thuật, hoạt động không tốt (nhưng xếp hạng 9/30 số đội tham gia)
- Giới hạn:
 - Làm việc tốt với các câu hỏi về sự kiện (fact)
 - Giới hạn bởi:
 - Loại câu hỏi
 - Kiểu câu trả lời
 - Tập luật viết lại câu hỏi

So khớp mẫu bề mặt (Ravichandran and Hovy, ISI)

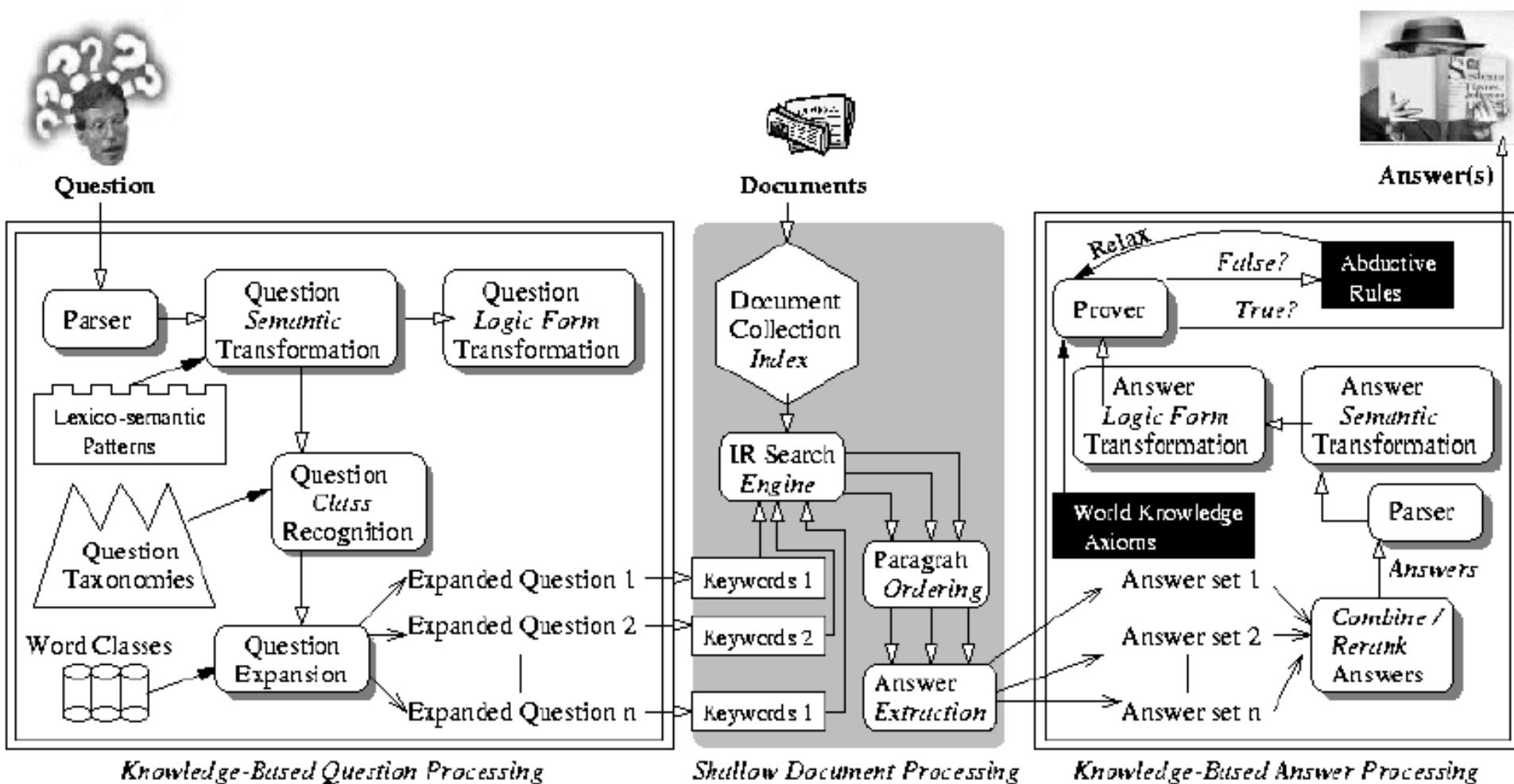
- When was X born?
 - Mozart was born in 1756
 - Gandhi (1869—1948)
- <NAME> was born in <BIRTHDATE>
- <NAME> (<BIRTHDATE>-
- Sử dụng cặp Q-A để truy vấn trên search engine
- Trích ra các mẫu và tính độ chính xác

Ví dụ: INVENTOR

- <ANSWER> invents <NAME>
 - the <NAME> was invented by <ANSWER>
 - <ANSWER> invented the <NAME> in
 - <ANSWER>'s invention of the <NAME>
 - ...
-
- Phần lớn các mẫu có độ chính xác cao
 - Nhưng vẫn có lỗi

Full NLP QA

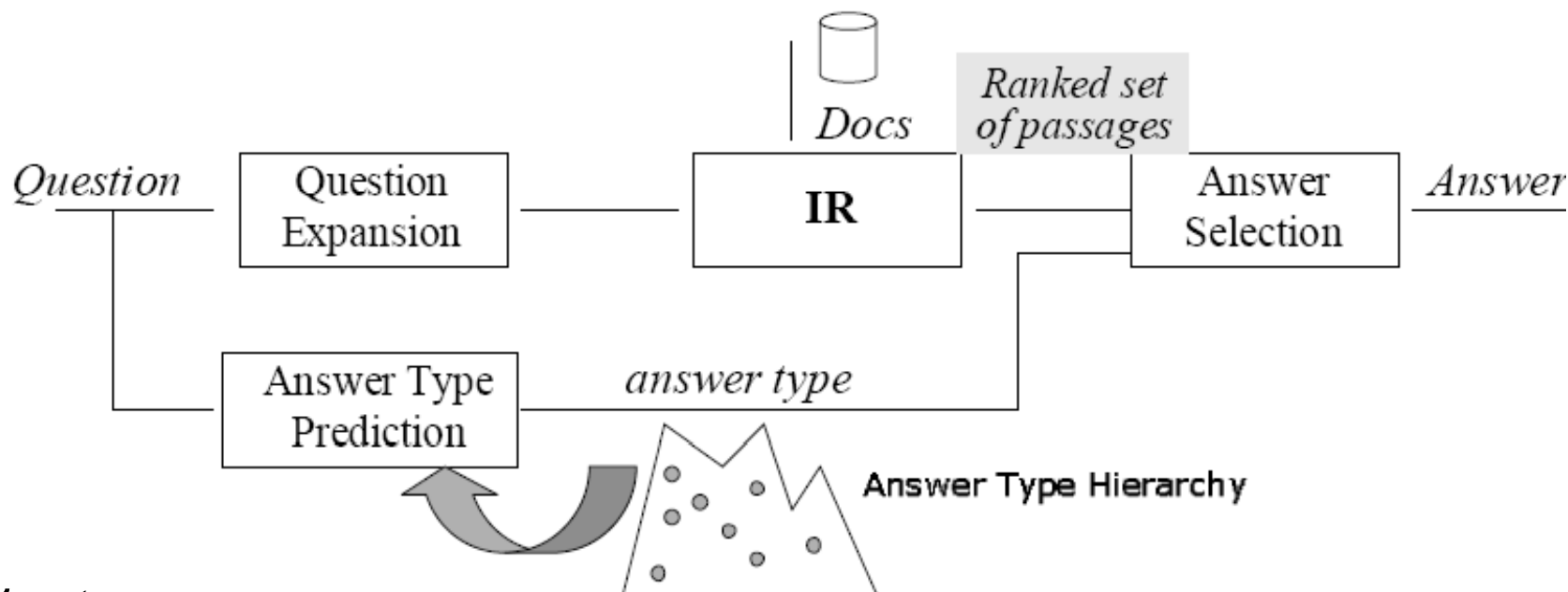
LCC: Harabagiu, Moldovan et al.



Hệ thống phức tạp NLP – Pasca & Harabagiu (2001)

- Cần công cụ tìm kiếm tốt để tìm ra các đoạn liên quan
- Cần một bộ từ vựng (taxonomy) các loại câu hỏi và dạng câu trả lời
- Bộ PTCP thống kê được dùng để phân tích câu hỏi và các văn bản liên quan để tìm câu trả lời, và xây dựng CSTT.
- Lặp lại việc mở rộng câu hỏi (về hình thái, từ đồng nghĩa, quan hệ ngữ nghĩa)
- Xếp hạng câu trả lời dựa trên học máy

Các dạng câu trả lời



Đặc trưng:

- Kiểu trả lời:
 - Gán nhãn câu hỏi và dạng câu trả lời dựa trên tập taxonomy
 - Phân loại câu hỏi (vd., sử dụng maximum entropy)

Các dạng câu trả lời

- Câu hỏi “Who” có thể có câu trả lời là tổ chức
 - Who sells the most hybrid cars?
- Câu hỏi “Which” có thể có câu trả lời là người
 - Which president went to war with Mexico?

Thuật toán lựa chọn từ khóa

Lựa chọn tất cả...

- Từ không phải từ dừng trong “ ”
- Tên riêng
- Tên riêng đi kèm với tính từ
- Danh từ đi kèm với tính từ
- Danh từ
- Động từ
- Từ chỉ dạng câu trả lời

Vòng lặp trích rút đoạn

- Thành phần trích rút đoạn
 - Trích rút các đoạn chứa tất cả các từ khóa
 - Kích thước, vị trí đoạn là động
- Tinh chỉnh chất lượng đoạn và điều chỉnh từ khóa
 - Vòng lặp 1: sử dụng 6 từ khóa có điểm cao nhất
 - If #passages < θ \rightarrow query chặt quá \rightarrow bỏ bớt 1 từ khóa
 - If #passages > θ \rightarrow query lỏng quá \rightarrow thêm 1 từ khóa

Tính điểm đoạn

Bao gồm 3 điểm:

- Số từ trong câu hỏi theo đúng trật tự trong cửa sổ
- Số từ nằm giữa 2 từ khóa xa nhất trong cửa sổ
- Số từ không khớp trong cửa sổ

Xếp hạng các câu trả lời ứng cử trong các đoạn kết quả tìm kiếm

- Name the first private citizen to fly in space
- Kiểu câu trả lời: Person
- Đoạn:

“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...”
- Câu trả lời ứng cử tốt nhất: Christa McAuliffe

Nhận diện thực thể

- Các hệ thống QA hiện tại phụ thuộc nhiều vào việc nhận dạng thực thể

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		

- Độ chính xác nhận dạng
- Độ phủ của các lớp tên
- Ánh xạ sang cây phân cấp khái niệm
- Tham gia vào các quan hệ ngữ nghĩa (vd, cấu trúc predicate-argument của khung ngữ nghĩa)

Ngữ nghĩa và lập luận cho QA: Cấu trúc predicate-argument

- *When was Microsoft established?*

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

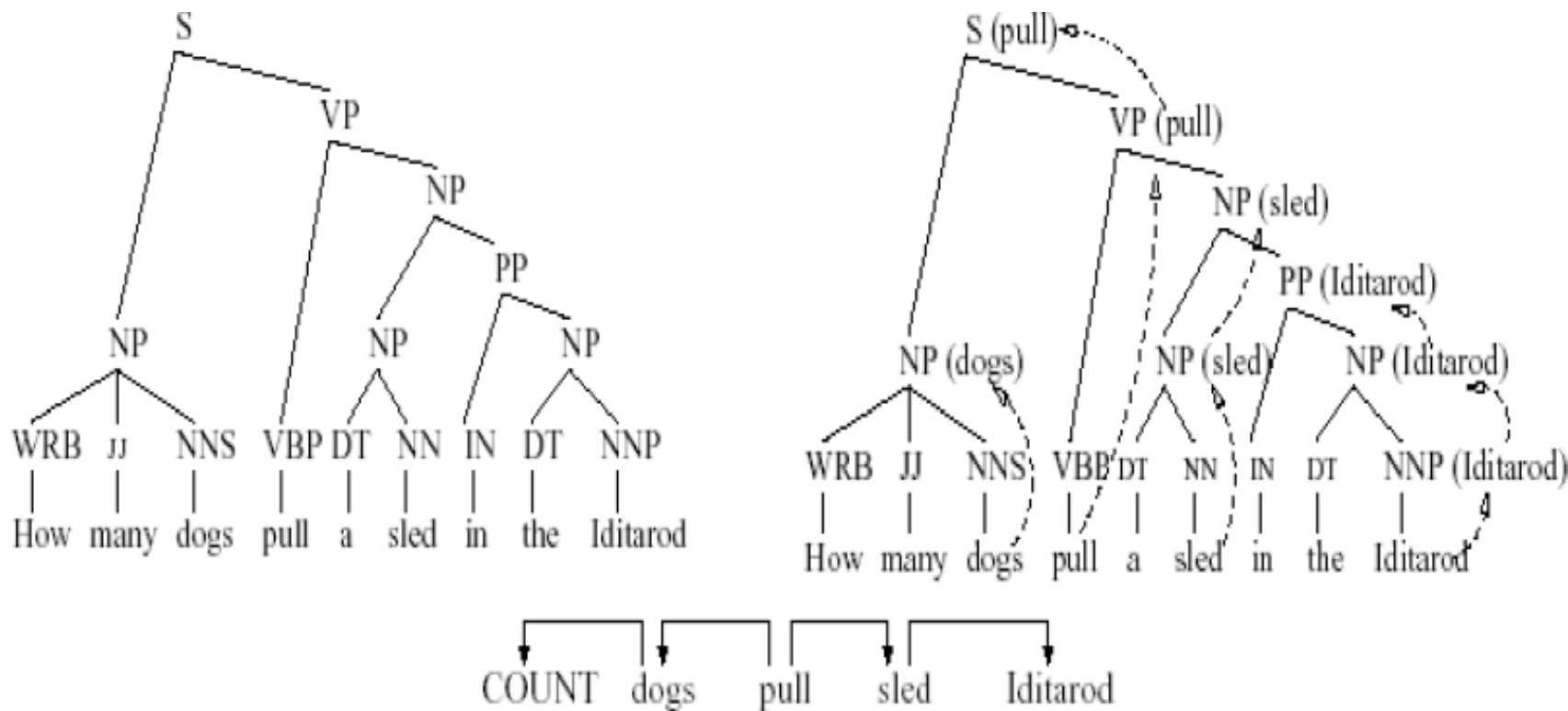
- Cần phát hiện câu trong đó ‘Microsoft’ là đối tượng của ‘establish’ hoặc từ đồng nghĩa gần.

- Câu khớp:

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

- Cần phân tích cú pháp/ngữ nghĩa của câu

Ngữ nghĩa và lập luận cho QA: từ cú pháp đến biểu diễn logic



- PTCP + ngữ nghĩa → logical form
- Khớp câu hỏi và câu hỏi tiềm năng để tìm kết quả phù hợp

Suy diễn

- Hệ thống suy diễn để xác định câu trả lời (thường dùng lexical chains)
- Suy diễn là trung gian giữa logic và so khớp từ khóa
- Khá hiệu quả: cải thiện 30%
- Q: When was the internal combustion engine invented?
- A: The first internal-combustion engine was built in 1867.
- Invent → create_mentally → create → build

Một số platform

- Tập dữ liệu mẫu:
 - Squad
 - VLSP 2020

Electra – mô hình QA

- Thu thập dữ liệu đủ lớn (wikipedia)
- Huấn luyện mô hình
 - Code từ Hugging Face
 - <https://huggingface.co/blog/how-to-train>
 - Code từ blog
 - <https://towardsdatascience.com/understanding-electra-and-training-an-electra-language-model-3d33e3a9660d>
 - Thường cần GPU hoặc TPU
- Sử dụng mô hình vào các bài toán cụ thể