

# Question Answering Systems & Chatbots!

Conal Sathi and Sameer Singh  
(featuring guest lecture by Arvind Sathi!)

---

BANA 290: ADVANCED DATA ANALYTICS

MACHINE LEARNING FOR TEXT

SPRING 2018

May 29, 2018

\*Much thanks to Professor Dan Jurafsky for the content of these slides

# Upcoming...

---

## Homework

- Homework 3 due today!
- Homework 2 grades will be released today

## Project

- Feedback regarding proposals sent
- Please meet with us if you need/want to!
- Progress presentations next class!

## Office Hours

- Conal's office hours 4:15pm-6:15pm tomorrow (Wed 30<sup>th</sup>)
- Feel free to come by to discuss Final Project or Data Science post-graduation

# Question Answering Systems

---

CONCEPT & ARCHITECTURE

# Question Answering Systems

---

Given a question inputted by human, have a system to automatically provide an answer

What do worms eat?


- Potential answers:
  - Worms eat grass
  - Horses with worms eat grass
  - Birds eat worms
  - Grass is eaten by worms
- How to choose the best answer?

# Examples of Question Answering Systems?

---

# Examples of Question Answering Systems?

---



Search

About 904,000 results (0.30 seconds)

Everything

Images

Maps

Videos

News

Best guess for Louvre Museum Location is **Paris, France**  
Mentioned on at least 7 websites including [wikipedia.org](#), [answers.com](#) and [east-buc.k12.ia.us](#) - [Show sources](#) - [Feedback](#)

[Musée du Louvre - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Musée\\_du\\_Louvre](#)  
Musée du **Louvre** is **located** in Paris. **Location** within Paris. Established, 1793. **Location, Palais Royal, Musée du Louvre, 75001 Paris, France.** Type, Art **museum** ...  
[Louvre Palace](#) - [List of works in the Louvre](#) - [Category:Musée du Louvre](#)

# Examples of Question Answering Systems?

---



# Examples of Question Answering Systems?

---

IBM's Watson won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S  
"AN ACCOUNT OF THE PRINCIPALITIES OF  
WALLACHIA AND MOLDOVIA"  
INSPIRED THIS AUTHOR'S  
MOST FAMOUS NOVEL



Bram Stoker



# Examples of Question Answering Systems?

 **WolframAlpha**<sup>™</sup> computational knowledge engine

how many calories are in two slices of banana cream pie

Examples Random

Assuming any type of pie, banana cream | Use pie, banana cream, prepared from recipe or pie, banana cream, no-bake type, prepared from mix instead

Input interpretation:

pie	amount	2 slices	total calories
	type	banana cream	

Average result: Show details

702 Cal (dietary Calories)

# Types of Questions in QA systems

---

## Factoid questions

- *Who wrote “The Universal Declaration of Human Rights”?*
- *How many calories are there in two slices of apple pie?*
- *What is the average age of the onset of autism?*
- *Where is Apple Computer based?*

## Complex (narrative) questions

- *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
- *What do scholars think about Jefferson’s position on dealing with pirates?*

*For these examples discussed previously, which types of questions do they deal with?*

# Commercial systems mainly deal with factoid questions

Where is the Louvre Museum located?	In Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What is the telephone number for UCI?	(949) 824-5011

# Approaches for QA

---

Information-Retrieval approaches

Knowledge-based and hybrid approaches

# Approaches for QA

---

## Information-Retrieval approaches

- System stores a set of documents
- Given a user question, information retrieval techniques extract passages directly from this set of documents, guided by the text of the question

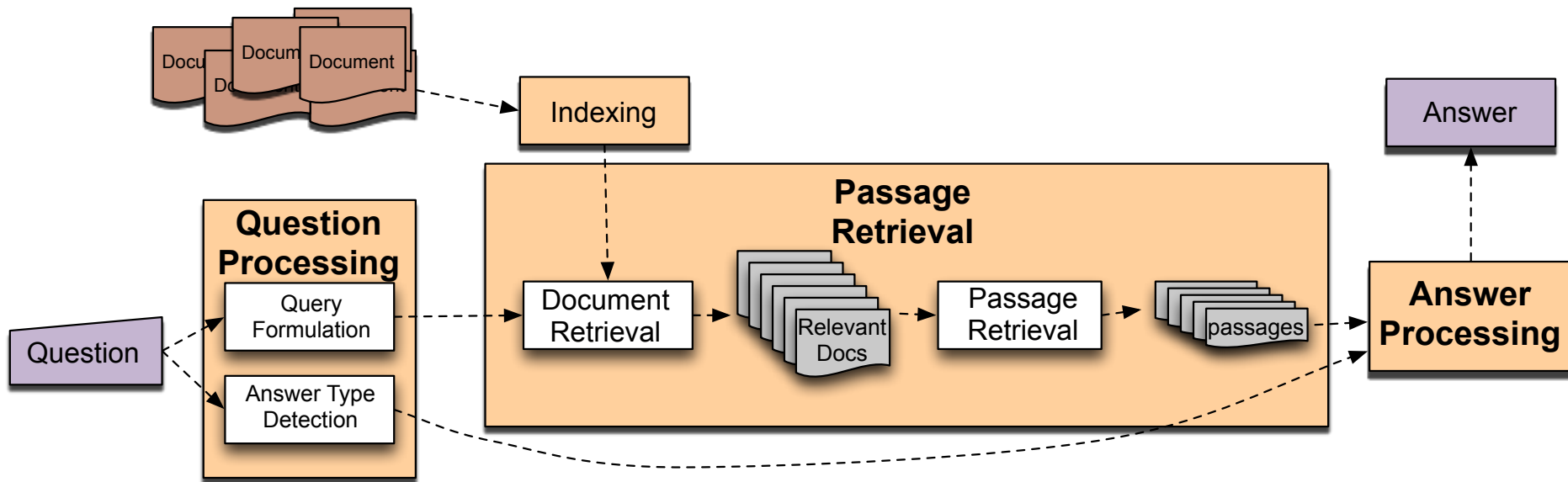
## Knowledge-based

- System has a database of facts/knowledge about the world
- Convert user question to a semantic query (like SQL)
- Query database of facts

Which approaches do the examples discussed previously match to?

Examples discussed: Google, Siri, Wolfram Alfa, IBM Watson for Jeopardy

# IR-based Factoid QA



# Question Processing

---

# Question Processing

---

## Query formulation

- Choose keywords for IR-system

## Answer-type detection

- Decide what kind of answer needs to be returned (e.g. person, city, quantity, etc.)



# Query formulation

---

Examples of questions that users could enter:

- "Who founded the company Virgin Airlines"
- "When was the laser invented"
- "Which two states you could be reentering if you're crossing Florida's northern border"
- "Who coined the term 'data science' in the tech community"

Why not just give the entire question – why do we need to select keywords?

- Might need to remove words or change the order of the words
- May need to apply query expansion (morphological variants or synonyms)

# Answer-type detection

---

*Who founded Virgin Airlines?*

- PERSON

*What Canadian city has the largest population?*

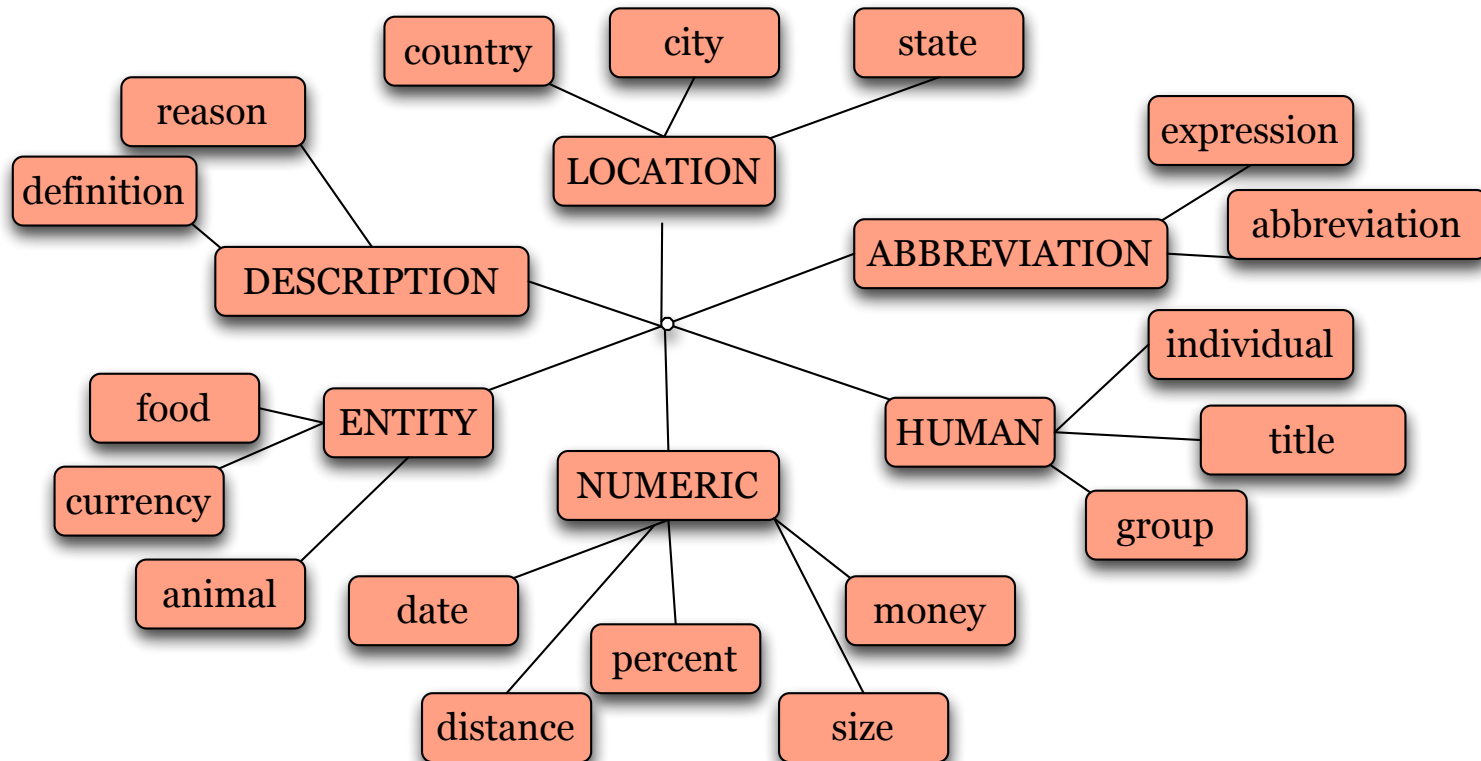
- CITY

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

- Two level answer-type category taxonomy
- 6 categories at level 1
- 50 categories at level 2

# Answer-type detection

---



# Answer-types in Jeopardy

---

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types

he, country, city, man, film, state, she, author, group, here, company, president, capital, star, novel, character, woman, river, island, king, song, part, series, sport, singer, actor, play, team, show, actress, animal, presidential, composer, musical, nation, book, title, leader, game

See Ferrucci et al. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. Fall 2010. 59-79.

# How would you do answer-type detection?

---

Hint, what did we cover in this class...

- Hand-written rules
- Machine Learning
- Hybrid solutions

# How would you do answer-type detection?

---

Regular expression-based rules can help us in some cases

- If the question starts with “Who” -> PERSON

Can use a parser to find the headword of the first noun phrase after the wh- word

- Which **city** in China has the largest number of foreign financial companies?
- What is the state **flower** of California?
- You can play with <http://corenlp.run/>

# Treating it as a machine learning text classification problem

---

- Define a taxonomy of answer types
- Get labeled training data for each answer type
- Define a set of features
  - Can use the rules as features
- Train a classifier

# Passage Retrieval

---



# Passage Retrieval

---

Finding a block of text in a document that likely contains the answer

- Document Retrieval
  - Which document is likely to contain the answer
- Passage Segmentation
  - Segment the document into shorter units
- Passage ranking
  - Which passage is likely to contain the answer

# Document retrieval

---

- Use your favorite search engine algorithm to find the documents that match the query
- How would you do this? We mentioned a solution in a previous class?
  - Remember tf-idf and cos similarity?

# Passage ranking

---

- Which passages match the query
- How would you do this?
  - Hint... what have we covered in this class?

# Passage ranking

---

- Treat it like a supervised text classification problem and sort by confidence value
- What kinds of features would you use?
  - Number of Named Entities of the right type in passage
  - Number of query words in passage
  - Number of query N-grams also in passage
  - Proximity of query keywords to each other in passage
  - Longest sequence of question words
  - Rank of the document containing passage

# What's left?

---

We've processed the question and now have a passage that likely contains the answer

Still need to extract the actual answer. People are too impatient to read the entire passage...

# Answer Extraction

---

Run an answer-type named-entity tagger on the passages

- i.e. if you classify the answer-type of the question to be PERSON, look for a PERSON in the passage

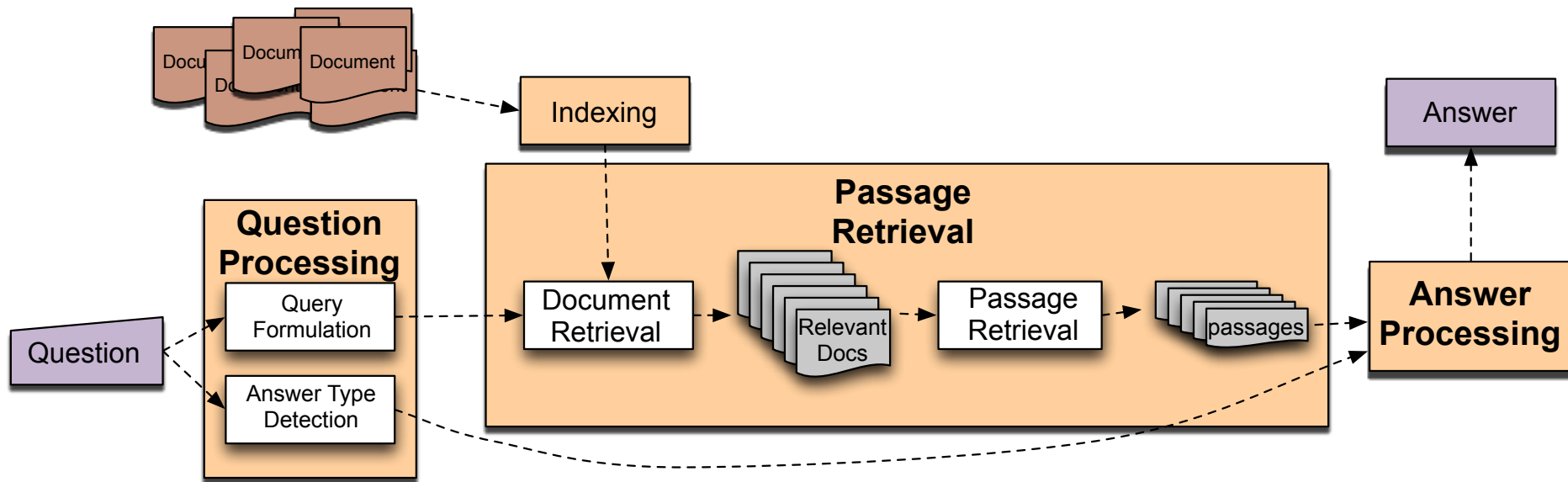
- Who is the prime minister of India? -> PERSON

Manmohan Singh, Prime Minister of India, had told leaders that the deal would not be renegotiated.

- How tall is Mt. Everest? -> LENGTH

The official height of Mount Everest is 29035 feet

# IR-based Factoid QA



# What are cons of the IR based approach discussed?

---

Only input is unstructured text documents

- Not incorporating knowledge of the world
- Not normalizing the data so may get different answers for different questions

There are benefits of using databases of facts/relations -> knowledge-based solution

Watson for Jeopardy was a hybrid solution incorporating both IR based approach and knowledge in the world



# In-Class Activity 1

---