Slide 1

CS 6120/CS 4120: Natural Language Processing

Instructor: Prof. Lu Wang
College of Computer and Information Science
Northeastern University
Webpage: www.ccs.neu.edu/home/luwang

1

Slide 2

Question Answering

2

Slide 3

Questions in Search

3

Slide 4

Questions in Search

4

Slide 5

Question Answering (Some Background)

One of the oldest NLP tasks (punched card systems in 1961)

Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

5

Slide 6

Question Answering

One of the oldest NLP tasks (punched card systems in 1961)

6

Question Answering

One of the oldest NLP tasks (punched card systems in 1961)

Question: Potential Answers:

What do worms eat?   Worms eat grass   Horses with worms eat grass

```
worms                    worms                horses
  ↑                        ↑                    ↑
 eat                      eat                 with  eat
  ↑                        ↑                    ↑     ↑
 what                    grass               worms  grass
```

Birds eat worms   Grass is eaten by worms

```
birds                    worms
  ↑                        ↑
 eat                      eat
  ↑                        ↑
 worms                   grass
```

7

Question Answering: 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

8

Apple's Siri

9

Types of Questions in Modern 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?*

10

Types of Questions in Modern Systems

• Factoid questions   ← **This lecture**
  • *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?*

11

Commercial systems: mainly factoid questions

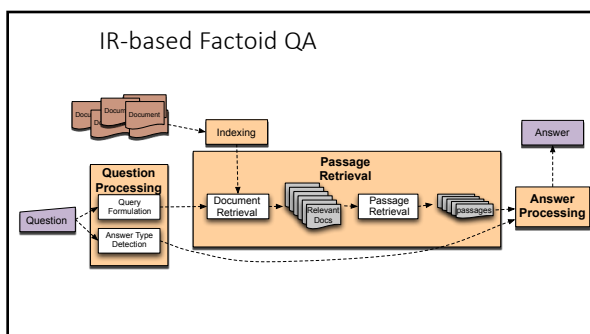|	
Where is the Louvre Museum located?	In Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums

12

Paradigms for Factoid QA

- Information Retrieval (IR)-based approaches
  - IBM Watson (some parts); Google
- Knowledge-based and Hybrid approaches
  - IBM Watson; Apple Siri; Wolfram Alpha
- Built upon the above two:
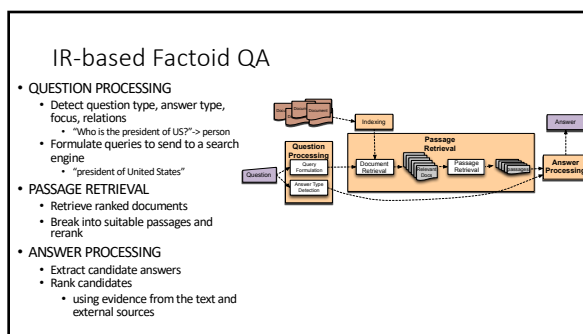 - Data-driven, neural network-based approaches (more recent)

13

Information Retrieval (IR)-based QA

- Factoid QA pipeline
- Factoid QA evaluation
- Common Knowledge used in QA
- Recent QA tasks

14

IR-based Factoid QA

15

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
    - "Who is the president of US?"-> person
  - Formulate queries to send to a search engine
    - "president of United States"
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

16

Question Processing:
Things to extract from the question

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction (if there are more than one entities)
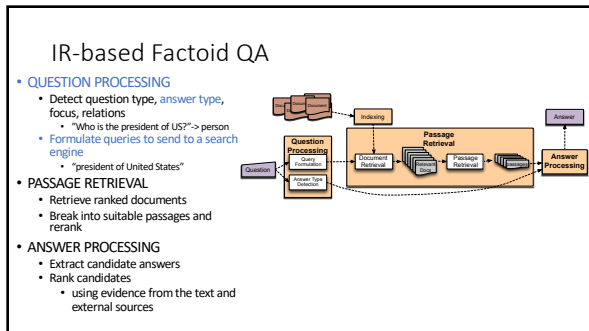 - Find relations between entities in the question

17

Question Processing

Jeopardy!: They're the two states you could be reentering if you're crossing Florida's northern border
*You should answer: what are the states of **Georgia and Alabama**?*

- Answer Type: US state
- Query Formulation: two states, border, Florida, north
- Focus: the two states
- Relations: borders(Florida, ?x, north)

18

Slide 19

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
    - "Who is the president of US?"-> person
  - Formulate queries to send to a search engine
    - "president of United States"
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

19

Slide 20

Answer Type Detection: Named Entities

- *Who founded Virgin Airlines?*

20

Slide 21

Answer Type Detection: Named Entities

- *Who founded Virgin Airlines?*
  - PERSON
- *What Canadian city has the largest population?*

21

Slide 22

Answer Type Detection: Named Entities

- *Who founded Virgin Airlines?*
  - PERSON
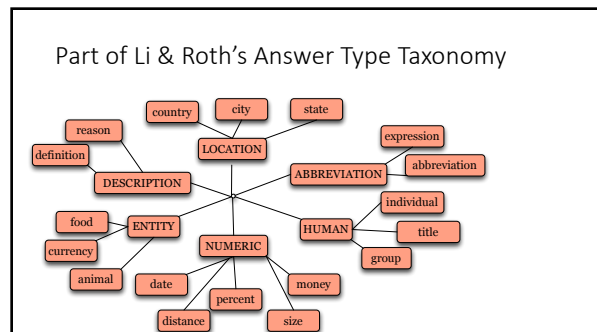- *What Canadian city has the largest population?*
  - CITY

22

Slide 23

Answer Type Taxonomy

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

- 6 coarse classes
  - ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title, description…
  - ENTITY: animal, body, color, currency…

23

Slide 24

Part of Li & Roth's Answer Type Taxonomy

24

Answer Types

ENTITY

animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

25

More Answer Types

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

26

Answer types in Jeopardy

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

- 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

country, city, man, film, state, 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

27

Answer Type Detection

- Hand-written rules
- Machine Learning

28

Answer Type Detection

- Regular expression-based rules can get some cases:
  - Who {is|was|are|were} PERSON
  - PERSON (YEAR – YEAR)
- Other rules use the **question headword**:
  (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?

29

Answer Type Detection

- Most often, we treat the problem as machine learning classification
  - **Define** a taxonomy of question types
  - **Annotate** training data for each question type
  - **Train** classifiers for each question class using a rich set of features.
    - features include those hand-written rules!

30

Features for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

Which **city** in China has the largest number of foreign financial companies?

What is the state **flower** of California?

31

Query Formulation

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
    - "Who is the president of US?"-> person
  - Formulate queries to send to a search engine
    - "president of United States"
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

32

Keyword Selection Algorithm

Dan Moldovan, Sanda Harabagiu, Marius Paca, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. 1999. Proceedings of TREC-8.

1. Select all non-stop words 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 their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the question focus word (skipped in all previous steps)
10. Select all other words

33

Choosing keywords from the query

Slide from Mihai Surdeanu

Who coined the term "cyberspace" in his novel "Neuromancer"?

1

1

4

4

7

cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

34

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
    - "Who is the president of US?"-> person
  - Formulate queries to send to a search engine
    - "president of United States"
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

35

Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - E.g. paragraphs or consecutive sentences
- Step 3: Passage ranking
  - Use answer type to help rerank passages

36

Features for Passage Ranking

Either in rule-based classifiers or with supervised machine learning

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question 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

37

Passage Retrieval as Query-focused Summarization

Which country has the largest part of the Amazon rain forest?

[The chaotic development that is gobbling up the Amazon rain forest could finally be reined in with a new plan developed by leading scientists from around the world.] ["That's some of the most encouraging news about the Amazon rain forest in recent years," said Thomas Lovejoy, an Amazon specialist.] ["It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon."]

[Sixty percent of the Amazon, the world's largest tropical rain forest, lies in Brazil.]

Extract passages that best summarize each document w.r.t. the query

38

Passage Retrieval as Query-focused Summarization

- Decide on a summary length (10% of document length).
- Use standard ad-hoc retrieval algorithm to retrieve top k documents.
- Treat each sentence/paragraph in top N documents as a document itself.
  - Use standard document similarity equations to assign a similarity score to the sentence/paragraph.
- Return highest-scoring sentences/paragraphs as the summary, subject to the length constraint.

39

Passage Retrieval as Query-focused Summarization

Which country has the largest part of the Amazon rain forest?

[The chaotic development that is gobbling up the Amazon rain forest could finally be reined in with a new plan developed by leading scientists from around the world.] ["That's some of the most encouraging news about the Amazon rain forest in recent years," said Thomas Lovejoy, an Amazon specialist.] ["It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon."]

[Sixty percent of the Amazon, the world's largest tropical rain forest, lies in Brazil.]

Sort summary extracts across top k documents

ordered list of summary extracts

answer hypotheses

40

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
    - "Who is the president of US?"-> person
  - Formulate queries to send to a search engine
    - "president of United States"
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

41

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
    - "Who is the president of US?"-> person
  - Formulate queries to send to a search engine
    - "president of United States"
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

42

Answer Extraction

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - Who is the prime minister of India (PERSON)
    Manmohan Singh, Prime Minister of India, had told left leaders that the deal would not be renegotiated.
  - How tall is Mt. Everest? (LENGTH)
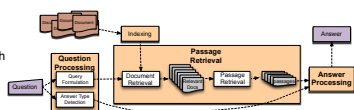   The official height of Mount Everest is **29035 feet**

43

Adding Analysis Patterns

- "Who is Elvis?"
  - Question type: "who"
  - Named-entity tagging: "Who is <person-name> Elvis</person-name>"
  - Analysis pattern: if question type = "who" and question contains <person-name> then
- Desired answer probably is a description
- Likely answer extraction patterns
  - "Elvis, the X", e.g., "Elvis, the *king of rock and roll*!"
  - "the X Elvis", e.g., "the *legendary entertainer* Elvis"

44

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
    - "Who is the president of US?"-> person
  - Formulate queries to send to a search engine
    - "president of United States"
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

45

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?
- Answer Type: **Person**

Passage:

The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

46

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?
- Answer Type: **Person**

Passage:

The **Marie biscuit** is named after **Marie Alexandrovna**, the daughter of **Czar Alexander II of Russia** and wife of **Alfred**, the second son of **Queen Victoria** and **Prince Albert**

47

Use machine learning:
Features for ranking candidate answers

Answer type match: Candidate contains a phrase with the correct answer type.

Pattern match: Regular expression pattern matches the candidate.

Question keywords: # of question keywords in the candidate.

Keyword distance: Distance in words between the candidate and query keywords

Novelty factor: A word in the candidate is not in the query.

Apposition features: The candidate is an appositive to question terms

Punctuation location: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

Sequences of question terms: The length of the longest sequence of question terms that occurs in the candidate answer.

48

Candidate Answer scoring in IBM Watson

- Each candidate answer gets scores from >50 components
  - (from unstructured text, semi-structured text, triple stores)

  - logical form (parse) match between question and candidate
  - passage source reliability
  - geospatial location
    - California is "southwest of Montana"
  - temporal relationships
  - taxonomic classification

49

Information Retrieval (IR)-based QA

- Factoid QA pipeline
- Factoid QA evaluation
- Common Knowledge used in QA
- Recent QA tasks

50

Common Evaluation Metrics

1. *Accuracy* (does answer match gold-labeled answer?)
2. *Mean Reciprocal Rank*
   - For each query return a ranked list of M candidate answers.
   - Query score is 1/Rank of the first correct answer
     - If first answer is correct: 1
     - else if second answer is correct: ½
     - else if third answer is correct: ⅓, etc.
     - Score is 0 if none of the M answers are correct
   - Take the mean over all N queries

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

51

Information Retrieval (IR)-based QA

- Factoid QA pipeline
- Factoid QA evaluation
- Common Knowledge used in QA
- Recent QA tasks

52

Knowledge in QA

- What are other types of knowledge useful for a QA system?
  - Relations
  - Temporal information
  - Dialogue context

53

Relation Extraction

- Answers: Databases of Relations
  - born-in("Emma Goldman", "June 27 1869")
  - author-of("Cao Xue Qin", "Dream of the Red Chamber")
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions
  Whose granddaughter starred in E.T.?
  ```
  (acted-in ?x "E.T.")
     (granddaughter-of ?x ?y)
  ```

54

9

Temporal Reasoning

- Relation databases
  - (and obituaries, biographical dictionaries, etc.)
- IBM Watson
  - "In 1594 he took a job as a tax collector in Andalusia"
  Candidates:
    - Thoreau is a bad answer (born in 1817)
    - Cervantes is possible (was alive in 1594)

55

Context and Conversation in Virtual Assistants like Siri

- Coreference helps resolve ambiguities
  - U: "Book a table at Il Fornaio at 7:00 with **my mom**"
  - U: "Also send **her** an email reminder"
- Clarification questions:
  - U: "Chicago pizza"
  - S: "Did you mean pizza restaurants in Chicago or Chicago-style pizza?"

56

Limitations of Factoid Q/A

- Question must query a specific fact that is explicitly stated somewhere in the document corpus.
- Does not allow aggregating or accumulating information across multiple information sources.
- Does not require "deep compositional" semantics, nor inferential reasoning to generate answer.

57

Information Retrieval (IR)-based QA

- Factoid QA pipeline
- Factoid QA evaluation
- Common Knowledge used in QA
- Recent QA tasks

58

What are recent tasks for QA?

- Reading comprehension (machine reading)

- Visual Question Answering

59

Reading Comprehension Q/A

- Answer questions that test comprehension of a specific document.
- Use standardized tests of reading comprehension to evaluate performance (Hirschman et al. 1999; Rilo & Thelen, 2000; Ng et al. 2000; Charniak et al. 2000).

60

Sample Reading Comprehension Test

School Kids Clean Up Creek!

(EVERETT, WASHINGTON, June, 1988) – It has taken five years of hard work.

But a group of children have cleaned up Pigeon Creek.

In 1985, the creek was so dirty that the fish had all died. A group of kids at Jackson School decided to do something about it. It became a project for their whole class.

First they cleaned out the bottom of the creek, called the creek bed. There were many weeds. They also found lot of trash and litter. Then the students raised fish eggs in a tank of water called an aquarium. When the eggs hatched, they emptied the new fish into the creek. They wanted others to stop making the creek dirty. So they made signs that said "Don't Dump". They put the signs near the creek. They also made little books, called booklets. The booklets told people how to keep the creek clean. Now, the water is clean and new fish are being born in the creek. Pigeon Creek is a sign of the good things young people can do.

Q1: *Who cleaned up the creek?* Q2: *What was the name of the creek?* Q3: *When did the work begin?*
Q4: *Where is the bed of a creek?* Q6: *Why did the kids make booklets?*

61

Large Scale Reading Comprehension Data

- DeepMind's large-scale data for reading comprehension Q/A (Hermann et al., 2015).
  - News articles used as source documents.
  - Questions constructed automatically from article summary sentences.

	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
# queries	380,298	3,924	3,198	879,450	64,835	53,182
Max # entities	527	187	396	371	232	245
Avg # entities	26.4	26.5	24.5	26.5	25.5	26.0
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

62

Sample DeepMind Reading Comprehension Test

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched *ent58* shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " …
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

Table 3: Original and anonymised version of a data point from the Daily Mail validation set. The anonymised entity markers are constantly permuted during training and testing.

63

Deep LSTM Reader

- DeepMind uses LSTM recurrent neural net (RNN) to encode document and query into a vector that is then used to predict the answer.

Incorporated various forms of attention to focus the reader on answering the question while reading the document.
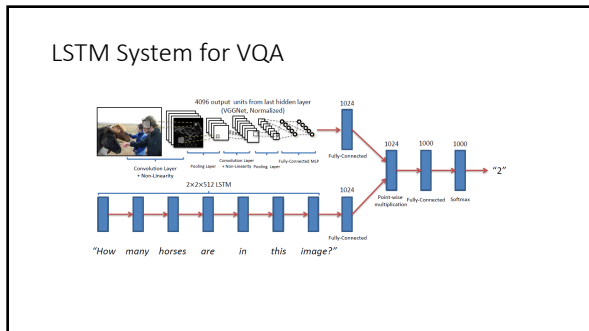
64

Visual Question Answering (VQA)

- Answer natural language questions about information in images.
- VaTech/MSR group has put together VQA dataset with ~750K questions over ~250K images (Antol et al., 2016).

65

VQA Examples

66

11

LSTM System for VQA

67