Logistics

• Project progress discussion meeting with instructor
  • Meetings on March 29 morning will take place in Rm 110, WVH (our classroom).
  • Meetings on March 29 afternoon will take place in Rm 911, 177 Huntington Ave.

Project presentation

• Each team will present for 10 minutes, with 2 minutes for QA.

• After all presentations, we will vote for favorite project. Each team has two votes.

• The team that gets the most votes wins. Each team member will get 1% bonus towards the final grade.

• One winning team is selected on each of April 9, 12, and 16.

Project presentation

• The presentation order will be posted on piazza.

• Please upload your slides on blackboard after presentation.

• Feedback will be sent to the team through blackboard after the presentation.

• Final reports are expected to resolve the issues (hopefully not too many!) raised in the feedback. Due on April 21st 11:59pm.

Presentation and final report

• Problem Description (10%)
  What is the task?
  System input and output
  Examples will be helpful

• Reference/Related work (20%)
  Put your work in context: what has been done before? You need to have reference!
  What’s new in your work?

• Methodology: What you have done (30%)
  Preprocessing of the data
  What are your data? Features used? What are effective, and what are not?
  What methods do you experiment with? And why do you think they’re reasonable and suitable for the task?

• Experiments (40%)
  Datasets size, train/test/development
  Evaluation metrics: what are used and are they proper to calibrate system performance?
  Baseline: what are they?
  Results, tables, figures, etc

Question Answering
Questions in Search

Question Answering (Some Background)

Question: What do worms eat?

Potential Answers:
- Worms eat grass
- Worms eat grass
- Grass is eaten by worms
- Grass is eaten by worms

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

Question Answering: IBM’s Watson

- Won Jeopardy on February 16, 2011!
Apple’s Siri

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?

Commercial systems: mainly factoid questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is the Louvre Museum located?</td>
<td>In Paris, France</td>
</tr>
<tr>
<td>What’s the abbreviation for limited partnership?</td>
<td>L.P.</td>
</tr>
<tr>
<td>What are the names of Odin's ravens?</td>
<td>Huginn and Muninn</td>
</tr>
<tr>
<td>What currency is used in China?</td>
<td>The yuan</td>
</tr>
<tr>
<td>What kind of nuts are used in marzipan?</td>
<td>Almonds</td>
</tr>
<tr>
<td>What instrument does Max Roach play?</td>
<td>Drums</td>
</tr>
</tbody>
</table>

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)

Information Retrieval (IR)-based QA

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

**Jeopardy!:** They’re the two states you could be reentering if you’re crossing Florida’s northern border

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

**Answer Type Detection: Named Entities**

- **Who founded Virgin Airlines?**
Answer Type Detection: Named Entities

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

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...

Part of Li & Roth’s Answer Type Taxonomy

Answer Types

<table>
<thead>
<tr>
<th>ENTITY</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>animal</td>
<td>What are the names of China’s animals?</td>
</tr>
<tr>
<td>body</td>
<td>What part of your body contains the corpus callosum?</td>
</tr>
<tr>
<td>location</td>
<td>Is there a book about the story of Atlantis?</td>
</tr>
<tr>
<td>person</td>
<td>What language is used in China?</td>
</tr>
<tr>
<td>country</td>
<td>What does the term person mean?</td>
</tr>
<tr>
<td>description</td>
<td>What was the first book you read?</td>
</tr>
<tr>
<td>currency</td>
<td>What was Einstein's favorite city?</td>
</tr>
<tr>
<td>date</td>
<td>What is the official language of Algeria?</td>
</tr>
<tr>
<td>food</td>
<td>What is the city of King Arthur’s death?</td>
</tr>
<tr>
<td>drug</td>
<td>What is the nation of King Arthur’s death?</td>
</tr>
<tr>
<td>profession</td>
<td>What is the capital of France?</td>
</tr>
<tr>
<td>religion</td>
<td>What is the capital of France?</td>
</tr>
<tr>
<td>sport</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>technology</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>travel</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>war</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>weather</td>
<td>What is the capital of the United States?</td>
</tr>
</tbody>
</table>

More Answer Types

<table>
<thead>
<tr>
<th>CLASS</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN</td>
<td>What is the telephone number for the University of Colorado?</td>
</tr>
<tr>
<td>country</td>
<td>How are the states of India connected?</td>
</tr>
<tr>
<td>description</td>
<td>How many countries are in the world?</td>
</tr>
<tr>
<td>location</td>
<td>How much land did Montana claim in 1867?</td>
</tr>
<tr>
<td>money</td>
<td>Who was the first person to use a computer?</td>
</tr>
<tr>
<td>object</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>person</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>time</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>date</td>
<td>What is the capital of the United States?</td>
</tr>
<tr>
<td>weight</td>
<td>What is the capital of the United States?</td>
</tr>
</tbody>
</table>
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
country, city, man, film, state, author, group, hero, 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

Answer Type Detection
• Hand-written rules
• Machine Learning

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?

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!

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?

Query Formulation
• QUESTION PROCESSING
  • Detect question type, answer type, focus, relations
  • "Who is the president of the 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
Keyword Selection Algorithm

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

Choosing keywords from the query

Who coined the term “cyberspace” in his novel “Neuromancer”?

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

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - e.g., “Who is the president of USA?” - 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
  - Use evidence from the text and external sources

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

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

Passage Retrieval as Query-focused Summarization

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

(The chaotic development that is engulfing the Amazon rain forest could finally be reined in with a new plan developed by leading scientists from around the world.” “That’s one of the most encouraging news about the Amazon in recent years,” said Thomas Lovejoy, an Amazon specialist. “It contrasts sharply with a year ago, when there was nothing to read about conservation in the Amazon.”

Ninety 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.
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.

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - "Who is the prime minister of India?" -> person
  - Formulate queries to send to a search engine
  - "prime minister of India"
- 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

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
  - "How tall is Mt. Everest?" -> length
- The official height of Mount Everest is 29035 feet.
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”

IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - 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 candidate answers using evidence from the text and external sources

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

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

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.
- 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.
Information Retrieval (IR)-based QA

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

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
     - If second answer is correct: $1/2$
     - If third answer is correct: $1/3$, etc.
     - Score is 0 if none of the M answers are correct
   • Take the mean over all N queries

\[
\text{MRR} = \frac{\sum_{i=1}^{\infty} \frac{1}{\text{rank}_i}}{N}
\]

Knowledge in QA

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

Relation Extraction

• **Answers: Databases of Relations**
  - born-in("Emma Goldman", "June 27 1869")
  - author-off("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)

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)
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?”

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.

Information Retrieval (IR)-based QA

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

What are recent tasks for QA?

- Reading comprehension
- Visual Question Answering

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; Riso & Thelen, 2000; Ng et al. 2000; Charniak et al. 2000).

Sample Reading Comprehension Test

- Multiple choice questions based on a short text.
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.

<table>
<thead>
<tr>
<th>CNN</th>
<th>DailyMail</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>valid</td>
</tr>
<tr>
<td># articles</td>
<td>0.269</td>
</tr>
<tr>
<td># questions</td>
<td>300.995</td>
</tr>
<tr>
<td>Male-Female</td>
<td>217</td>
</tr>
<tr>
<td>Real-False</td>
<td>764</td>
</tr>
<tr>
<td>True-False</td>
<td>151</td>
</tr>
<tr>
<td>Train set</td>
<td>108,447</td>
</tr>
</tbody>
</table>

Sample DeepMind Reading Comprehension Test

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).

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.

LSTM System for VQA

VQA Examples