CS 6120/CS4120: Natural Language Processing
Instructor: Prof. Lu Wang
College of Computer and Information Science
Northeastern University
Webpage: www.ccs.neu.edu/home/luwang

Project presentation
• Each team will present for 8 minutes, with 2 minutes for QA.
• After all representations, 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 final grade.

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 raised in the feedback. Due on Dec 10th, 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?
  Baselines: what are they?
  Results, tables, figures, etc

Question Answering

IR-based Question Answering
**IR-based Question Answering**

**Question Answering**

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 15:30-204

**Question Answering**

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 15:30-204

**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 QA

- Information Retrieval (IR)-based approaches
  - IBM Watson (some parts); Google
- Knowledge-based and Hybrid approaches
  - IBM Watson; Apple Siri; Wolfram Alpha
- Data-driven, neural network-based approaches

IR-based Factoid QA
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

Knowledge-based approaches (Siri)

- Build a semantic representation of the query
- Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
- Geospatial databases
- Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
- Restaurant review sources and reservation services
- Scientific databases

Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
- Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

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**
  - 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
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)
Answer Type Detection: Named Entities

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

Answer Type Taxonomy

- 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

<table>
<thead>
<tr>
<th>ENTITY</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>nation</td>
<td>What is the name of China’s capital?</td>
</tr>
<tr>
<td>body</td>
<td>What part of (your body) contains the organs collagen?</td>
</tr>
<tr>
<td>sequence</td>
<td>Is what book from? Is the story of Alice?</td>
</tr>
<tr>
<td>entity</td>
<td>What entity is used in China?</td>
</tr>
<tr>
<td>description</td>
<td>What does bulk cancer prevent?</td>
</tr>
<tr>
<td>group</td>
<td>What was the word of Cybergues?</td>
</tr>
<tr>
<td>instrument</td>
<td>What instrument does Black Beauty play?</td>
</tr>
<tr>
<td>language</td>
<td>What is the official language of Algeria?</td>
</tr>
<tr>
<td>index</td>
<td>What letter appears in the title of Trivia?</td>
</tr>
<tr>
<td>other</td>
<td>What is the name of King Arthur’s sword?</td>
</tr>
<tr>
<td>product</td>
<td>What is the latest computer?</td>
</tr>
<tr>
<td>religion</td>
<td>What religion did the priest teach?</td>
</tr>
<tr>
<td>sport</td>
<td>What was the name of the ball game played by the Mayans?</td>
</tr>
<tr>
<td>appearance</td>
<td>What had his appearance?</td>
</tr>
<tr>
<td>name</td>
<td>What is the name of the English River?</td>
</tr>
<tr>
<td>technique</td>
<td>What is the best way to answer a question?</td>
</tr>
<tr>
<td>time</td>
<td>How do you say “Children’sチェック”?</td>
</tr>
<tr>
<td>website</td>
<td>What was the name of England’s ship?</td>
</tr>
<tr>
<td>word</td>
<td>What is the nature of the day?</td>
</tr>
</tbody>
</table>

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING’02
More Answer Types

<table>
<thead>
<tr>
<th>Question</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2500 answer types in 20,000 Jeopardy question sample</td>
<td></td>
</tr>
<tr>
<td>The most frequent 200 answer types cover ~ 50% of data</td>
<td></td>
</tr>
<tr>
<td>The 40 most frequent Jeopardy answer types</td>
<td></td>
</tr>
<tr>
<td>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</td>
<td></td>
</tr>
</tbody>
</table>

Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

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

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

Factoid Q/A

Passage Retrieval and Answer Extraction

Passage Retrieval

• Step 1: IR engine retrieves documents using query terms
• Step 2: Segment the documents into shorter units
  • something like paragraphs
• Step 3: Passage ranking
  • Use answer type to help rerank passages
Features for Passage Ranking

- 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

- 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.

Factoid Q/A

- 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**.
The noun phrase filter

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"

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

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.

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
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 \( 1/\text{Rank of the first correct answer} \)
     - If first answer is correct: 1
     - else if second answer is correct: \( \frac{1}{2} \), etc.
     - Score is 0 if none of the \( M \) answers are correct
   - Take the mean over all \( N \) queries

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

Knowledge in QA

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.?*
  - \( \text{acted-in} \ ?x \ “E.T.” \)
  - \( \text{granddaughter-of} \ ?x \ ?y \)

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.
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).

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.

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.

Sample Reading Comprehension Test

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).
VQA Examples

LSTM System for VQA