Outline

• Introduction

• History

• QA Architecture
Outline

- Introduction
- History
- QA Architecture
Motivation

• Finding small segments of text which answer users’ questions
Motivation

How many presidents do the USA have?

List of Presidents of the United States by occupation...

How many presidents have we had in the United States?

How many presidents has the United States had?

6.3.1. The Presidents by Number and Dates - Comcast.net

More results from answers.yahoo.com
Motivation

- How many presidents do the USA have?
  - 44
  - List of names in chronological order

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>Ronald Reagan</td>
<td>Actor &amp; broadcaster, President of the Screen Actors Guild</td>
</tr>
<tr>
<td>41</td>
<td>George H. W. Bush</td>
<td>Pilot (Navy Lieutenant, Junior Grade), Businessman (Oil)</td>
</tr>
<tr>
<td>42</td>
<td>Bill Clinton</td>
<td>Lawyer, Law lecturer</td>
</tr>
<tr>
<td>43</td>
<td>George W. Bush</td>
<td>Businessman (Oil, baseball)</td>
</tr>
<tr>
<td>44</td>
<td>Barack Obama</td>
<td>Public Interest professional, Lawyer, Constitutional Law Professor</td>
</tr>
</tbody>
</table>

(http://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_occupation)
Motivation

- Information retrieval
  - Keywords (short input)
  - Document (long output)

- Question Answering
  - Natural language question (long input)
  - Short answer (short input)
QA Types

• Closed-domain
  – Answering questions from a specific domain

• Open-domain
  – Answering any domain independent question
Outline

• Introduction

• History

• QA Architecture
History

• BASEBALL [Green et al., 1963]
  - One of the earliest question answering systems
  - Developed to answer users’ questions about dates, locations, and the results of baseball matches

• LUNAR [Woods, 1977]
  - Developed to answer natural language questions about the geological analysis of rocks returned by the Apollo moon missions
  - Able to answer 90% of questions in its domain posed by people not trained on the system
History

- **STUDENT**
  - Built to answer high-school students’ questions about algebraic exercises

- **PHLIQA**
  - Developed to answer the user’s questions about European computer systems

- **UC (Unix Consultant)**
  - Answered questions about the Unix operating system

- **LILOG**
  - Was able to answer questions about tourism information of cities in Germany
Closed-domain QA

- Closed-domain systems

- Extracting answers from structured data (database)
  (Labor-intensive to build)

- Converting natural language questions to database queries
  (easy to implement)
Open-domain QA

- Closed-domain QA ⇒ Open-domain QA
  - Using a large collection of unstructured data (e.g., the Web) instead of databases
- Advantages
  - Many subjects are covered
  - Information is constantly added and updated
  - No manual work is required to build databases
- Problems
  - Information is not always up-to-date
  - Wrong information cannot be avoided
  - Much irrelevant information is found
Open-domain QA

- START [Katz, 1997]
  - Utilized a knowledge-base to answer the user’s questions
  - The knowledge-base was first created automatically from unstructured Internet data
  - Then it was used to answer natural language questions

http://start.csail.mit.edu/index.php
Open-domain QA

United States

Executive branch:
chief of state: President Barack H. OBAMA (since 20 January 2009); Vice President Joseph R. BIDEN (since 20 January 2009); note - the president is both the chief of state and head of government
head of government: President Barack H. OBAMA (since 20 January 2009); Vice President Joseph R. BIDEN (since 20 January 2009)
cabinet: Cabinet appointed by the president with Senate approval

(For more information visit the World Leaders website)
elections: president and vice president elected on the same ticket by a college of representatives who are elected directly from each state; president and vice president serve four-year terms (eligible for a second term); election last held 6 November 2012 (next to be held on 8 November 2016)
election results: Barack H. OBAMA reelected president; percent of popular vote - Barack H. OBAMA 50.6%, Mitt ROMNEY 47.9%, other 1.5%:
IBM Watson

- Playing against two greatest champions of Jeopardy
- Challenges
  - Knowledge
  - Speed
  - Confidence
- Final Jeopardy, Stephen Backer

https://www.youtube.com/watch?v=WFR3IOm_xhE
Outline

- Introduction
- History
- QA Architecture
How many presidents do the USA have?

44
Question Analysis

- Named Entity Recognition
- Surface Text Pattern Learning
- Syntactic Parsing
- Semantic Role Labeling
Question Analysis: Named Entity Recognition

- Recognizing the named entities in the text to extract the target of the question
- Using the question’s target in the query construction step

- Example:
  - Question: “In what country was Albert Einstein born?”
  - Target: “Albert Einstein”
Question Analysis: Pattern Learning

- Extracting a pattern from the question
- Matching the pattern with a list of pre-defined question patterns
- Finding the corresponding answer pattern
- Realizing the position of the answer in the sentence in the answer extraction step
- Example:
  - Question: “In what country was Albert Einstein born?”
  - Question Pattern: “In what country was X born?”
  - Answer Pattern: “X was born in Y.”
Question Analysis: Syntactic Parsing

- Using a dependency parser to extract the syntactic relations between question terms
- Using the dependency relation paths between question terms to extract the correct answer in the answer extraction step

(http://nlp.stanford.edu:8080/corenlp/process)
Question Analysis: Semantic Role Labeling

- FrameNet: a lexical database for English
- More than 170,000 manually annotated sentences
- Frame Semantics: describes the type of event, relation, or entity and the participants in it.

- Example:
  - John\textsubscript{[Cook]} grills a fish\textsubscript{[Food]} on an open fire\textsubscript{[Heating-Instrument]}
Question Analysis: Semantic Role Labeling

- Frame assignment
- Role labeling

- Jim[Driver] flew his plane[Vehicle] to Texas[Goal]  
  OPERATE-VEHICLE

- Alice[Destroyer] destroys the item[Undergoer] with a plane[Instrument]  
  DESTROYING
Question Analysis: Semantic Role Labeling

- Finding the question’s head verb
- Example:
  - “Who\_\text{[Buyer]} purchased YouTube\_\text{[Goods]}?”
- COMMERCE−BUY
  - Buyer [Subj,NP] verb Goods [Obj,NP]
  - Buyer [Subj,NP] verb Goods [Obj,NP] Seller [Dep,PP-from]
  - Goods [Subj,NP] verb Buyer [Dep,PP-by]
  - ...
- Example:
  - “In 2006, YouTube\_\text{[Goods]} was purchased by Google\_\text{[Buyer]} for $1.65 billion.”
Question Classification

• Classifying the input question into a set of question types
• Mapping question types to the available named entity labels
• Finding strings that have the same type as the input question in the answer extraction step

• Example:
  – Question: “In what country was Albert Einstein born?”
  – Type: LOCATION - Country
Question Classification

• Example (NER):
  - S1: “Albert Einstein was born in 14 March 1879.”
  - S2: “Albert Einstein was born in Germany.”
  - S3: “Albert Einstein was born in a Jewish family.”
Question Classification

- Classification taxonomies
  - BBN
  - Pasca & Harabagiu
  - Li & Roth
    - 6 coarse- and 50 fine-grained classes
      - ABBREVIATION
      - ENTITY
      - DESCRIPTION
      - HUMAN
      - LOCATION
      - NUMERIC
Question Classification

\(HUMAN\)
- group
- ind
- title
- description

\(LOCATION\)
- city
- country
- mountain
- other
- state

human beings
a group or organization of persons
an individual
title of a person
description of a person
locations
cities
countries
mountains
other locations
states

(http://cogcomp.cs.illinois.edu/Data/QA/QC/definition.html)
## Question Classification

<table>
<thead>
<tr>
<th>Question</th>
<th>Type</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>How did serfdom develop in and then leave Russia?</td>
<td>DESC</td>
<td>manner</td>
</tr>
<tr>
<td>What films featured the character Popeye Doyle?</td>
<td>ENTY</td>
<td>cremat</td>
</tr>
<tr>
<td>What fowl grabs the spotlight after the Chinese Year of the Monkey?</td>
<td>ENTY</td>
<td>animal</td>
</tr>
<tr>
<td>What is the full form of .com?</td>
<td>ABBR</td>
<td>exp</td>
</tr>
<tr>
<td>What team did baseball 's St. Louis Browns become?</td>
<td>HUM</td>
<td>gr</td>
</tr>
</tbody>
</table>

(http://cogcomp.cs.illinois.edu/Data/QA/QC/train_1000.label)
Question Classification

Coarse Classifier

ABBREVIATION
ENTITY
DESCRIPTION
HUMAN
LOCATION
NUMERIC

Fine Classifier

group
ind
title
description
Question Classification

• Using any kinds of supervised classifiers
  – K Nearest Neighbor
  – Support Vector Machines
  – Naïve Bayes
  – Maximum Entropy
  – Logistic Regression
  – ...

• Benefiting from available toolkits: Weka, SVM-light, etc.

• Considering the confidence measure of the classification to filter the result
Query Construction

• Goal:
  – Formulating a query with a high chance of retrieving relevant documents

• Task:
  – Assigning a higher weight to the question’s target
  – Using query expansion techniques to expand the query
Document Retrieval

• Importance:
  - QA components use computationally intensive algorithms
  - Time complexity of the system strongly depends on the size of the to be processed corpus

• Task:
  - Reducing the search space for the subsequent components
  - Retrieving relevant documents from a large corpus
  - Selecting top n retrieved document for the next steps
Document Retrieval

• Using available information retrieval models
  – Vector Space Model
  – Probabilistic Model
  – Language Model

• Using available information retrieval toolkits: Lucene, Lemur, etc.
Sentence Retrieval

- Task:
  - Finding small segments of text that contain the answer

- Benefits beyond document retrieval:
  - Documents are very large
  - Documents span different subject areas
  - The relevant information is expressed locally
  - Retrieving sentences simplifies the answer extraction step

- Main problem: sentence brevity
Sentence Retrieval

- Information retrieval models for sentence retrieval
  - Vector Space Model
  - Probabilistic Model
  - Language Model
    - Jelinek-Mercer Linear Interpolation
    - Bayesian Smoothing with Dirichlet Prior
    - Absolute Discounting
Sentence Retrieval

• Approaches to overcome the sentence brevity problem:
  - Term relationship models
  - Translation model
  - Term clustering model
Sentence Annotation

- Similar to Question Analysis
- Annotating relevant sentences using linguistic analysis
  - Named entity recognition
  - Syntactic parsing
  - Semantic role labeling
- Example (NER):
  - S1: “Albert Einstein was born in 14 March 1879.”
  - S2: “Albert Einstein was born in Germany.”
  - S3: “Albert Einstein was born in a Jewish family.”
Answer Extraction

- Extracting candidate answers based on various information
  - Question
    - Question Analysis: patterns
    - Question Analysis: syntactic parse
    - Question Analysis: semantic roles
    - Question Classification: question type
  - Sentence
    - Sentence Annotation: all annotated data
Answer Extraction

- Using extracted patterns
- Example:
  - Question: “In what country was Albert Einstein born?”

- Question Pattern: In what country was X born?
- Answer Pattern: X was born in Y.

- Example (Pattern):
  - S1: “Albert Einstein was born in 14 March 1879.”
  - S2: “Albert Einstein was born in Germany.”
  - S3: “Albert Einstein was born in a Jewish family.”
Answer Extraction

- Using question type and entity type
- Example:
  - Question: “In what country was Albert Einstein born?”
  - Question Type: LOCATION - Country
- Example (NER):
  - S1: “Albert Einstein was born in 14 March 1879.”
  - S2: “Albert Einstein was born in Germany.”
  - S3: “Albert Einstein was born in a Jewish family.”
Answer Extraction

- Using syntactic parsing
  - Different wordings possible, but similar syntactic structure

Q: Who founded the Black Panthers organization?

S1: Bobby Seale, a student at Merritt College, founded the Black Panther Party for self-defense.

S2: The Black Panther Party, co-founded by Seale and Newton, flourished...

(based on the work by Dan Shen)
Answer Extraction

- Using syntactic parsing
  - Many syntactic variations → need robust matching approach

Q: Who **founded** the Black Panthers organization?

S1: Bobby Seale, a student at Merritt College, **founded the Black Panther Party** for self-defense.

S2: The Black Panther Party, **co-founded by Seale and Newton**, flourished...

S3: Hilliard introduced Bobby Seale, who **co-founded the Black Panther Party** here.

S4: Black Panthers **Co-founder Bobby Seale** visits UMM.

(based on the work by Dan Shen)
Answer Extraction

• Using semantic roles

• Example:
  – “Who\textsubscript{Buyer} purchased YouTube? \textsubscript{Goods}”

• Example:
  – “In 2006, YouTube\textsubscript{Goods} was purchased by Google\textsubscript{Buyer} for $1.65 billion.”
Answer Validation

• Using Web as a knowledge resource for validating answers

• Required steps
  - Query creation
  - Answer rating
Answer Validation

• Query creation
  – Combining the answer with a subset of the question keywords
  – Choosing different combinations of subsets
    • Bag-of-Word
    • Noun phrase chunks
    • Declarative form
Answer Validation

• Answer rating
  - Passing the query to a search engine
  - Analyzing the result of the search engine
    • Counting the results
    • Parsing the result snippets
  - Other possibilities:
    • Using knowledge bases to find relations between the question keywords and the answer
How many presidents do the USA have?

44
Further Reading

- Speech and Language Processing
  - Chapters 23.1, 23.2