Natural Language Question-Answering with Visualizations

Bianca Yu, Hannah DeBalsi

CS 294W Spr 2020
Outline

Visualization + Natural Language

The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants

Backend Implementation
NLP Query to Database Query
From Answer to Display
Visualization + Natural Language

The Purpose of Visualization

Natural Language Interfaces for Visualization

Application in Conversational Virtual Assistants

Backend Implementation

NLP Query to Database Query

From Answer to Display
What is visualization?

- “Transformation of the symbolic into the geometric” (McCormick et al. 1987)
What is visualization?

- “Transformation of the symbolic into the geometric”
  (McCormick et al. 1987)

- “... finding the artificial memory that best supports our natural means of perception.”
  (Bertin 1967)
What is **visualization**?

- “Transformation of the symbolic into the geometric”
  (McCormick et al. 1987)

- “… finding the artificial memory that best supports our natural means of perception.”
  (Bertin 1967)

- “The use of computer-generated, interactive, visual representations of data to **amplify cognition**.”
  (Card, Mackinlay, and Shneiderman 1999)
We use visualization to ...

- **Record** information
We use visualization to ...

- **Record** information
We use visualization to ...

- **Record** information
We use visualization to ...

- **Record** information
- **Analyze** information
The purpose of visualization

We use visualization to ...

- **Record** information
- **Analyze** information
  - See data in context
  - Make a decision
See data in context

- Example: Cholera outbreak, 1854
See data in context

- Example: Cholera outbreak, 1854
Make a decision

- Example: Challenger space shuttle launch, 1986
Make a decision

- Example: Challenger space shuttle launch, 1986

Tufte. Visual and Statistical Thinking 1997
Make a decision

- Example: Challenger space shuttle launch, 1986
We use visualization to ...

- **Record** information

- **Analyze** information
  - See data in context
  - Make a decision
We use visualization to ...

- **Record** information

- **Analyze** information
  - See data in context
  - Make a decision

- **Convey** information
We use visualization to ...

- **Record** information
- **Analyze** information
  - See data in context
  - Make a decision
- **Convey** information
Graphs in statistical analysis

“Graphs can have various purposes, such as:

(i) to help us perceive and appreciate some broad features of the data,

(ii) to let us look behind those broad features and see what else is there.”

(Anscombe 1973)
The Purpose of Visualization

Natural Language Interfaces for Visualization

Application in Conversational Virtual Assistants

Backend Implementation

NLP Query to Database Query

From Answer to Display
Visualization + Natural Language

The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants

Backend Implementation
NLP Query to Database Query
From Answer to Display
Why implement natural language interaction?
Why implement natural language interaction?

● “Natural language interaction allows users to **ask questions directly** in complex programs **without having to learn how to use an interface.**” (Gao et al.)

● Users of sophisticated visual analytic tools are “... usually domain experts with **marginal knowledge of visualization techniques.**” (Sun et al.)
Types of current natural language interfaces

1. Those that answer questions about existing visualizations
2. Those that create a new visualization
1) Answering questions about existing visualizations

**How Common Is Religious Extremism?**

<table>
<thead>
<tr>
<th>Religion</th>
<th>Common</th>
<th>Not Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muslims</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pagan/earth-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protestants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jews</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffiliated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catholics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mormons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orthodox Christian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buddhists</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Q1:** What is the percentage of response ‘Common’ for Catholics?

A(Sempre): **92**
A(Ours): **84**. I looked up the length of the orange bar for ‘Catholics’.

**Q2:** Which religion has the longest orange component?

A(Sempre): **Hindus**
A(Ours): **Muslims**. I looked up ‘Religion’ of the longest orange bar.

**Q3:** What does the blue field represent?

A(Sempre): **24**
A(Ours): **Not Common**. I looked up what blue represents by looking at the legend.
2) Creating new visualizations
2) Creating new visualizations

- Commercial
  - IBM
  - Microsoft
  - Wolfram Alpha
2) Creating new visualizations

- Commercial
  - IBM
  - Microsoft
  - Wolfram Alpha

- Research Projects
  - Articulate
  - DataTone
Outline

Visualization + Natural Language

The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants

Backend Implementation
NLP Query to Database Query
From Answer to Display
Outline

Visualization + Natural Language

The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants

Backend Implementation
NLP Query to Database Query
From Answer to Display
Motivation

- Current tools created for data analysts, not the general, curious public

- "Amplify cognition" (Card, Mackinlay, and Shneiderman 1999)

- Provide context to numerical responses to increase comprehension

- Encourage curiosity and “see what else is there” (Anscombe 1973)
Challenges

● Ambiguity
  ○ What is the user asking for specifically?

● Inferring **when** to include a chart in the response
  ○ When does a user benefit from viewing a chart?

● Determining **what** to display
  ○ What kind of additional data should be displayed?
  ○ What kind of chart is most effective?

● CUI vs GUI
Question #1

Besides standard graphs, what other types of visualizations do you think would be helpful to integrate into a virtual assistant?

For what domain(s)?
Visualization + Natural Language

The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants

Backend Implementation
NLP Query to Database Query
From Answer to Display
The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants

Backend Implementation

NLP Query to Database Query
From Answer to Display
Visualization + Natural Language

The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants
Backend Implementation
NLP Query to Database Query
From Answer to Display
DataTone System Architecture
Client Side: Web-based interface that operates in standard web browsers
Client Side Example

Data Overview:
- Athlete: Michael Phelps, Natalie Coughlin, Sun Yang, ...
- Age: 15-61
- Country: United States, Australia, Russia, ...
- Year: 2000-2012
- Sport: Swimming, Diving, Cycling, ...
- Gold Medal: 0-8
- Silver Medal: 0-3
- Bronze Medal: 0-3
- Total Medal: 1-8

Sample Queries:
- Olympic Athletes
- Data Overview:
- Chart Templates
- Color
  - Color by Country
  - Color by Sport
  - Single Color
- Group Order
  - Group by Sport then by Country
  - Group by Country then by Sport

Show me medals for hockey and skating by country
- NLP Query to Database Query
Server Side: handles translation of user input to a visualization
Query Analyzer

NLP Query to Database Query

Slide content from https://dl.acm.org/doi/pdf/10.1145/2807442.2807478
Query Analyzer
Tokenization

- Identify low-level language features (words and phrases) that have meaning within the context of the dataset and analysis tasks
  - Example: words that identify column names

1. Construct set of possible phrases
   - Extract all n-grams, ranging from 1 (single words) to k, the sentence length
   - Example: This is a sentence. => {this, is, a, sentence, this is, is a, a sentence, this is a, is a sentence, this is a sentence}

2. Identify n-grams with relevance to dataset/query
   - comparing each n-gram to a set of regular expressions and a lexicon consisting of general phrases
   - tag each matched n-gram with one of eight category labels
Category Labels

1. database attributes (i.e., column names)
2. database cell values
3. numerical values
4. time expressions
5. data operators and functions (greater than, less than, equal, sum, average, sort)
6. visualization key phrases (trend, correlation, relationship, distribution, time series, bars, stacked bars, line graph)
7. boolean operators (e.g., and, or),
8. “direct manipulation” terms (e.g., color)
Example

Query: What is the relationship between unemployment and family income for those families earning more than 20000 and less than 150000 between 2007 and 2010 for California and Michigan?

1. Break into N-grams
2. Identify relevant N-grams by matching to categories
Example

Query: What is the relationship between unemployment and family income for those families earning more than 20000 and less than 150000 between 2007 and 2010 for California and Michigan?

1. Break into N-grams
2. Identify relevant N-grams by matching to categories

Numerical Values
Example

Query: What is the relationship between unemployment and family income for those families earning more than $20000 and less than $150000 between 2007 and 2010 for California and Michigan?

1. Break into N-grams
2. Identify relevant N-grams by matching to categories
Example

Query: What is the relationship between unemployment and family income for those families earning more than 20000 and less than 150000 between 2007 and 2010 for California and Michigan?

1. Break into N-grams
2. Identify relevant N-grams by matching to categories
Example

Query: What is the relationship between unemployment and family income for those families earning more than 20000 and less than 150000 between 2007 and 2010 for California and Michigan?

1. Break into N-grams
2. Identify relevant N-grams by matching to categories
Example

Query: What is the relationship between unemployment and family income for those families earning more than 20000 and less than 150000 between 2007 and 2010 for California and Michigan?

1. Break into N-grams
2. Identify relevant N-grams by matching to categories
Example

Query: What is the relationship between unemployment and family income for those families earning more than 20000 and less than 150000 between 2007 and 2010 for California and Michigan?

1. Break into N-grams
2. Identify relevant N-grams by matching to categories
Relation Identification

- We now have a set of tokens with category tags
- We need to define relationships between these tokens in order to construct a query
Relation Identification Example

“Show me the states that had total sales greater than 20000.”

Stanford Core NLP Parser

Manually constructed set of patterns

- “total sales”: noun phrase
- “greater than 20000”: adjective phrase
  NP and ADJP are siblings of a sentence

Apply:
- SUM to Sales
- the operator “>” to 20000
- generate a filter \( \text{SUM(Sales)} > 20000 \).
Natural Language Parse ⇒ Data Specification (DSP)

- DSPs contain:
  - **Attributes**: all column names in the original query (ie unemployment, family income)
  - **Values**: all strings, numbers, times (ie California, Michigan, 20000, 150000)
  - **Filters**: as explained in the relation identification
  - **Aggregates**: “Show me average medal count by country per year” → AVG(MedalCount).
  - **Order**: “show me the sorted medal count by country from largest to smallest” → orderBy(MedalCount, DESC)

- Generate one database query for each DSP
Outline

Visualization + Natural Language

The Purpose of Visualization
Natural Language Interfaces for Visualization
Application in Conversational Virtual Assistants
Backend Implementation
NLP Query to Database Query
From Answer to Display
Query Analyzer
Visual Specification (VSP)

- Template for a graph
- Each template has constraints on how parameters can be filled
  - supported dimension and data types (categorical, quantitative, or time) for each parameter in the graph
- Map each DSP to the VSP template that can accept that specific DSP’s configuration
- Bar Chart VSP:
  - x-axis: one categorical dimension
  - y-axis: one quantitative measure
  - color: a color encoding (mapping) of one dimension (optional)
- Given a DSP, there are may be several possible templates
VSP → Client → D3.js → Image
What are some of the category labels for n-grams in this string? What type of graph would you use to represent the answer?

*What were the trends of COVID-19 deaths in May between New York and California?*

1. database attributes (i.e., column names)
2. database cell values
3. numerical values
4. time expressions
5. data operators and functions (greater than, less than, equal, sum, average, sort)
6. visualization key phrases (trend, correlation, relationship, distribution, time series, bars, stacked bars, line graph),
7. boolean operators (e.g., and, or)
8. “direct manipulation” terms (e.g., color)