Carnegie Mellon University
Master of Computational
Data Science

Carnegie Mellon University
Language Technologies Institute

SayHear

Bhavya Karki, Fan Hu

Advisors: Anthony Tomasic, Matthias Grabmair

In collaboration with

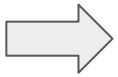
Lucile Callebert, Nithin Haridas, Suhail Barot, Zihua Liu

11-632 (Fall 2018)
MCDS Capstone Course

Introduction

Main Objective : Answer questions from tables collected on the web.

Question: Alexa, when is Maroon 5 coming to San Antonio for a concert?



SQL: SELECT “Date” FROM Maroon_5 WHERE “City” ~ “San Antonio”

Date	Location	City
12-May	Estadio Akron	Guadalajara, Mexico
30-May	Tacoma Dome	Tacoma, WA
1-Jun	Oracle Arena	Oakland, CA
2-Jun	Golden 1 Center	Sacramento, CA
4-Jun	The Forum	Inglewood, CA
5-Jun	The Forum	Inglewood, CA
7-Jun	Talking Stick Resort Arena	Phoenix, AZ
9-Jun	American Airlines Arena	Dallas, TX
10-Jun	Toyota Center	Houston, TX
12-Jun	AT&T Center	San Antonio, TX
14-Jun	Smoothie King Center	New Orleans, LA
16-Jun	Amalie Arena	Tampa, FL

Table: Maroon_5

- Question answering from structured data
- Convert natural language to SQL query

Hypothesis

- This system is constructed with the assumption that the tables contain information sufficient to answer the question.

- The framework is constructed in such a way that a sequence of steps will lead to identifying the location of the answer in the table.

- The underlying confidence is that it is possible to develop a question answering system for structured data and that it is possible to retrieve the relevant information using SQL query for a given question.

- The proposed framework establishes a baseline for the dataset which is the first of its kind

Development Goals

- **Vision** - Make the web more accessible and allow users to ask questions that could be answered using tabular information online.
- **Intended Users** - any user who inputs a natural language query.
- **System Requirements** - Framework must be robust, complete and able to precisely fetch the answer with a very low latency.

Fall Semester Scope

1. Collect more data
2. Dataset analysis
3. Data processing
4. Feature engineering
5. Model designing and training
6. Error analysis

Data Collection

- **Question collection**
  - 5 most recent questions each of AMT workers searched on web
  - 5 triples composed of (i) a natural language question, (ii) the URL of a page that contained the answer and (iii) the answer

- **Collecting tables and SQL queries**

  For each new question :
  - Search the web for a page with tabular data containing the answer
  - Extract the relevant table from that page using import.io and Smartwrap
  - Write and execute a SQL query to extract the answer from the table

- **Introduction of '~' operator**
  - Facilitates matching of terms occurring in the question and those in the table

Data

A total of 302 question-table pairs.

- 238 (78.8%) training data,
  64 (21.2%) testing data.
- 174 (57.6%) entity-instance type table,
  128 (42.4%) key-value type table.

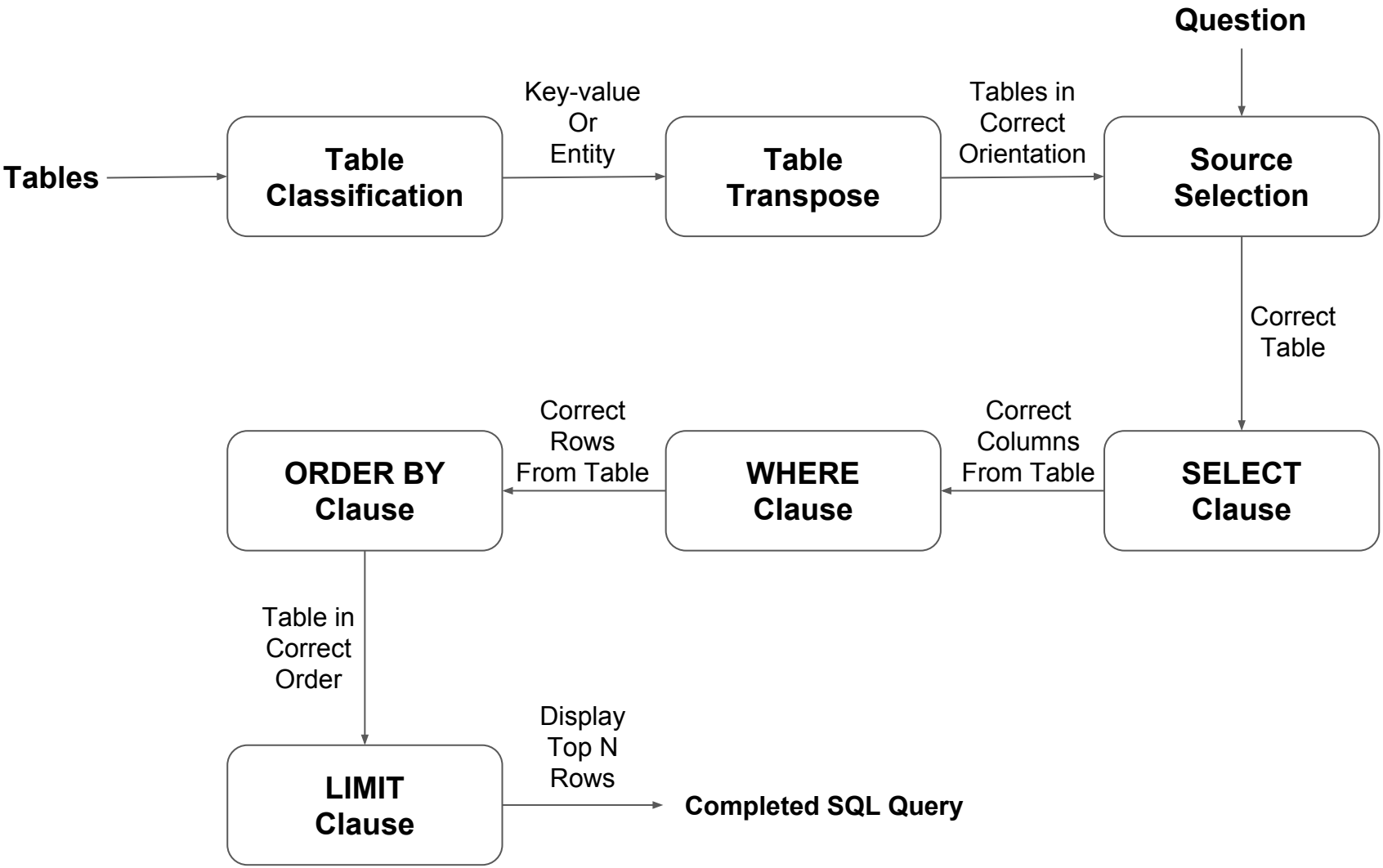
	Presidency[a]	President		Prior office[b]	Party[c]	Term[d]	Vice President
1	April 30, 1789 [e] – March 4, 1797		**George Washington** 1732–1799 (Lived: 67 years) [3][4][5]	Commander-in-Chief of the Continental Army (1775–1783)	*Unaffiliated* [2]	(1788–89) 1 (1789) – (1792) 2 (1793)	John Adams [f][g]
2	March 4, 1797 – March 4, 1801		**John Adams** 1735–1826 (Lived: 90 years) [6][7][8]	1st Vice President of the United States	Federalist	(1796) 3 (1797)	Thomas Jefferson [h]

Entity-instance type

Personal details	
Born	Donald John Trump June 14, 1946 (age 71) New York City
Political party	Republican (1987–99, 2009–11, 2012–present)
Other political affiliations	Democratic (until 1987, 2001–09) Reform (1999–2001) Independent (2011–12)
Spouse(s)	Ivana Zelníčková (m. 1977; div. 1992) Marla Maples (m. 1993; div. 1999) Melania Knauss (m. 2005)
Children	Donald Jr. · Ivanka · Eric · Tiffany · Barron
Parents	Fred · Mary Anne
Relatives	Trump family
Residence	White House (official/primary) Trump Natl. Bedminster (summer) Mar-a-Lago (winter) Trump Tower (private/secondary)
Alma mater	The Wharton School (BS in Econ.)
Occupation	Real estate developer (The Trump Organization) Television host/producer (*The Apprentice*)
Net worth	US$3.1 billion (March 2018)

Key-Value type

System Design



Experimental Evaluation

- **Table type recognition**
  - **Problem** - Machine Learning Classification - KNN, Decision Tree, Logistic Regression
  - **Evaluation Metrics** - Accuracy

- **Source selection**
  - **Problem** - Information Retrieval
  - **Evaluation Metrics -** Precision @ k

- **Column projection - SELECT Clause**
  - **Problem** - Machine Learning Classification - MLP
  - **Evaluation Metrics -** Confusion matrix, accuracy, precision, recall

- **Row filtering - WHERE Clause**
  - **Problem** - Machine Learning Classification - MLP
  - **Evaluation Metrics -** Confusion matrix, accuracy, precision, recall

Table Classification - Model

ML classification problem:

- Given a table, predict if the table is entity-instance type or key-value type

Features used:

- **The number of columns after removing URL columns**
- **The presence of “key” or “property” in the column headers**
- **Variation of cell length** (normalized)
- **Variation of presence of digits** (normalized)

Table Classification - Result

	Train Dataset	Test Dataset
Logistic Regression	97.9%	100.0%
Decision Tree	98.3%	100.0%
KNN	98.3%	100.0%

Table Transpose

Transpose all the key-value tables to entity-instance tables

Question: When is Donald Trump's birthday?

Original Table

Key	Value
spouse	melania trump (m. 2005), marla maples (m. 1993–1999), ivana trump (m. 1977–1992)
born	june 14, 1946 (age 71 years), jamaica hospital medical center, new york city, ny
height	6′ 3″
net worth	3.1 billion usd (2018)
education	wharton school of the university of pennsylvania (1966–1968) , more

Transposed Table

spouse	born	height	net worth	education
melania trump (m. 2005), marla maples (m. 1993–1999), ivana trump (m. 1977–1992)	june 14, 1946 (age 71 years), jamaica hospital medical center, new york city, ny	6′ 3″	3.1 billion usd (2018)	wharton school of the university of pennsylvania (1966–1968) , more

Source Selection - Method

Information Retrieval Approach:

1. Pre-process the given question and all tables into bag-of-words (BoW), including stop-word removal, stemming etc.

2. Calculate TF-IDF of word stems in the given question and tables.

3. Measure similarity between the given question and each table, and then select the best matched table.

Machine Learning Approach:

1. Convert the given question and all tables to vectors using pre-trained Doc2Vec model.

2. Calculate the cosine similarity between the given question and each table, and then select the best matched table.

Source Selection - Result

Information Retrieval approach provides a better performance (~76%) compared to Machine Learning approach (~56%) for this sub-problem.

	Train Dataset			Test Dataset		
	Cosine Similarity	Dot Product	Inverse Euclidean	Cosine Similarity	Dot Product	Inverse Euclidean
P@1	60.5%	68.9%	~~70.6%~~ 76.5%	73.4%	71.9%	~~73.4%~~ 76.6%
P@3	79.0%	81.5%	81.9%	85.9%	87.5%	87.5%
P@5	84.5%	85.3%	86.1%	90.6%	90.6%	92.2%
P@10	89.9%	91.2%	92.0%	92.2%	95.3%	95.3%

Source Selection - Error Analysis

1. **Some questions are semantically equivalent**

   e.g. *When is Easter this year?* v.s. *What day is Easter on this year*?

2. **Some questions are in the same type and are very close**

   e.g. *What is the capital of Louisiana?* v.s. *What is the capital of New Jersey?*

3. **Some questions are asking different information about the same thing**

   e.g. *What time does the Super Bowl start?* v.s. *Where is the Super Bowl being played this year?*

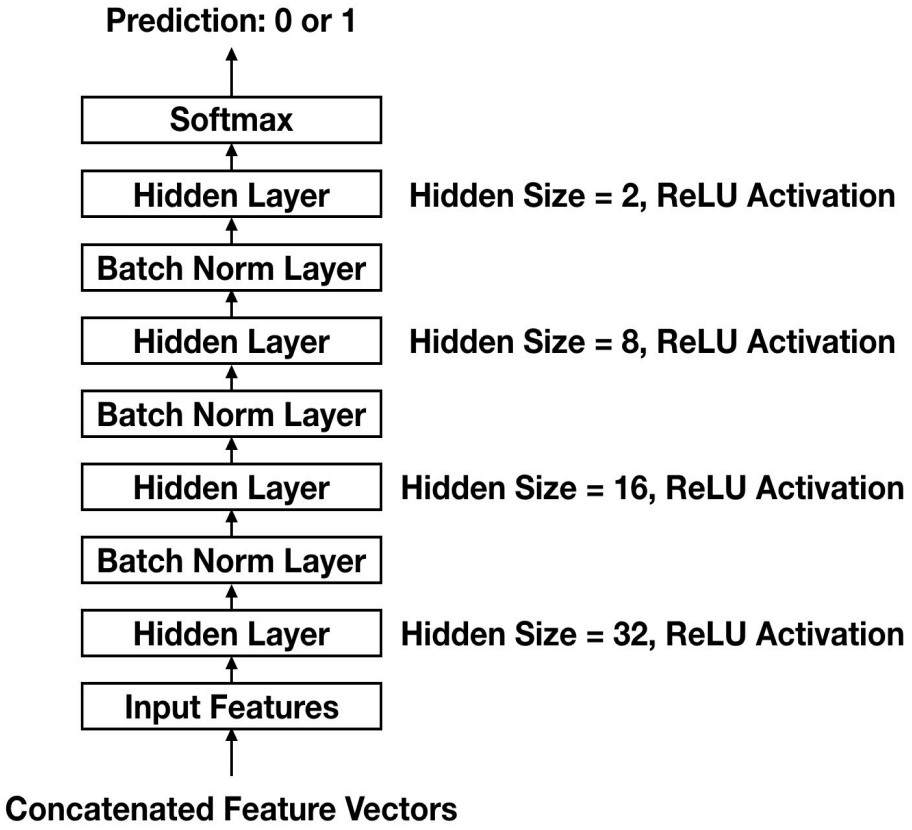
SELECT Clause - Model

ML classification problem:

- Given a question, for each column in the table, predict if the column should be included in the SELECT clause.

Method:

- Learn and extract important features for the question and each column.
- Build a ML classification model to predict the relevance score.



SELECT Clause - Features

- **Question type**
  (NUMERIC, HUMAN, ENTITY, LOCATION, ABBREVIATION, DESCRIPTION, BINARY QUESTION etc.)

- **Column type**
  (DateTime, Currency, Boolean, Number, Percentage, URL, Text etc)

- **Edit distance between question and column header**

- **Proximity between question and column content using word vectors**

SELECT Clause - Result

	Train Dataset		Test Dataset	
	Predicted: 1	Predicted: 0	Predicted: 1	Predicted: 0
Actual: 1	238	35	56	10
Actual: 0	305	1468	94	399

	Train Dataset	Test Dataset
Accuracy	83.4%	81.4%
Recall	87.2%	84.9%
Precision	43.8%	37.3%

SELECT Clause - Error Analysis

Among 64 cases in test dataset, we have 6 (9.4%) exact matches between the predicted SELECT clause and the actual SELECT clause. Besides, we have another 48 (75.0%) cases that we include all the required columns but also provide additional columns.

Case 1: Question type detected wrongly (7 cases)

Question: What is Washington Wizards record?
Actual: NUMERIC Predicted: ABBREVIATION

Case 2: Column type detected wrongly (3 cases)

Question: How long do cats lives? Targeted Column: Lifespan [4-5 years (In the wild)]
Actual: Numeric Predicted: Text

Case 3: Multiple columns in targeted type (24 cases)

Question: How many centimeters in an inch?
Predicted Columns: Centimeter & Inch

Question: What is the population of Boston MA?
Predicted Columns: 2018 Population & 2016 Population

SELECT Clause - Error Analysis

Case 4: Location type detection needed for WHERE-type questions (5 cases)

Case 5: Human Entity type detection needed for WHO-type questions (2 cases)

Current column type: DateTime, Currency, Boolean, Number, Percentage, URL, Text
Ideally we would like to break down Text into “Location”, “Person”, “Description”, “Instruction” etc.

Case 6: Binary questions requires more semantic features (4 cases)

Question: Does self employment require you to pay taxes?
Question: Are there any science jobs in Rochester NY?

Not able to establish a fixed relationship between question type and column type like other Wh-type question
May require more semantic features to get the model understand the question and each column fully.

WHERE Clause - Model & Features

ML classification problem:

- Format: **WHERE Column ~ Keyword**
- Given a question, for each column in the table and each word in the question, predict if (column, q_word) pair should be included in the WHERE clause.

Features:

- **Question type & Column type**
- **Minimum edit distance between q_word and column content** (Check if column X contain q_word)
- **Average Cell Length**
- **Number of Rows** (Single-row tables do not need row filtering)
- **Columns selected in SELECT clause**
- **POS/NER/Dependency parsing for q_word**

WHERE Clause - Result

	Train Dataset		Test Dataset	
	Predicted: 1	Predicted: 0	Predicted: 1	Predicted: 0
Actual: 1	117	3	18	12
Actual: 0	26	5199	36	1731

	Train Dataset	Test Dataset
Accuracy	99.5%	97.3%
Recall	97.5%	60.0%
Precision	81.8%	33.3%

WHERE Clause - Error Analysis

Among 64 cases in test dataset, we have 34 (53.1%) exact matches between the predicted WHERE clause and the actual WHERE clause. We also have another 8 (12.5%) cases are considered as not really wrong after manual checking.

Case 1: Single row tables (7 cases)

Returned results would not be different whether a WHERE clause is included or not.

Case 2: Slight different on search keyword (1 case)

Question: What is washington wizards record?

Actual SQL: SELECT W, L FROM NBA_Southeast_Standings
WHERE Team ~ washington_wizards

Precited SQL: SELECT W, L FROM NBA_Southeast_Standings
WHERE Team ~ washington

Team	W	L	Pct	Gb	Strk
Miami	44	38	0.537	-	W1
Washington	43	39	0.524	1	L1
Charlotte	36	46	0.439	8	W1
Orlando	25	57	0.305	19	W1

WHERE Clause - Error Analysis

Case 3: External information required (4 cases)

Question: Who is the actress that plays Sheldon's mother?
Actual SQL: SELECT Portrayed_by FROM Character_Appearances WHERE Character LIKE Mary Cooper

Question: What is my weight in kilograms?
Actual SQL: SELECT Kilograms FROM Table_1 WHERE Pounds = 100.00lb

Case 4: Incomplete search keyword due to stop-word removal (3 cases)

Question: How many feet are in a mile?

Actual SQL: SELECT \"Foot\" FROM \"Table_1\" WHERE \"Mile\" ~ \"a_mile\"
Precited SQL: SELECT \"Foot\" FROM \"Table_1\" WHERE \"Mile\" ~ \"mile\"

Discussion

- We have successfully divided the the problem into several well-defined sub-problems, and implemented a baseline model for each sub-problem.

- We also performed error analysis for each model to identify the future improvement areas.

- We have proven that the current approach works for most general cases. However, there are some complex cases and edge cases need to be further studied after we collect more data.

Discussion - Complex cases

Case 1: AND/OR operator

Question: What year was Brock Lesnar's last UFC fight?

SQL: SELECT "Date" FROM "Brock_Lesnar_Tabe" WHERE (("Fighter_1" ~ "Brock_Lesnar") OR ("Fighter_2" ~ "Brock_Lesnar")) AND "Event_Name_1" ~ "UFC" ORDER BY "Date" DESCENDING LIMIT 1

Case 2: Sub query

Question: Who became president after John Kennedy?

SQL: SELECT "President" FROM "List_of_Presidents_of_the_United_States" WHERE "Number" > (SELECT "Number" FROM "List_of_Presidents_of_the_United_States" WHERE "President" ~ "John_Kennedy") ORDER_BY "Number" ASCENDING LIMIT 1

Case 3: Aggregate function

Question: How many science jobs in Rochester NY?

SQL: SELECT COUNT (Title) FROM "Table_1" WHERE "Location" ~ "Rochester_NY"

Discussion - Edge cases

Case 4: JOIN operator

Question: Who is the actress that plays Sheldon's mother?

SQL: SELECT Portrayed_by FROM Character_Appearances WHERE Character LIKE Mary Cooper

Solution: Use JOIN operator to incorporate external information

Case 5: Questions with answers change over time

Question: Who won the super bowl?

Solution1: ORDER BY year DESCENDING LIMIT 1

Solution2: WHERE year = EXTERNAL("current year")

Lessons Learned

- **Technical learning**
  - Semantic inference is an unsolved problem
  - Better machine learning models require more data
  - Models built on train set may not be fully extended to test set due to different data distribution

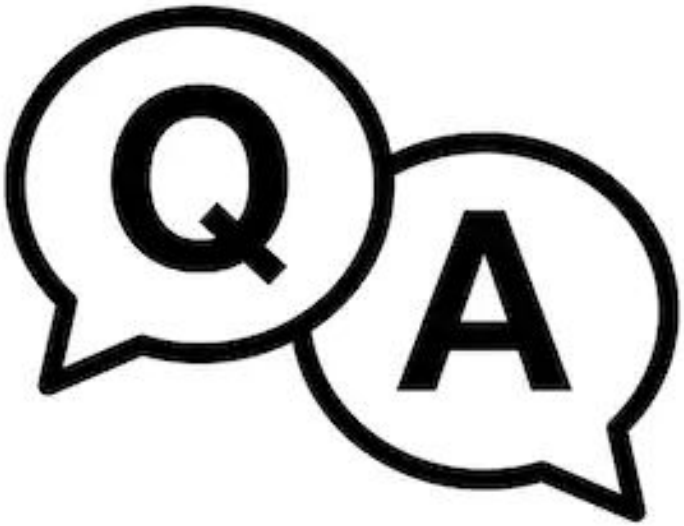
- **Other key take-aways**
  - A problem is composed of many small tasks
  - Ideal to work on sub-tasks and then stitch them together
  - Time management
  - Team dynamics
  - Presentation skills

Future Work

- Collect more data
- Handle differences in data distribution in train set and test set
- Comparative study using pre-trained models
- Identify more column types
- Incorporate more semantic features
- Generate features and build models for ORDER BY and LIMIT clause
- System integration

Conclusion

- Curated a dataset that benchmarks baseline performance metrics for structured question answering
- Dividing the problem into several well-defined sub-problems was a winning strategy, it helps us to better understand the challenges for each sub-problem.
- The source selection problems was solved with IR techniques.
- The table type recognition, the SELECT clause and WHERE clause problems were treated as a ML classification problems.
- Challenges encountered:
  - SELECT CLAUSE: question type classification, more granularity in column type classification, disambiguation with multiple matches, and semantic inference problem etc.
  - WHERE CLAUSE: data sparsity issue, data distribution issue, incorporating external information etc.


Q
A