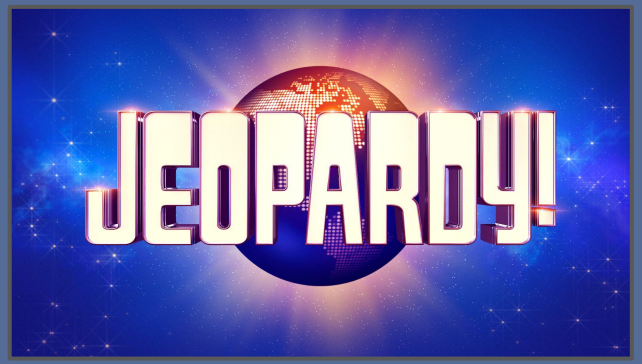
Deriving Jeopardy! Categories

Ben Burger

TABLE OF CONTENTS

01 Introduction



Problem

Derive the Jeopardy! category from the questions and answers.

Data Source

- J! Archive -- fan operated collection of Jeopardy! questions and answers
- Over 400,000 data points

Challenges

- Preprocessing
- Category name is often unique and of variable length
- Word play
  - The Flies of the Lord
  - The "Pit"

02

Baseline

```
Precisely Correct: 85
Partially Correct: 373
Total Correct:     458
---
Precise Accuracy:
0.018027571580063628
Partial Accuracy:
0.07910922587486745
Combined Accuracy:
0.09713679745493108

Overall Performance: 9.71%
```

Naive Approach

Idea: the category is likely related to the most common word in the questions and answers.

Methodology

- No need to train
- Just count the occurrences of words and emit the most common

Efficacy

- Can work well for proper nouns (especially unigrams)
- Naturally it cannot generate (>1)-grams

03

Rules Based

```
Precisely Correct: 250
Partially Correct: 361
Total Correct:     611
---
Precise Accuracy:
0.053022269353128315
Partial Accuracy:
0.07656415694591728
Combined Accuracy:
0.1295864262990456

Overall Performance: 12.96%
```

Improved Approach

Idea: take a rules based approach to categorizing the data. That is, record words that are related to a category and match a set of questions and answers with the best fitting category.

Methodology

- Use training set to build a dictionary of words for each category
- Look at new data, align it to a category based on dictionaries

Efficacy

- Can work well for common categories
- Naturally it cannot generate new categories

04

RNN

```
Precisely Correct: 37
Partially Correct: 661
Total Correct:     698
---
Precise Accuracy:
0.007847295864262 99
Partial Accuracy:
0.14019088016967127
Combined Accuracy:
0.14803817603393427

Overall Performance: 14.80%
```

Improved Approach

Idea: use a RNN to generate the categories. Category as an AMR to the questions and answers.
Key Goal: Improve upon the shortcomings of the previous two models: (>1)-grams and new categories

Methodology

- Machine translation with copy functionality
- Scheduled learning rate reduction

Efficacy

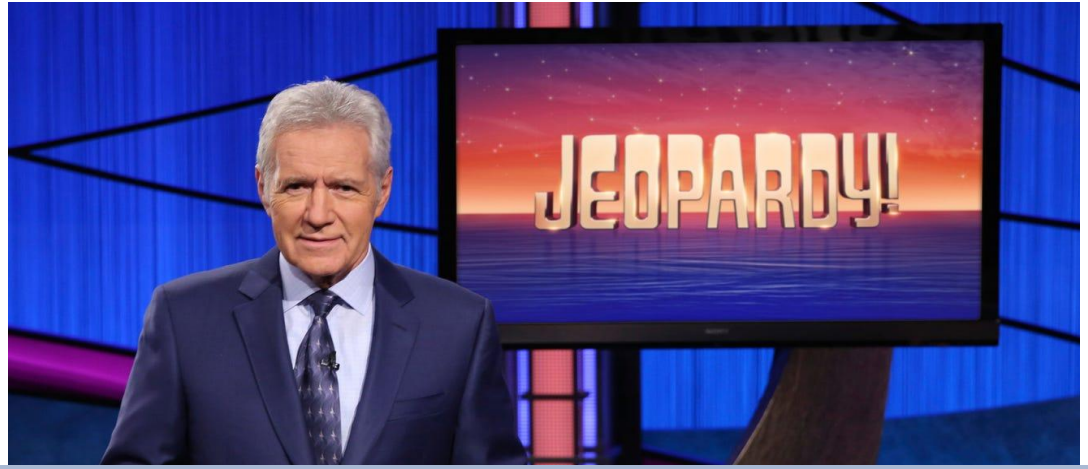
- Work in progress

Some Guesses

- the big to my independence
- the phrase phrase
- the last 2003
- be a salami
- computer music

05

Performance Summary



Baseline

Total Correct: 458
Performance: 9.71%

Rules Based

Total Correct: 611
Performance: 12.76%

RNN

Total Correct: 698
Performance: 14.75%

Questions?
Feedback?

Code available on  at https://github.com/bburger11/jeopardy-categories

Find the Jeopardy! datasets at https://j-archive.com/