This is a Task to Analyze Entities Within *Jeopardy!* Questions — What is: NER?

Megan Chiang, Sunghee Park, and Trang Tran
INFO 498 B | Autumn 2023

Introduction & Motivation

Jeopardy! is a popular quiz show where contestants are presented with general knowledge clues in the form of answers and must phrase their responses in the form of a question. Some of our group members have watched episodes of *Jeopardy!* during the 2010s-2020s, and we wanted to learn more about the show's history and see how the question topics have changed over time. Not only will our analysis provide us more insight into the show, but our findings might also be helpful for potential *Jeopardy!* contestants who want to review past questions. This dataset contains over 200,000 questions, so it would be very time-consuming for someone to manually parse through each question. Our analysis will help them better understand the general style and common topics of questions so they can focus on the most important topics to review.

Our overarching question for this analysis is: **Do *Jeopardy!* questions tend to be up to date with current events, or do they focus more on historical events? Has this changed over time?** To answer this, we trained an NER model using spaCy to extract the most common entities, then analyzed how the entities changed per year and determined if they have more of a historical or current context.

Corpus

The corpus contains 216,930 *Jeopardy!* questions, answers and other data from 1984 to 2012.

The columns of the dataset include question, category, value of the question, answer to the question, round, question number, and the air date.

- category' : the question category, e.g. "HISTORY"
- 'value' : $ value of the question as string, e.g. "$200"
  Note: This is "None" for Final Jeopardy! and Tiebreaker questions
- 'question' : text of question
  Note: This sometimes contains hyperlinks and other things messy text such as when there's a picture or video question
- 'answer' : text of answer
- 'round' : one of "Jeopardy!","Double Jeopardy!","Final Jeopardy!" or "Tiebreaker"
  Note: Tiebreaker questions do happen but they're very rare (like once every 20 years)

- 'show_number' : string of show number, e.g '4680'
- 'air_date' : the show air date in format YYYY-MM-DD

However, In order to use the NER model on both questions and answers, we combined the two columns with space between them, and added a new column called ‘Document’.

There were 216,930 documents in the dataset, but after dropping the NaN values in the answer and question columns, there were 216,928 rows.

The most common answers are portrayed in figure 1.

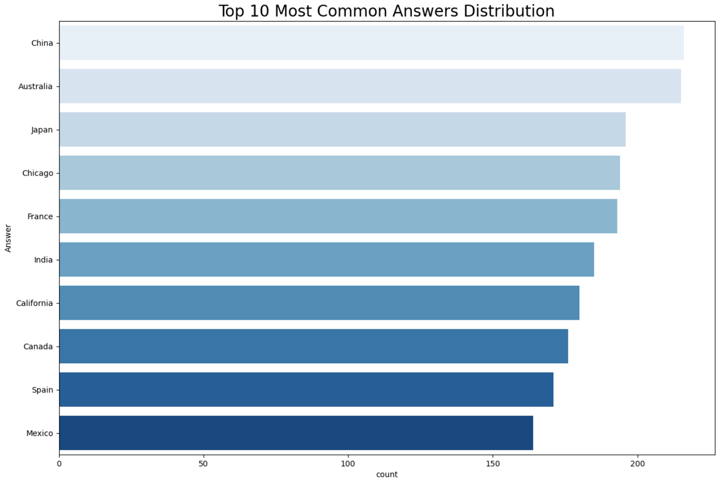

Figure 1. Top 10 Most Common Answers Distribution

The most common categories are portrayed in figure 2.

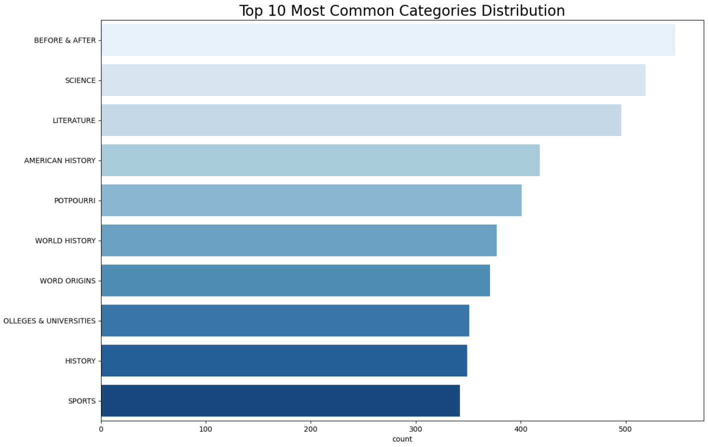

Figure 2. Top 10 Most Common Categories Distribution

Modeling

Training Data

We used **en_core_web_lg**, a spaCy model trained on web text, such as blogs, news, and comments. This web text came from OntoNotes 5, ClearNLP Constituent-to-Dependency Conversion, WordNet 3.0, and Explosion Vectors (OSCAR 2109 + Wikipedia + OpenSubtitles + WMT News Crawl). Some of the model's tasks include named entity recognition, part-of-speech tagging, and sentence recognition. It was created by Explosion, a software company specializing in NLP and AI.

For the named entity recognition task, the labels are the following (source for descriptions):

- **ORG**: Organizations, such as companies, agencies, institutions, etc.
- **PERSON**: People, including fictional
- **CARDINAL**: Numerals that do not fall under another type
- **GPE**: Geopolitical entities (countries, cities, states)
- **DATE**: Absolute or relative dates or periods
- **PRODUCT**: Objects, vehicles, foods, etc. (not services)
- **NORP**: Nationalities or religious or political groups
- **PERCENT**: Percentage, including "%"
- **LOC**: Non-GPE locations, mountain ranges, bodies of water
- **FAC**: Buildings, airports, highways, bridges, etc.
- **LAW**: Named documents made into laws
- **ORDINAL**: "first", "second", etc.
- **QUANTITY**: Measurements, as of weight or distance
- **MONEY**: Monetary values, including unit
- **WORK_OF_ART**: Titles of books, songs, etc.
- **TIME**: Times smaller than a da
- **EVENT**: Named hurricanes, battles, wars, sports events, etc.
- **LANGUAGE**: Any named language

The accuracy evaluation for the NER task is high, with a precision score of 0.85, recall score of 0.86, and F-score of 0.85.

[Sunghee] Model Architecture

Our primary task is named entity recognition. We will be using spaCy's English language model, which is pre-trained on a diverse range of text and incorporates large word embeddings. We believed that spaCy was the most appropriate for our use case because one of our goals was to model NER labels for categories in our *Jeopardy!* dataset. It is ideal for this task as we are able to use its entity recognition system to assign labels.

The task involves using spaCy's pre-trained large English language model (en_core_web_lg) to perform NER on textual data extracted from the 'Document' column in the *Jeopardy!* dataset. This column is a combination of the 'Question' and 'Answer' columns, capturing comprehensive information for each instance in the dataset. The code iterates through this 'Document' column, utilizing spaCy's NLP pipeline to identify named entities and compiles the results into a list named ‘ner_results’.

The input was the ‘Document’ column. The output was a list of ‘ner_results’ containing the named entities identified by spaCy's NLP pipeline. The labels the model can output are DATE, PRODUCT, ORG, CARDINAL, PERSON, GPE, TIME, WORK_OF_ART, LAW, ORDINAL, QUANTITY, FAC, LOC, NORP, EVENT, PERCENT, MONEY, LANGUAGE.

Training Considerations

While analyzing the corpus, we noticed there were some NaN values, [audio clue], [filler], [video clue], and other irrelevant entries that did not contribute to our analysis. Prior to performing Named Entity Recognition (NER), we systematically removed all instances containing these extraneous values to ensure a meaningful analysis.

Metrics of model performance on CoNLL dataset

We evaluated the model using the CoNLL2003 dataset, since it contains ground truth labels. The CoNLL2003 dataset is made up of Reuters news stories published from August 1996 to August 1997. Each article has been annotated with named entities of types location (I-LOC, B-LOC), organization (I-ORG, B-ORG), person (I-PER, B-PER), and miscellaneous (I-MISC, B-MISC).

In the evaluation of the model, entity recognition was conducted by utilizing spaCy model. Subsequently, an evaluation of the generated labels was performed by comparing them with the ground truth labels using the metrics.classification_report() method.

```
              precision    recall  f1-score   support

         LOC       0.45      0.02      0.04      1015
         ORG       0.48      0.54      0.51       782
         PER       0.70      0.88      0.78       867

   micro avg       0.60      0.45      0.52      2664
   macro avg       0.55      0.48      0.44      2664
weighted avg       0.54      0.45      0.42      2664
 samples avg       0.30      0.24      0.26      2664
```

Evaluating Considerations
Throughout the process of model evaluation, it was observed that the spaCy model yielded labels distinct from those present in the CoNL2003 ground truth. Consequently, we opted to exclude labels other than LOC, ORG, and PER. Additionally, due to the spaCy NER's inability to recognize the MISC label we chose to omit this particular label from the CoNLL2003 ground truth labels. Furthermore, because CoNLL2003 ground truth labels included location (I-LOC, B-LOC), organization (I-ORG, B-ORG), person (I-PER, B-PER), and miscellaneous (I-MISC, B-MISC), we combined both I- and B- instances into a unified representation. In summary, the evaluation focused on the three labels: PER, LOC, and ORG.

Due to the refinement process, the evaluation scores are relatively low, especially for the LOC label. The presumed reason is that the spaCy NER model outputs a broader range of labels, for example, GPE, LOC, ORG, and NORP, whereas the CoNLL ground truth data appears to categorize all or some of them exclusively as LOC and ORG.

Describing/visualizing results

[Megan] First Visualization: Word Clouds

Method and Process
To create the following word clouds, we first used spaCy's model to extract entities in each document, then added these named entities as a new column in the dataset. We then organized the entities into dictionaries based on their type, with the keys being the entity spans and the values being their counts. We initially made dictionaries for *every* entity type, but since our main goal was to determine whether *Jeopardy!* questions focused more on historical or current events, we ultimately decided to narrow it down to five entity types: organization (ORG), person (PERSON), geopolitical entities (GPE), events (EVENT), and works of art (WORK_OF_ART). Below are the number of entities for each type that the model identified.

Entity Type	Number of Entities
Organization (ORG)	37886
Person (PERSON)	46331
Geopolitical entities (GPE)	8183
Events (EVENT)	2075
Works of art (WORK_OF_ART)	25080

We found that these five types made up about 77% of the total entities.

Word Clouds

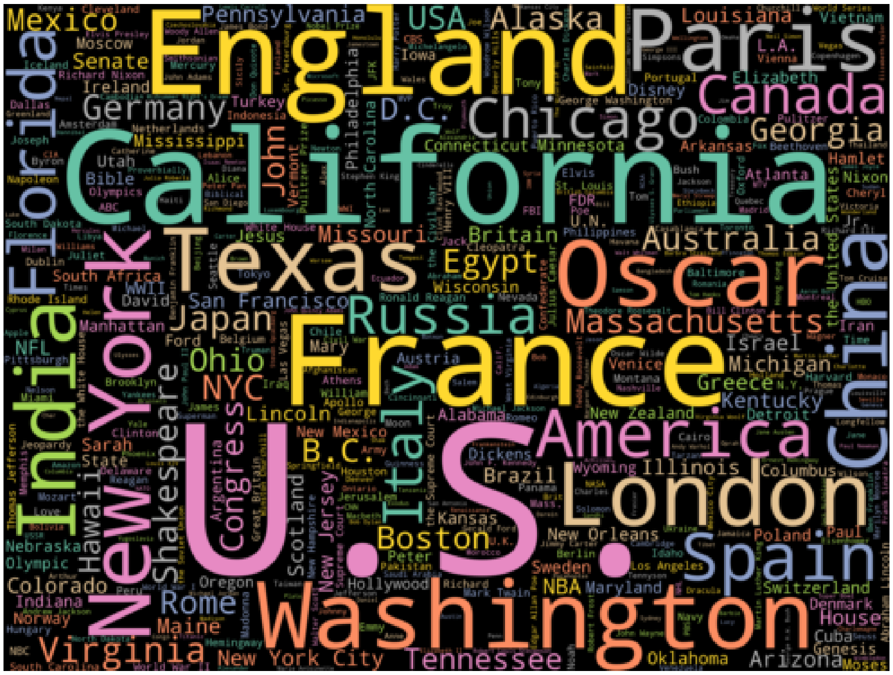

Figure 3. Word cloud for entity types ORG, PERSON, GPE, EVENT, and WORK_OF_ART

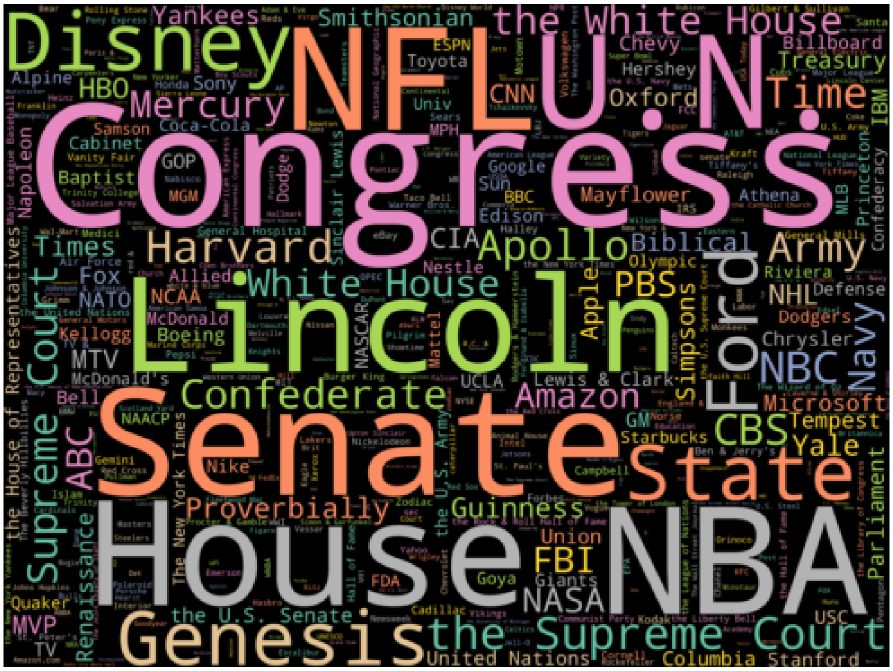

Figure 4. Word cloud for entity type ORG

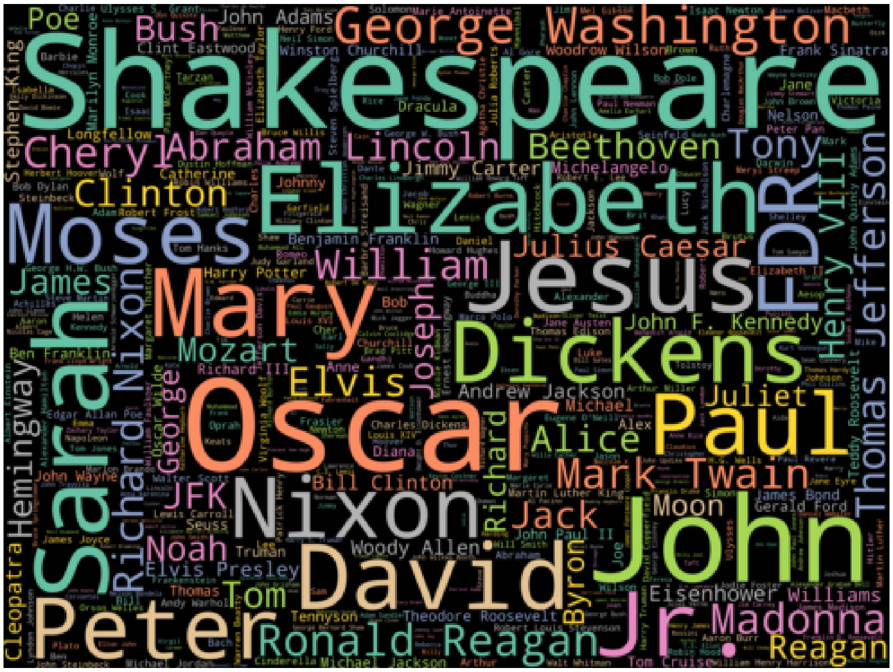

Figure 5. Word cloud for entity type PERSON

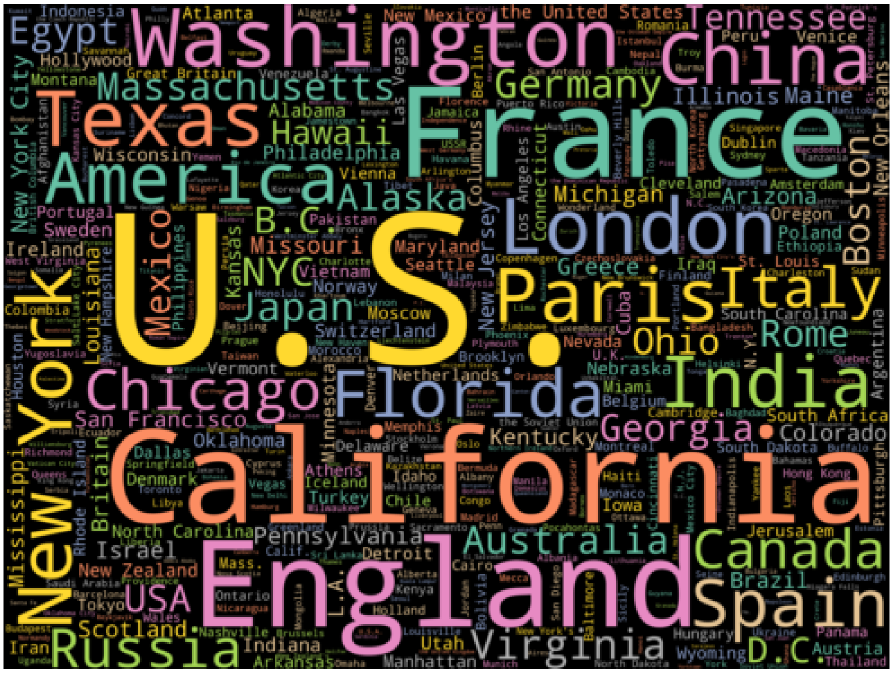

Figure 6. Word cloud for entity type GPE

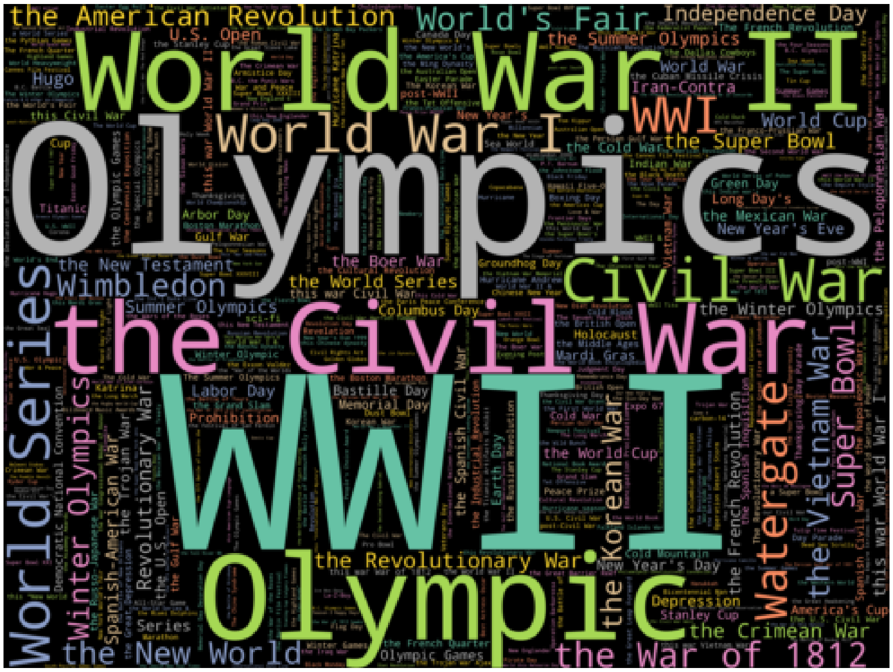

Figure 7. Word cloud for entity type EVENT

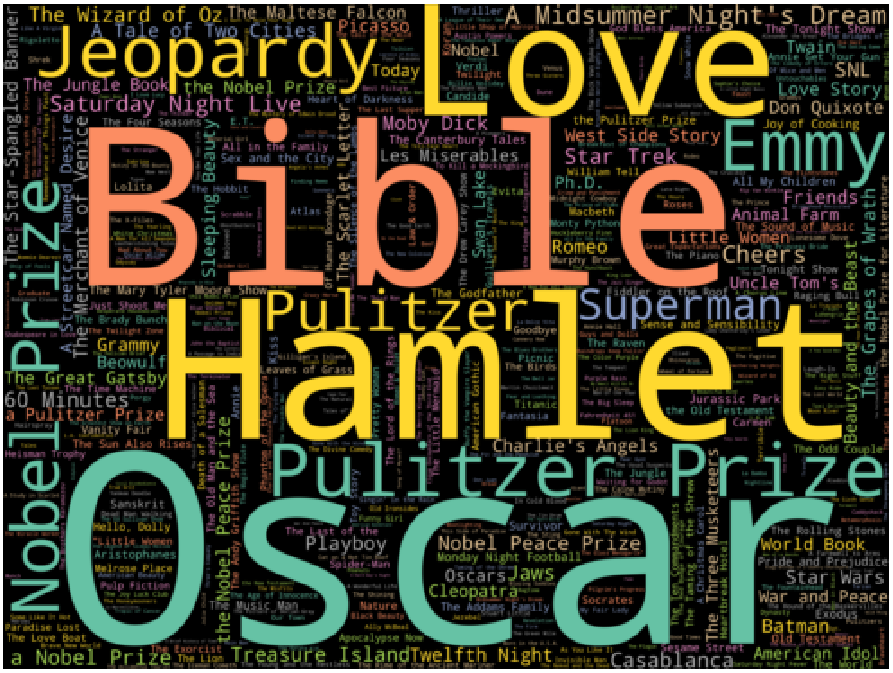

Figure 8. Word cloud for entity type WORK_OF_ART

Key Takeaways

Overall, the model was able to identify entities and tag them with their correct label, since the words and phrases in all six word clouds generally make sense with their respective entity type.

After finding that the ten most common answers (Fig. 1) were all countries, states, and cities, we predicted that the model's most common named entities would be similar. In the word cloud for all five entity types (Fig. 3), many of the most frequent spans were countries, states, and cities as we had predicted.

Additionally, in Fig. 5, Fig. 7, and Fig. 8, many of the most frequent spans had a more historical context than a current one, such as "Shakespeare" (Fig. 5), "Dickens" (Fig. 5), "World War II" (Fig. 7), "Civil War" (Fig. 7), and "Hamlet" (Fig. 8). This suggests that the topics in *Jeopardy!* questions and answers are more focused on historical events.

[Trang] Second Visualization: Bar Charts

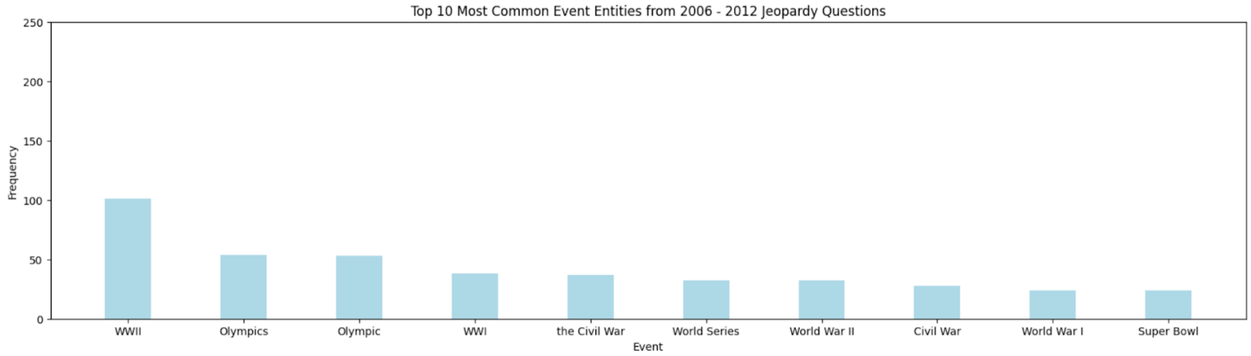

Figure 9: Top 10 most common event entities from 2006 - 2012 Jeopardy Questions

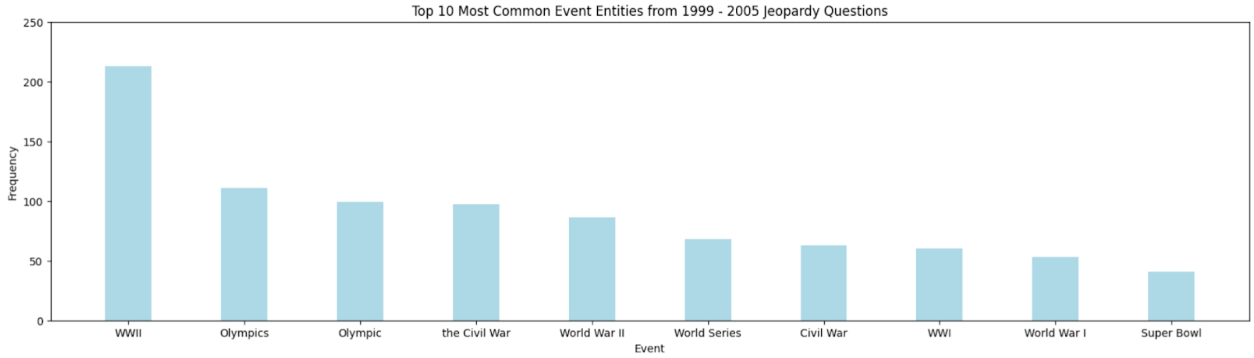

Figure 10: Top 10 most common event entities from 1999 - 2005 Jeopardy Questions

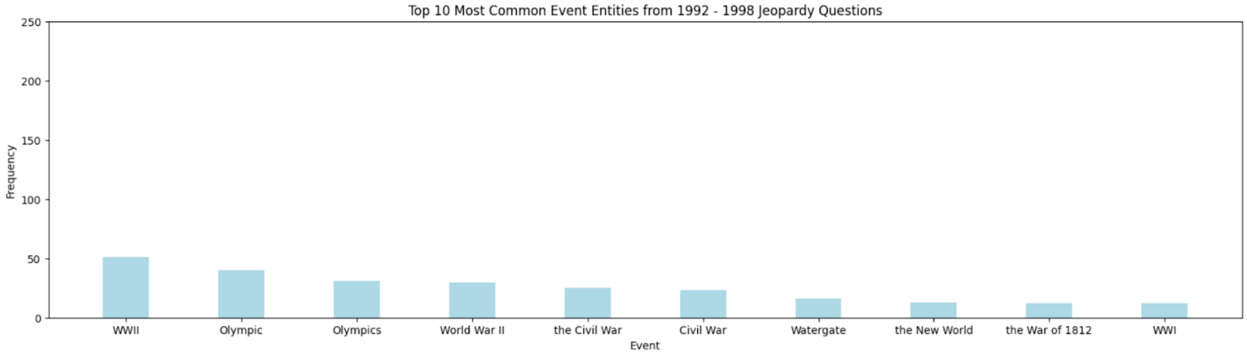

Figure 11: Top 10 most common event entities from 1992 - 1998 Jeopardy Questions

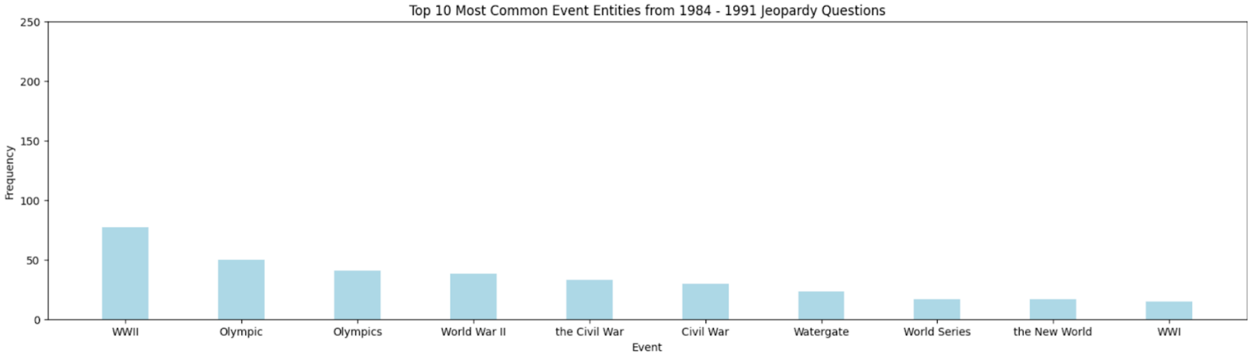

Figure 12: Top 10 most common event entities from 1984 - 1991 Jeopardy Questions

Based on the results of our model and visualizations, we can conclude that **most *Jeopardy!* questions tend to be based on historical events instead of current events**. After dividing the dataset into four different year ranges (2006 - 2012 , 1999 - 2005, 1992 - 1998, 1984 - 1991) and calculating the most common event entities on the *Jeopardy!* dataset in each range, we observed that there was a pattern of some common historical events that repeated in the top most common entities.

The events in the questions from the years 1999 to 2012 (see Figure 9 and Figure 10) appeared to be similar to each other, and the events from 1984 to 1992 (see Figure 11 and Figure 12) appeared to be similar to each other. Additionally, World War II, Olympics, and the Civil War were historical events that continued to show up as common events in *Jeopardy!* questions in recent years. There are some new events that appear such as the Super Bowl, however, the majority of the event entities based on frequencies convey that there are more questions that focus on historical events.

Discussion

Some interesting observations that we found are that the historical events that are most common in the questions tend to stay similar across the different years but have different frequencies on how often they appear. Although there were mostly historical events as common entities, we found it interesting to see events that occurred near the year range get incorporated into the *Jeopardy!* questions even though they were not as common such as the Super Bowl and Watergate. Overall, seeing significant historical events continue to be in *Jeopardy!* questions in recent years demonstrate the lasting impact that history has, and it is intriguing to see how new events continue to get incorporated into *Jeopardy!* as time goes on.

Challenges

We faced challenges in our model evaluation process. Initially, we had difficulties manually annotating ground truth labels for the sampled *Jeopardy!* dataset, which led to limitations in our ability to conduct a more precise evaluation. Furthermore, because the CoNLL ground truth labels and the spaCy NER model output labels were different, it was a difficult task to align them.

Additional Questions

Some additional questions our results raised include:

1. What are some other patterns seen in the entities of these Jeopardy questions across the four year ranges other than in the historical events and historical figures?
2. What are the most common entities for each of the *Jeopardy!* categories (e.g., Before and After, Science, Literature, etc.)?

The categories in the *Jeopardy!* dataset appeared to be specific making it harder to categorize the topics they fit in. Some next steps that we hope to take are creating broader categories (e.g., science, art, history) and to label each question in these categories. We also hope to see how the model performs for each of these broader categories and not just the provided category.