What Can We Do with Large Language Models and Cultural Heritage Texts?

When AI meets Irish drama

Project Advisors: Professor Catherine Flynn, Kent Chang. **Students:** Jacob Aldrich, Vera Guo, Shujing Hu, Meher Kajaria, Laura Ma, Melissa Tsaowimonsiri, Demi Wang.

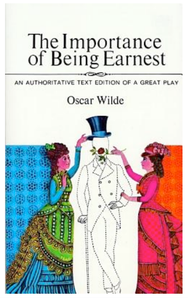
Problem and Significance

This project aims at enabling text analysis with a specific focus on Irish plays. By extracting and standardizing structural elements from OCR'd Irish plays, we intend to deal with unstructured data and make computational analysis of the texts easier.

The significance of this project lies in its potential to enhance our understanding of Irish drama and literature, which can thus facilitate research in digital humanities and computational linguistics. Furthermore, the comparison of different approaches illustrate various ways of handling unstructured data, which can be applied to other domains beyond Irish drama.

Data

We used 400+ Irish plays for the project. These were scanned or retrieved online. For scanned plays, an OCR pipeline was used to convert scanned pdfs into text files. Both were then processed further (see LLM-Assisted Data Cleaning).



LLM-Assisted Data Cleaning

Procedure

To assist our data cleaning process, we used a LLM (ChatGPT 3.5) with specific prompts to do things like:

- Remove typos.
- Remove page headers/footers.
- Standardize format.

Results

ChatGPT appeared to clean the texts very well. However some issues still occurred, such as:

- Hallucinations.
- Inconsistencies in speaker labels.
- Modernization of very old plays.

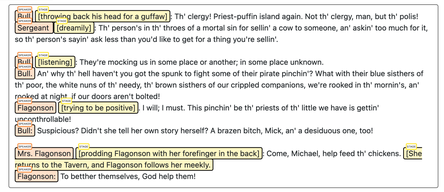
Some advantages of using ChatGPT are:

- It's good at removing typos, footers, headers, page numbers, and other metadata.
- It removes textual noise from scanning.
- It can correctly infer labels from context.

Structural Element Labeling

Manual Annotation

- Used Potato (Portable Text Annotation Tool).
- Labeled elements manually to implement the BERT model.



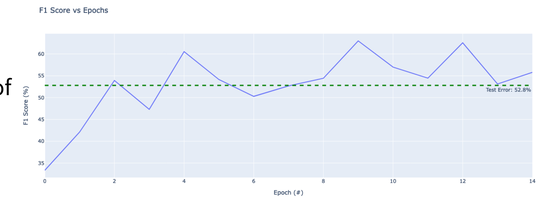
Regex

- Identified structural elements and all syntactic variations of elements.
- Converted syntactic variations into Regex.
- Used Python and Regex script to label structural elements.
- Happened to be our best structural element labeler.

```
(3, 'SPEAKER'): [
    r'(?<=\<br\> )([A-Za-z0-9 ]+:)',
    r'^[a-zA-Z]+[\.]*[\s]*[a-zA-Z]+$',
    r'^[a-zA-Z]+[\.]*[\s]*[a-zA-Z]+\.$',
    r'^[a-zA-Z]+[\.]*[\s]*[a-zA-Z]+\.$',
    r'^[a-zA-Z]+[\.]*[\s]*[a-zA-Z]+\:$',
    r'^[a-zA-Z]+[\.]*[\s]*[a-zA-Z]+\s$',
    r'^\_[a-zA-Z]+[\.]*[\s]*[a-zA-Z]+\:\_$',
    r'^\_[a-zA-Z]+[\.]*[\s]*[a-zA-Z]+\_\.$'
],
```

BERT

- Implemented in PyTorch, ran on Google Colaboratory.
- Achieved F1 Training/Test scores of 63% and 53% respectively.
- Potential for improvement!



ChatGPT

- Prompted GPT with a few annotated structural elements of plays.
- ChatGPT lacks understanding of structural elements, tokens, labeling.
- Nonviable method of element labeling at the moment.
- All elements represented in prompts.
- Specificity and rules are key to using GPT.

O O O O B-SPEAKER I-SPEAKER O O O O O O O O O O B-STAGE I-STAGE I-STAGE I-STAGE I-STAGE O O

Main Findings

- We processed texts using both ChatGPT-stripped texts and regular stripped texts.
- Regex was the best method to label structural elements, BERT has some promise, and ChatGPT is not viable yet.
- LLMs worked best in rule-based static environments, like with text processing.
- Manual methods work best in dynamic situations like play writing or annotation.

Towards Generation and Understanding?

GPT-2 Generated Play

SEBASTIAN: My brother's living, you little boy!
RAMON: He's asleep, he's dead, you big man!
PEDRO: My little angel!
(After an exhalation)
SYLVESTER: (shaking hands with PEDRO) 'Bye man, bye man, bye man.
(The band starts up. MICHAEL enters.)

- Trained a fine-tuned GPT-2 model capable of generating Irish play texts by providing a prefix.
- The generated text demonstrates the manifestation of various structural elements and the general style of an Irish play script.

Chat with GPT-4 to gain understanding

- Utilized LLM such as GPT-4 to analyze an Irish play script, ask it directed questions to better understand the different structural elements and textual meanings.
- Teachers can use LLM to assist classroom teaching and encourage students to think more deeply and explore the text further.
- LLM can also help us delve more deeply into the structure, themes, and cultural background of plays, thereby facilitating better preservation of these cultural heritages.

1. How many characters are in this play text?

There are five characters in this play text: SEBASTIAN, RAMON, PEDRO, SYLVESTER, and MICHAEL.

2. What are some of the character interactions in this text?

- SEBASTIAN speaks to RAMON.
- RAMON speaks to SEBASTIAN.
- PEDRO is called a "little angel."
- SYLVESTER shakes hands with PEDRO and bids him farewell.
- SYLVESTER speaks to MICHAEL.
- MICHAEL speaks to SYLVESTER.

3. Are there any stage directions in this play text?

Yes, there is one stage direction: "(After an exhalation)".

Berkeley
UNIVERSITY OF CALIFORNIA