Word count of document (excluding this line): 1199

1. Introduction

We want to explore the differences in sentence formation within English between different cultures. As much as English is a single language, the various way that different cultures adapt the language are fascinating. How Yi Da and Jashwant have similar yet slightly different ways of addressing similar things, for example, chai and tea.

There are two objectives for this project. Using GPT-playground, we will be using GPT-4 to classify and generate english sentences based on the tone and word choice of different English dialects. As per the initial paragraph, different cultures have different words to refer to similar objects, greet each other differently, incorporate other languages within their region to their use of English. We would like to explore what GPT-4 is capable of accomplishing with regards to dialect based on classification and sentence generation.

Classification would involve passing a sentence into Chat-GPT and having it guess what kind of dialect the sentence originally originated from. Generation would use the system prompt to set the dialect of GPT-4 and having it produce a conversation between two people using that dialect.

2. Background

Commented [1]: Background - Describe 2-3 related papers you've found

"Language and Dialect Identification: A Survey" makes a comparison between American and British english pronunciations of consonant, vowels, fricatives. It also differentiate between 2 dialects by the way a sentence is framed:

- American or British English : "Go and turn off the heater!"
- Singaporean English :"Go and switch off the heater lah!"

The paper also taks about the methods used to differentiate between dialects. Such as clustering methods, ACCDIST for measuring the similarity and difference in people's accents.
"Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology" gives the difficulty in finding a dataset that has equal distribution of different dialects. An example such as English dialect referenced from social media might have more american english rather than Indian English, Singlish, Australian English or British English.
They also talk about sentence formations in different english speaking countries and an example of how a sentence can end in a preposition such as "What are you preparing for?"..
They build a model that can differentiate between languages and feature set, robustnes of the model, accuracy and the uniqueness of regional varieties. The model provided the difference between different parts of speech and their importance and the diversity between dialects of a particular language.

3. Source of data and processing

Commented [2]: Source of Data and Processing - Where will you get the data for part of project? What work will you do in the collection/labelling

Our source of data will likely come from kaggle datasets involving conversations with people, OpenSLR datasets, while are mostly audio datasets, come with audio transcripts which we can use as datasets. These datasets while generally abundant, are open done using grammatically correct formal English, which might have little to no dialect or cultural influence.

The other option will be to find and take the transcripts of street interviews. This will provide a more genuine and natural way of how someone might speak and converse. The problematic part to this is that most publications would fix their transcript in post and the transcript reflects the meaning of what is said but is not the exact wording. We would likely need intervene ourselves to record down the actual sentences verbatim to make sure we capture it as well as we can.

The last thing we will be doing is to include in our own sentences, Jashwant and Yi Da come from different cultures, with Jashwant being Indian and Yi Da being southeast-asian. We will both be writing up sentences that are from our cultures and including them into the dataset

4. Architecture of the model

Commented [3]: Architecture of the model - Rough guesses of type and structure of model; If you're doing a Class 2 project, then give the structure of the system you're proposing.

For the classification of dialects, we will be specifying in the system prompt that we would like the model to specify what dialect it thinks the sentence given in the prompt is. The model would then output a 3 character abbreviation that should signify the prediction. The python code that will be used to call the GPT API will then compare the output of the model with the label of the sentence so we can get the accuracy of the prompt/model. We will then improve the prompts to improve the accuracy of the output of the model as much as possible.

We will be running the classification architecture with 2 variants. One variant where the model doesn't have the option of classifying the sentence as formal English and another model that can. Since there is a scientific and formal way of writing English, we would also like to test if having another label for the model to put uncertain prompts into.

For the dialect conversation generation, we would input in the system prompt to generate a conversation using a dialect based on the topic that will be passed in through via the input prompt. Two different variants will be tested, zero-shot and few-shots. For the few-shots variant, we will be passing the examples of conversation held in that dialect.

5. Comparison

Commented [4]: Baseline Model or Comparison - Describe a simple baseline model that you'll compare against; in the case of a Class 2 project, then say how you'll know that you've succeeded.

For the classifier, we will be judging it based on the accuracy. We are hoping for **95%** for a 2-class classification and a **85%** for an n-class classification.

For the generator, **we will need to come up with a rubrics**. At the moment, we are not looking at grammar as being a key point of comparison as in certain dialects, the english grammar isn't followed as strictly and it adapts some of the native language's grammar into it, e.g. malay. We will use a 5 point scale based on the use of specific vocabulary that is inherent to the dialect and any adaptation of the local grammar into the conversation.

Depending on how well the classifier does, we might also pass the generated conversation back into GPT-playground to classify and make sure it gets it right.

6. Plan

Week of 30th Oct	Sourcing of Database and labeling datasets
Week of 6th Nov	Sourcing of Database and labeling datasets
Week of 13th Nov	Classification prompt based testing using zero-shot, single-shot, multi-shot, etc
Week of 20th Nov	Generation prompting using zero shot, single-shot etc..
Week of 27th Nov	Finetuning our prompt and preparing for presentation.
Week of 4th Dec	Tidying up presentation slides and final report

Commented [5]: Plan - List of sub-tasks - and your guess as to how much time each task will take; Use to create estimate of end-to-end time; Discuss how you're going to work together

7. Risks and mitigations

Commented [6]: 7. Risks - Predict what might go wrong & how you'd recover

There are a couple of risks going into the project.

The first risk would be if OpenAI for some reason restrict the use of the GPT-playground. This would severely hinder the progress of our project and we would likely have to pivot to an open-source version, such as GPT-2 or Llama 2. If this ends up being the case, we will limit the project to just doing dialect classification, since there is less complexity in completing that.

The other risk is that the datasets end up being really difficult to obtain, in which case we would likely need to spend more of our time transcribing the data to be used for the project. We have been having difficulties finding a good dataset that doesn't use English that is too formal.

The last risk is that OpenAI might just update their models and the results become difficult to replicate. If the GPT-4 weights are modified during the project duration, our experimental data will become obsolete. The mitigation is to save all our codes and have it be as automated as possible so we are able to rerun and get our new results.