What Dialect La

Kang Yi Da and Jashwant Raj Gunaseelan
Word Count : 1997
Penality : 0%

December 12, 2023

1 Introduction

What Dialect La has two main objectives, using GPT-4, we are looking to classify dialects based on written text and to regenerate sentences in one dialect into another dialect. For a foreign individual moving to a new place, the best way to adapt is by understanding the region's culture and getting accustomed to its way of life. Language is one significant way to adapt, tailoring responses that are most respectful to people of that culture and ensuring we do not tap into anything that may be taboo within that culture shows significant effort.

The best way to understand these nuances requires an in-depth knowledge of the sentence structure and the slang common to each English-speaking region. Machine learning helps us understand those attributes and assists in translating or restructuring sentences, thereby omitting the need for manual assistance.

2 Background & Related Work

"Language and Dialect Identification: A Survey" [1] compares American and British English pronunciations of consonants, vowels, and fricatives. It distinguishes between dialects based on sentence structure, illustrated with examples:

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

The paper also explores methods for differentiating dialects, including clustering techniques and ACCDIST, a tool for measuring similarities and differences in accents.

"Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology" 2 discusses the challenges in compiling a dataset with an equal distribution of various dialects. For instance, English dialects represented on social media may predominantly feature American English over Indian English, Singlish, Australian English, or British English. This study also examines sentence structures across different English-speaking countries. It highlights how sentences can end with a preposition, as in "What are you preparing for?"

Furthermore, the research details the development of a model capable of differentiating languages. It assesses the model's robustness, accuracy, and ability to capture regional varieties' uniqueness. The model also distinguishes between different parts of speech, emphasizing their significance and the diversity within dialects of a specific language.

3 Data and Data Processing

We obtained the datasets from online street interviews and picked out specific sentences that we thought were telling of the dialect. Some of the sentences were very obviously of a certain dialect while other sentences are vaguer and can share vocabulary and sentence structures with other dialects. We collected 30 sentences from 6 different dialects and 30 sentences that we felt were more neutral. Examples:

Table 1: Dialect Examples

Dialect	**Sentence**
Australian	Yeah nah, not in the mood for it tonight
British	Aye it's not nice. Lots of druggies, scroungers, you know.
Indian	My cousin brother lives in Canada.
Irish	That film was brilliant, so it was.
Singaporean	I'm going to the market, you want come along?
South African	I'm after getting a new job in the city, so I am.
Standard	As a safety feature, a fail-safe feature is enabled by default.

Based on experimentation conducted afterwards, some of the sentences had another label added due to reasons explained in section 6.

4 Architecture & Software

GPT-4 was used to produce the results for both the classifier and the translator. The final prompt for both systems used will be provided in the Github codebase due to length constraints.

4.1 Classifier

The context prompt tells GPT-4 to be a classifier and the kinds of labels that are available. The sentence that we want GPT-4 to classify is the input prompt and GPT-4 will provide the classification label.

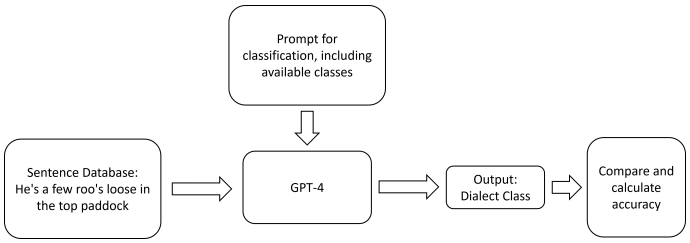

Figure 1: Classifier Structure

4.2 Translator

The context prompt tells GPT-4 to translate a specific dialect to another specific dialect. The sentence that we want GPT-4 to translate is the input prompt and GPT-4 will provide the translated sentence.

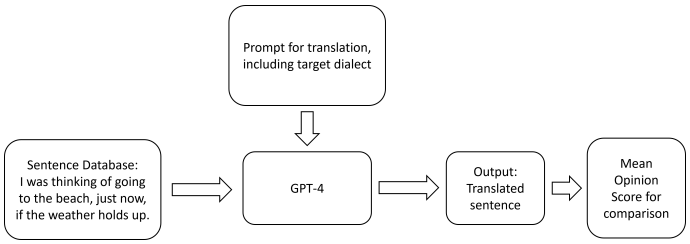

Figure 2: Translator Structure

5 Comparison

We compare to GPT-4's zero-shot prompt variant.

The context prompt used for the classifier is:

```
'You are a dialectologist specializing in English. Your job is to classify
the dialect. The classification labels are:
1 = British dialect
2 = Singapore dialect
3 = Indian dialect
4 = Australian dialect
5 = South African dialect
6 = Irish dialect
0 = No dialect
Just output the labels'
```

The context prompt used for the translation is:

```
f'You are a dialectologist specializing in English. Your job is to translate the
prompt sentence into {targetDialect} English.'
```

For chain-of-though, the classifier prompt provides a unique example for each dialect and explains the reasons behind it by highlighting the unique features of the sentence. The translation prompt provides an example of a single sentence translated to the target dialect, along with an explanation of the thought process. You can find the example and methodology on the "What Dialect La" GitHub project group.

6 Results

6.1 Classifier

6.1.1 Quantitative

As seen in table 2 the performance of our dataset using zero-shot and chain-of-thought prompt is similar. After analysis, we deduced the issues arose because of 2 reasons.

1. Sentences can have more than one class.

2. Some sentences are more formal and have no differentiable dialect

To overcome this, we introduced an extra label to sentences that fall under the above categories. Thus making this a multi-label classification dataset. The team notes that the classifier performs remarkably compared to a randomly-selecting system.

Table 2: Classification Accuracy

Classes	**Original Zero-Shot Accuracy (Baseline)**	**Original Chain-of-Thought Accuracy**	**Revised Chain-of-Thought Accuracy**
Australian	96.67%	96.67%	96.67%
British	86.67%	86.67%	93.33%
Indian	80%	90%	90%
Irish	46.67%	46.67%	76.7%
Singaporean	73.33%	63.33%	83.3%
South African	43.33%	46.67%	80%
Standard	100%	100%	100%
Total Accuracy	75.24%	75.71%	88.57%

6.1.2 Qualitative

The classification was largely accurate. Correct classification examples are given in 3. The Indian and Singaporean sentences are prime examples of the system being able to classify languages even without dialect-specific vocabulary.

Table 3: Correct Classification

Classes	**Sentences**
Australian	He's a few roo's loose in the top paddock
British	You all know that if there's insulting to be done, then it's me what should be doing it! No-one's allowed to be rude to my son, except me!
Indian	They is more better than what you did before. Atleast you took the the time to change it
Irish	Would you be after stopping at the shop for some milk?
Singaporean	Have you finished your homework already or not?
South African	This curry is lekker strong, isn't it?
Standard	It leverages both an autoregressive decoder and a diffusion decoder; both known for their low sampling rates.

Table 4 shows a few sentences that have been classified incorrectly. The system struggles when unique vocabulary like 'bloke' and 'gobsmacked' are shared between linguistically similar dialects. The system also tends to veer towards predicting standard cases more often, but this will be discussed further in section 7.

Table 4: Incorrect Classification

Original Class	**Classifed As**	**Sentences**
Australian	British	He's a top bloke, really decent.
British	Standard	He looks at the telephone. He picks up the receiver to check if it is still working
Indian	Standard	Don't worry about the predicted results, ill try my level best to get something close enough
Irish	British	I was absolutely gobsmacked by the news!
Singaporean	Indian	This year, we started having seventeen volunteers each row
South African	British	I used to could swim very fast when I was younger.

6.2 Translator

6.2.1 Quantitative

As the team possesses the knowledge to evaluate the quality of Indian and Singaporean dialects, translating sentences from other dialects into either Indian or Singaporean is the primary focus. We translated 60 sentences from dialects other than the target dialect.

A 1 to 5 mean opinion score is used to grade the results. Table 5 show the rubrics.

Table 5: Mean Opinion Score

Points	**+0**	**+1**	**+2**
Meaning preserved	Meaning of original sentence was lost	Meaning of original sentence was lost	The meaning of the sentence got preserved entirely.
Sentence structure and vocabulary of the target dialect	The generated sentence does not contain any aspects of the target dialect.	Some aspects of the target dialect were shown but there were either grammatical or sentence structural mistakes	The generated sentence has the sentence structure and some vocabulary from the target dialect.

Based on the rubrics, table 6 shows the results. Scores were aggregated for 30 translations. Our target model performed significantly better than our baseline by 0.4.

Table 6: Translation Accuracy

Prompt Style	Zero-Shot	Zero-Shot (Baseline)	Chain-of-Thought	Chain-of-Thought (Best)
Temperature	1	0	1	0
Singaporean	3.55	3.98	3.88	4.20
Indian	3.47	3.90	3.75	4.45

6.2.2 Qualitative

The best configuration provided good chain of thought for the translation, leading to better translation. Table 7 provide a few examples translations:

Table 7: Translated sentences

Original Sentence	Accadaccas is something the Aussie feel proud of ,since it's a rock band that has its origins in Australia	Lets get a burger at maccas, might put a smile on the kids face.
Zero-Shot, Temp = 1	Accadaccas is something the Aussies take a lot of pride in because it's a rock band that comes from Australia only, na.	Let's go grab a burger at MacD, maybe can make the kid happy one.
Zero-Shot, Temp = 0 (Baseline)	Accadaccas is something the Aussies take pride in, yaar, because it's a rock band that started in Australia only.	Let's go eat burger at Macs, sure can make the kids happy one.
Chain-of-Thought, Temp = 1	AC/DC is something we Indians feel proud of, since it's a rock band that originated in Australia.	Let's go eat burger at McD lah, can make the kids happy."
Chain-of-Thought, Temp = 0 (Best)	AC/DC is something that Australians take pride in, since it's a rock band that originated in Australia.	Let's go eat burger at Mac, maybe can make the kid happy.
Target	Indian	Singaporean

The poor results were either due to poor use or overuse of ***lah and shiok*** and ***yaar and only*** or when the target dialect uses different sentence structures. The system translated *"dearest wife"* into *"sayang wife"*, *"sayang"* when used before a noun is a verb but when used after, it is an adjective, which the system might not have enough information on.

7 Discussions & Learnings

Dialects are hard to define as classes.

People pick up dialect traits from wherever they have lived. A spoken dialect could have attributes from multiple dialects. Since the main purpose of the classifier is for cultural sensitivity, classifying a sentence or conversation as multi-dialect can be a viable option.

Providing examples in the context prompt makes GPT-4 more risk-averse.

When provided examples of each dialect, GPT-4 was more keen on classifying certain sentences as standard English. Even though our overall accuracy was largely the same, the single-label chain of thought classifier predicted more sentences as standard even when it had classified those sentences correctly in the zero-shot variant. This might be due to the wide definition of standard English and by providing examples, we limited the scope of a dialect even though it might differentiate it better from other dialects.

8 Individual Contribution

Table 8: Task Distribution

Singlish, Irish, South African and Standard Dataset	Yi Da
Indian, Australian and British Dataset	Jashwant
Prompting	Jashwant
Coding with the API	Yi Da
Report Writing	Yi Da & Jashwant

All the data was manually labelled. The prompt for classification was written by Jashwant and assessed by Yi Da. The prompt was initially tested on GPT Playground, and all subsequent testing was done using the GPT4 API. Code utilizing API was done by Yi Da . All results were stored in Excel sheets for analysis. The report was a collaborative effort.

References

[1] A. Etman and A. A. L. Beex, "Language and dialect identification: A survey," in *2015 SAI Intelligent Systems Conference (IntelliSys)*, 2015, pp. 220–231.

[2] J. Dunn, "Global syntactic variation in seven languages: Toward a computational dialectology," *Frontiers in Artificial Intelligence*, vol. 2, Aug. 2019. [Online]. Available: http://dx.doi.org/10.3389/frai.2019.00015

Jashwant Raj G

1. permission to post video: YES
2. permission to post final report: YES
3. permission to post source code: YES

Kang Yi Da

1. permission to post video: Wait to see
2. permission to post final report: YES
3. permission to post source code: YES