1 Introduction

In this project, we developed an **n-gram based language model** to predict the next character in a given text input. Our primary objective was to train a model using real conversational data and evaluate its performance in predicting the next three characters in a word or phrase.

While we initially considered using neural networks or LSTMs, we decided that n-gram models were the most applicable to our task.We started with an approach focused on yielding high accuracy for English, after which we would apply our findings to other languages. This is also because we initially had difficulties with encoding errors when faced with combinations of unicode and utf-8 encoded characters. Then, we applied our findings to other languages.

2 Data

We used the OpenAssistant 1 dataset from Hugging Face, which consists of human-generated, multi-turn conversations across 35+ of the most commonly spoken languages.

We looked into Wikipedia data dumps, Project Gutenberg, and assorted LLM training datasets, but ultimately decided on the OpenAssistant dataset from HuggingFace due to the conversational nature of the text. While Wikipedia and Project Gutenberg had much larger datasets, we felt that the often formal, academic, or literary tone of the texts were not representative of the conversation prediction task of the project.

We wanted to process the text in a way that removed irrelevant characters while ensuring the multilingual capabilities were preserved given the scope of our project: an astronaut, in space, not bound by one single country or language!

- Filtered for **English, Spanish, Russian, German, French, Chinese, Thai, Portuguese, and Catalan**
- Used UTF-8 encoding to remove emojis and non-printable characters while preserving multilingual text.
- Normalized by removing newlines, but retained whitespace, punctuation and capitalization since the model predictions were expected to those into consideration
- Analyzed the dataset to determine the average word length (5 characters), leading us to use a **5-gram model**.
- Extracted character sequences using default dictionaries and Counter objects from Python's 'collections' module.

3 Method

We used these n-grams to predict the next character by taking the last 4 characters of the input, finding their corresponding n-grams, then returning the 3 highest frequency characters for that n-gram. However, this method could not account for strings of lengths less than 5, which we initially solved using a naive random character guesser.

However, we then decided that a unigram model taking the three most common characters in the dataset would allow for a more accurate prediction than completely random guesses. At this stage, our combined unigram model and 5-gram model allowed for relatively accurate predictions on our validation set. After expanding our training data to include all the other languages, we found that the method worked on the other languages without much other tweaking.

Why Not NLTK? In terms of existing code libraries and packages, we used the HuggingFace datasets library to access our dataset and other helper libraries like json and collections to help streamline the data storage process. Rather than using common libraries such as NLTK to train our n-gram model, we used a simpler approach using default dictionaries and Counter objects from the collections library. Since the Counter object is optimized for faster counting, we believed that this would reduce the processing speed in comparison to using a built-in function from NLTK. In addition, by removing the dependency on NTLK, we were able to reduce the processing speed of our model because we didn't have to account for the download time of the library.

4 Results

Our n-gram model achieves 54.75% accuracy on a [1]**multilingual validation dataset**.

- **ASCII character accuracy**: 62.94% (340 samples)
  Strong performance on letter-based languages such as English, Spanish, German, French, Vietnamese, and Swedish.
- **Non-ASCII character accuracy**: 8.33% (60 samples)
  Struggled with character-based languages such as Chinese, Korean, Japanese, and non-Latin scripts like Russian, Hebrew, and Arabic.

These results highlight the effectiveness of n-grams for Latin-script letter-based languages but reveal limitations in handling logographic systems, which are not well captured by our English-trained model.

Data Statement

1 Data Source

We source our data from OpenAssistant, "a human-generated, human-annotated assistant-style conversation corpus." It is an open-source effort, and there have been over 13,500 contributors across 35 different languages (although we don't use them all). The dataset is hosted here:

`https://huggingface.co/datasets/OpenAssistant/oasst1`

There are also links to their paper.

2 Data Size and Processing

Data Size: 161,433 total messages (91,829 prompter, 69,614 assistant)—8,576 messages were synthetic. Language information breakdown below.

Preprocessing: From this dataset, we only use English, Spanish, Russian, German, French, Chinese, Thai, Portuguese, and Catalan messages, normalize all newlines to be spaces, and remove any characters that aren't utf-8 friendly (i.e. emojis).

3 Language Breakdown

- English (42.8%)
- Spanish (31.4%)
- Russian (5.7%)
- German (3.6%)
- French (2.9%)
- Chinese (2.5%)
- Thai (1.9%)
- Portuguese (1.7%)
- Catalan (1.6%)
- Other (5.8%)

4 Curation Rationale

We wanted to best match what the test dataset would be, so removing newlines and non-utf-8 characters allowed us to modify the data to best match the test data. We thought that this dataset in general would be a good fit because the astronaut is interacting with others via text messages mimicking a conversational style.

5 Contributor Demographics

A survey of 270 of the 13,500 contributors was conducted showing that 89% of contributors identified as male, had a median age of 26. 70% of the contributors use AI tools on a regular basis and 40% had fine-tuned their own models, showing that there was a strong familiarity with AI/NLP among participants. 63% were fluent or native speakers of English, only .7% had no proficiency in English (although I assume this survey may not be representative of true statistics due to potential selection bias, as the survey was a google form in a discord channel, presumably in English only). Due to this selection bias, it is potentially likely that top contributors, who are probably more likely to respond to a feedback form, could be over-represented in this form, which is especially important given that the top 12 contributors (of 13,500) accounted for a whopping 10% of the data. No data on contributor's race/ethnicity or economic background. All information found from their paper, more details here: https://arxiv.org/pdf/2304.07327

6 Speech Situation

Written text of assistant-style conversation mimicking human-chatbot interaction. Contributors were told to be polite and have a "friendly and approachable manner," make text readable (i.e. proper paragraph breaks), and keep a consistent tone when speaking as the AI assistant. When writing prompts, contributors were told to have variety, including varying formatting and "degrees of politeness."

7 Text Characteristics

The dataset specifically aimed for variety by encouraging contributors to "ask questions that reflect real-life situations and needs," "ask questions that might be directed towards search engines or specialists," and "make requests that encourage lateral thinking and/or require specialized knowledge." This created a wide variety of topics and information from biology and cooking to quantum physics and music—more information in Figure 11 of their paper.

8 Ethical Considerations

There were some pretty good guidelines provided to the contributors and moderation is performed. However, moderation can never be perfect, so there is potential for some content, privacy, and safety

issues even while the large majority of that information is flagged or filtered. I think that there is a bigger concern about potential biases in the dataset. At least from the survey, the contributors are almost exclusively male and have (or are pursuing) college degrees. This is one of the biggest open-source datasets on assistant-styles conversations, so many models are probably trained on this. This demographic is not representative of everyone, and so by having a dataset without a good amount of diversity we could be perpetuating bias.

References

1. We tested our model using 400 samples from the multilingual validation dataset shared by classmate Haoquan Fang on EdStem.

   https://edstem.org/us/courses/70783/discussion/6295572