CSC Recommendation Chatbot

By Nithya Chandran, Casey Ng, and Brandon Chan

Introduction

We are interested in utilizing methods in machine learning and KDD to design a chatbot to accurately converse with a user, recognizing textual patterns in user input and responding with necessary information. More specifically, for the purposes of our chatbot, we want to be able to help recommend courses for CSC majors at Cal Poly to take. Our goal is to create a retrieval-based model that is trained to provide the best response from a database of predefined responses.

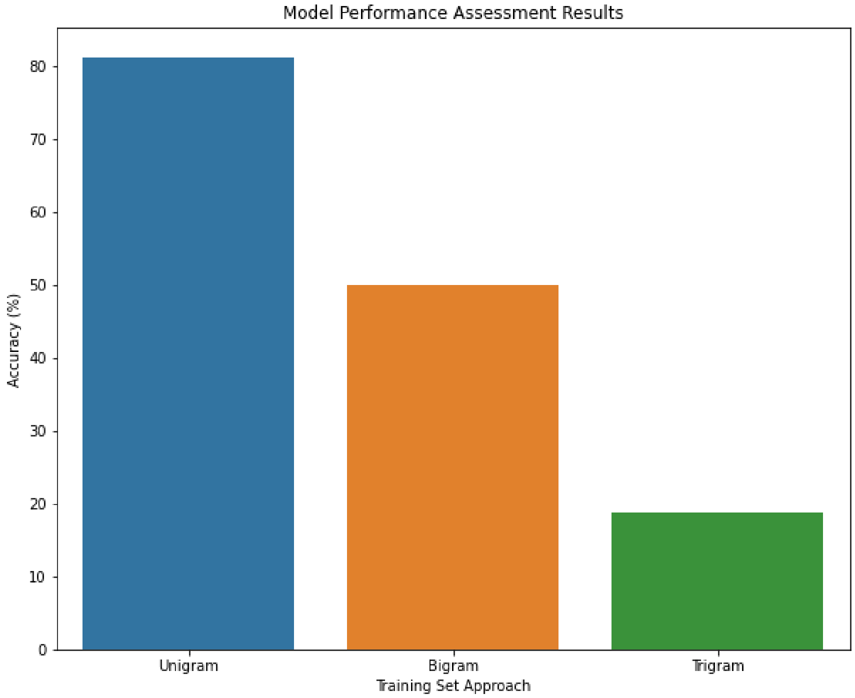
Methods

Because there is no readily available data on chatbot behaviour for Cal Poly-specific CSC course recommendations, we had to create our own data set. The data set was created in a JSON format via Google Colab. The data contains four general areas, basic chatbot behavior (greetings, goodbyes, blank responses, etc.), course recommendations, technical electives, and prerequisites. Recognizable patterns and recommended responses with corresponding tags were then written based on the information from the official Cal Poly CSC department flowcharts and course offerings pages.

The training set was created using three different approaches: unigram, bigram, and trigram. Three special recurrent neural network (LSTM) models were created. Each model was trained with a training set created from either the unigram, bigram, or trigram approaches. A testing set was also created using a list of prepared user inputs with expected response tags, mimicking a conversation. Accuracy was then determined based on whether or not the correct response tags were chosen by the model.

Results

Model accuracies were determined based on the proportion of correct response tags returned by each model based on the testing set. The unigram model had an accuracy of 81.25%, the bigram model had an accuracy of 50% and the trigram model had an accuracy of 18.75%.



Discussion

Comparing the performances of our models, the unigram model had by far the best performance of any of the models. This is likely because the smaller the sequence of words, the more easily the model can detect the patterns in the correct tags from the data. This is evident in the extremely low performance from the trigram model which searches for patterns based on sequences of three words, but much of our testing data, as is with natural conversation, contains user inputs with less than three words. This causes the trigram, and to a lesser extent the bigram model to be very inaccurate, especially when investigating patterns with fewer than three and two words, respectively, in the input sequence.

Some limitations of our model and our entire study certainly lie within our data set. Given that we had to create the entirety of the data set on our own, we were limited in the sample and scope of what ideas and topics we could cover. This in turn limited the range of topics our chatbot could cover. The limitations in our data also could very well have led to some of the inaccuracies in our model as we may have simply lacked an adequate variety of patterns in each tag for recognition from our models.

Some other additions to our project could be that we could've filtered out stopwords during the training of our model, which may have increased the accuracy of our predictions. We also could have lemmatized words to reduce words to their root meaning and decrease the likelihood of duplicate words. We could have tried a TF-IDF approach to creating the training set. And finally, we could have tried developing a model that enables the chatbot to remember past inputs to improve its responses.

References

Cal Poly CSC department flowchart
CSC degree requirements
CSC course catalog

Hello
Good

```
I need help deciding on what CSC classes to take next quarter
I am a Senior and I need help for Fall
Can you tell the prereqs for those classes?
CSC 141
CSC 445
What about some tech electives?
My concentration is Game development
cool thanks
see ya
quit
```