NLP Project Part 3: Data Collection
Grant Fass; Nick Kaja

Report

For this project, we will be training a language model to complete sentences. Sentence completion is a task that is used to get an idea of how well a language model can respond to questions that language researchers have deemed important. Many of the models we researched in existing documentation were trained on very large and broad sets of documents. For example, one of the models was trained on just over one million words from newspapers.

We have elected to gather our own data for this project. Because this task measures how well the language model can respond to important questions, we wanted to use a more educational domain for training. We will be using a dataset composed of all the lectures for five curated domains on the Khan Academy website as of January 2023.

We originally collected this dataset for use in our senior design project. That is why it was collected four months ago. This data was collected by scraping the site using pythons BeautifulSoup library. We judged the early schooling lectures to not be important for our task because they would not contain complex enough language. For this reason, we restricted the Khan Academy domains to “Math: High School & College”, “Science”, “Computing”, “Arts & Humanities”, and “Economics”.

We collected the data to pass into BeautifulSoup using the python html requests library. We began by using the requests library and BeautifulSoup to create the URI links to the individual lectures. We then used requests and Beautiful Soup to get the transcripts for the videos. These transcripts were then parsed using the BeautifulSoup html parser. We used this same process to grab all the course tags, unit tags, lesson tags, and video tags. This lets us

determine exactly where each video comes from and what it is related to. We then saved this information to a CSV file for easy use.

We plan to read in the CSV file of our training data using the python pandas library. This will involve reading the data into a DataFrame. This allows us to easily parse and mask the data as needed. It also allows us to easily use TQDM for progress bars on any operations we need to apply to all rows of the data.

We plan to make some modifications to our training data before using it in our sentence completion task. First of all, we will need to perform some cleaning related to this dataset specifically. These transcripts often include tags or names indicating who is talking at the beginning (such as [Instructor] or [Narrator]) so these will need to be removed, likely with regex.

In addition, two more modifications we plan to make involve running named entity recognition using python SpaCy library and running word tokenization using python NLTK. This will be done to help combine terms such as New York City. The tokenization would then be done to assist in the creation of our N-grams. Tokenization using NLTK also has the advantage of tokenizing punctuation.

We plan on training a N-gram model for our early prototypes. Once we have determined that the model works as expected we will attempt to implement alternative strategies to help improve performance. The early model will utilize N-grams to represent the sentence passed as input. We can then get the probability of each possible N-gram to replace the blank word in the sentence. We can then extract the missing word from the highest possibility N-gram. This will give us a simple implementation of a sentence completion model.

We plan on testing, and measuring our performance, on the Microsoft Sentence Completion Challenge Dataset. This dataset is no longer published on the Microsoft site but can still be found on Kaggle and GitHub according to StackOverflow. If we have issues using this dataset we plan to attempt to find and use the SAT questions dataset used in one of the research papers reviewed last week.

Raw Document

- [Instructor] For those of you who are just starting to learn about the history of China in the first half of the 20th century, it can be a little bit confusing. So the goal of this video is really to give you an overview, to give you a scaffold, of the history of the first half of the 20th century in China. So as we go into the early 1900s, you have the end of imperial dynastic rule in China. This is a big deal. China has been ruled by various dynasties for multiple thousands of years. But as you get into the 1900s, the dynastic rule, in particular the Qing Dynasty, was getting weaker and weaker. It had suffered at the hands of the Japanese during the first Sino-Japanese War at the end of the 1800s. There was growing discontent amongst the opposition that the dynasty, that the emperors, were not modernizing China enough. Remember, this is the early 1900s. The rest of the world was becoming a very, very modern place. China in the 1800s had suffered at the hands of Western powers who were essentially exerting their own imperial influence in China. Many people felt that this was because China was not as modernized economically, politically, technologically as it needed to be. … (rest of document truncated)

Processed Document

(Note: this doesn't visualize tokenization)