MAIS 202 Winter - Final Project Deliverable 2

Problem Statement

The goal of the project is to input a question given a paragraph of text (context) and to predict corresponding segment of text that answers the given question. This project focuses on NLP methods so that the computer can “understand” the input questions and text and produce an answer that is fitting.

Dataset preprocessing

The dataset used for this project is the SQuAD v.2 training dataset, which contains article paragraphs from wikipedia as well as question and answer sets for each paragraph. The data is grouped by article, and each article is divided by its paragraphs, providing the context for a given questions and answers set. For each question and answer set, there is also the corresponding id, and whether or not the question has an answer. Since the model bases itself on two inputs (question and context) for a given output (answer location in the context), I preprocessed the data by obtaining a list of questions, the corresponding contexts, and a list of corresponding answer indices for each question. Then, I had to vectorize the inputs. In this case, a bag of words is not sufficient because we want to make sense out of the question, not merely which topics the question could pertain to. The same applied to word2vec and GloVe, since they do not consider the relationships between the words and context. A pre trained BERT was implemented to tokenize the text inputs and then vectorize them, but there are still some bugs to go over.

Training method

In my data selection proposal, the model I envisioned was not realistic for a task of this nature, where the computer has to understand a given input. My current aspiration is to implement RNNs that take two inputs, the question and its associated context. The output will give the answer indices within the context. I plan to do a 80/20 split for training and validation, respectively. They will be based on the training SQuAD data. As for the testing part, that will be done with the dev. dataset from SQuAD.

Next steps

My next steps will focus on debugging my BERT vectorization and implementing the model. I will also need to adjust my input and expected output sets to factor for questions without any answer to them. I will also be trying to find a new way to structure my input data because it is a bit too complicated right now: the input questions and contexts and the expected answers are comprised of nested lists.

References

Park, Do-Hyoung, and Vihan Lakshman. Question Answering on the SQuAD Dataset. Stanford University, web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761899.pdf.

Nicholson, Chris. “A Beginner's Guide to Word2Vec and Neural Word Embeddings.” Pathmind, pathmind.com/wiki/word2vec.

Kana, Michel. “BERT for Dummies - Step by Step Tutorial.” Medium, Towards Data Science, 26 Oct. 2019, towardsdatascience.com/bert-for-dummies-step-by-step-tutorial-fb90890ffe03.