Chapter 1

Discussion

This project was focused on understanding how a large language model such as GPT-2 XL after being trained for a task such as knowledge graph generation. In order to understand the differences between the two models, we compared their performances for the GLUE benchmark set and arrived at a mixed conclusion.

For the majority of the GLUE tasks, the models behave very similarly and have comparable performance. The only task where COMET model seems to have a lower performance was with the Corpus of Linguistic Acceptability task. In this case, the naive GPT-2 XL seems to perform better by a noticeable margin. Upon analyzing the predictions for this task, we noticed that COMET model seems to struggle with sentences in the past tense.

This points us in the right direction for conducting further research. Given the time constraints and the computational cost of training and testing LLMs, we had to limit this project to set of GLUE tasks.

In the future, it would be valuable to understand how newer models such as GPT-J and GPT-3 change after they have been trained for knowledge graph generation.