COE379L - Project 04 Proposal
Martin Infante, mbi244
Version History:

- Predator-Prey Danger Detector deployed on Multi-Model Inference Server
- Danger Proximity with Nearest Civilian Validation in response to Armed Threats

---

From Project Ideas section on Project 04 Assignment

Perform sentiment analysis, text summarization or other classical NLP tasks on commonly available datasets such as social media postings, product reviews, articles or papers, etc. What model(s)/techniques will you use? You might consider using LLMs/transformers as part of this project. How will you evaluate the model?

Fine-tuning a BERT Transformer for Better Punctuation Accuracy on Auto-Generated Captions

Introduction

My proposal focuses on improving YouTube's auto-generated captions by fine-tuning a BERT transformer model to correct punctuation errors using Hugging Face's YouTube Caption Corrections dataset. An evaluation script will quantify an accuracy score–rooted on the subset of punctuation-based errors found in the original auto-generated captions–by comparing model outputs against human-corrected captions.

Datasets

Hugging Face's YouTube Caption Corrections dataset

- *default_seq,* lists the auto generated captions
- *correction_seq,* lists the manually corrected word at index if *diff_type* at said index list a value > 0
- *diff_type,* lists a number between 0-8 based on error produced by auto generated captions.

For the context of this project, *diff_type = 2* is the only case to be considered, as this describes punctuation differences between auto-generated and corrected captions. A BERT model will then be trained to find a better punctuation to use, rather than the incorrect one. So for our use case, items in *default_seq* will be masked at indexes where *diff_type* has a value equal to 2. This modified string list will be fed to the BERT model to notify it that the original punctuation at index was incorrect, and should make an inference on the correct punctuation to use.

Technologies

BERT base model

- Pretrained via MLM

Products to be Delivered

The final deliverables include a fine-tuned BERT model specialized in punctuation correction for YouTube captions, and an evaluation report detailing performance improvements of the BERT model for caption generation.