Prompt-Based NLP

HW4 Written Report

Samantha Ryan-Lee

Pattern-Exploiting Training

Pattern-Exploiting Training (PET) is a useful technique when training datasets are limited. One can develop patterns that can occur before, after, or in the middle of one or two series of text inputs. Then a large pretrained language model can be leveraged to generate a word that is flagged as belonging to a particular class (verbalizer). Based on these verbalizer generations in the context of the input text, a data example can be classified or labeled to train the model. In the case of this assignment, we trained 10 different PET models for a series of data samples (10, 50, 100, and 500 data examples). Each pet model has a specific pattern and set of verbalizers which correspond to the binary labels non toxic or toxic.

Example:

- Text input: “I have 10 pictures of armed gaddafi loyalists (mostly from 2013-2014) they are all taken from facebook/YT videos would they be classified as fair use?”
- Pattern: text input + "I can't believe how " + <masked token> + “you are!”
- Verbalizer - negative class (nontoxic): “helpful" or “clear”
- Verbalizer - positive class (toxic): “stupid” or “horrible”

Depending on what the language model predicts for the masked token, based on the context provided by the text input and the pattern, the example will be classified accordingly. Using the above example, if the language model generates the word “helpful” to fill in the blank (so to speak), the example is labeled as the negative class. Then, during evaluation, the examples are examined for how they are correctly or incorrectly classified based on the PET model training.

Comparisons of Evaluation by Model Type

For this assignment, we were tasked with comparing training a large pretrained BERT based language model on nearly 160,000 comments. Then, we sampled this data in various increments (also known as data instances) and developed 10 different patterns for the PET model. Each data instance was trained on the individual 10 models and macro F1 scores were calculated. These F1 scores were then compared to the overall F1 score of the trained BERT based model (also known as the MiniLM model). Both the PET and regularly fine-tuned large language model have the same foundation - which can be found at https://huggingface.co/microsoft/MiniLM-L12-H384-uncased. Based on the scores seen in the bar chart, training a PET model on only 50 instances of data can achieve near (or even greater than) the performance of a regularly trained BERT based model. The

performance depends on the type of pattern used, with some (pattern 4, for example) achieving even higher performance than the generally trained miniLM model. **In general, it seems that training on 100 data examples can consistently achieve a similar performance regardless of which pattern is used.** At the same time, by training on 500 instances of data, **all** of the individual PET models achieve greater performance than the regularly trained miniLM model. At this time, I did not examine how the use of all patterns in training each instance impacts the PET model performance for each data instance, however this is something that can be explored in the future.