Project Proposal

Daniel Jin, Qingwei Meng, Zhihan Qin, Hyejun Shin

Due: March 13, 9:30 AM

Part 0: Collaborators (Written, 0 Points)

Daniel Jin, Qingwei Meng, Zhihan Qin, Hyejun Shin

Part 1: Research Question (Written, 15 Points)

Briefly introduce the research question you wish to answer in your project. Include the following information.

1. What question are you trying to answer, or what problem are you trying to solve?

   Like humans, LLM are also vulnerable to generate unlikely story with priming, but not ungrammatical sentences. And in this paper we explore different prompting methods in combination with linguistic phenomenon such as island, gender agreement, "meaningless" sentences... We want to evaluate LLM's generated text's grammar/structures are influenced by priming.

2. Why should people care about this question/problem?

   This test give us more insight on how LLM predict tokens in relation to it's context, and it provides linguistic insight for some formalism on priming in relations to types of unacceptable sentences.

3. What do we currently know about your question/problem? It is well know that modern LM are capable of generate grammatical text, and some newer ones are even capable of identify the grammatical errors with relative high accuracy. And we also have many experiments to show priming can affect models' token generation.

4. Identify something we currently *do not* know about your question/problem, and propose some method for investigating it.

   We don't fully understand how different types of priming affect the generation of complex or pragmatically odd sentences across various models.

5. Why do you think your method would work? We have done simple testing on gpt-2 with manually generated test case, and it has shown that gpt-2's behavior aligns with human instinct. So we want to explore more on the topic with different models and different linguistic phenomenons.

Part 2: Background/Literature Search (Written, 15 Points)

1. https://aclanthology.org/2020.tacl-1.25/
   BLiMP: The Benchmark of Linguistic Minimal Pairs for English
   The paper provides minimal pairs for accepted and unacceptable sentences, this paper serves as a benchmark and base dataset for our experiments.

2. https://doi.org/10.1162/tacl_a_00612 this paper provide us a methodology that can be used to evaluate how "expected" each token is for the languge model and this also

3. https://arxiv.org/abs/2010.05465 arXiv:2010.05465 [cs.CL] Kim's paper provides a few linguistic analysis on how LM understand language structures, we will test them on different models.

Part 3: Methodology (Written, 15 Points)

Describe two experiments that you plan to conduct in order to answer your research question.

We utilize the BLiMP data set and models like BERT and GPT-2. To compare the difference on the unacceptable word with it's minimal pair under different priming. And calculate the differences of the score on the token. During the process it is important to ensure that the priming makes sense in relation to the sentence in question. And possibly we need to generate more sentences by using CFG(provided with BLiMP)

For example, it is unlikely for someone directly says:" I ate an ANT" or "I ate an BANANA*". Where eating an ANT is acceptable but surprising, but "an BANANA" is unacceptable but maybe less surprising. But if we preceded the sentence change them into "*Let me tell you a wild story,* I ate an ANT." "*Let me tell you a wild story,* I ate an BANANA*." It is still unlikely for the second sentence to ever be said, but the first sentence became much likely to occur. And there are similar behavior for models to give tokens a higher score if it is intuitively expected by human, giving the word ANT a higher score than BANANA.

After the evaluation in a large scale, we will pick some representative cases, and evaluate the list of other possibly generated tokens that the model given a high score in the position of the token in question. We can give a more detailed evaluation on how the priming affected the ordering of the tokens to see what type of token gets affected in each linguistic phenomenon.

Part 4: Required Resources (Written, 5 Points)

Only HPC might be needed on large data set.