Experimental Protocol: Question Generation with Few Shot LLMs

Nitya Sampath
sampathnitya@gmail.com

1 Hypotheses

My project aims to investigate the use of LLMs (specifically, GPT-3 and Llama2) and few-shot learning on the task of question generation. My core hypothesis for this project is as follows:

(1) Llama2 will outperform GPT-3 on the task of question generation with both zero-shot and few-shot prompting.

2 Data

I will be using the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) for this project. SQuAD consists of over 100,000 question-answer pairs generated by crowdworkers on Wikipedia articles. The answers are spans from the corresponsing text.

3 Metrics

I want to use BLEU (Papineni et al., 2002) for evaluating the results of my models. BLUE very common metric used in generation tasks. BLEU is a machine translation metric that evaluates the similarity between generated translations and a set of high quality translations. The BLEU score is a number between 0 and 1 that represents the similarity of the prediction to the reference passage, with higher scores being more similar.

I also want to run a qualitative analysis of the results, to evaluate and compare the models' ability to generate questions of different types.

4 Models

My baseline models will be two GPT-3 models: one zero-shot and one few-shot. My focus will then be on two Llama2 models, one zero-shot and one few-shot.

5 General Reasoning

For this project, I will be comparing the performance of two LLMs, GPT-3 and Llama on the task of question answering using few-shot learning. Both GPT-3 and Llama2 are some of the most popular LLMs used for natural language tasks. Question generation is a task that requires taking a context passage and an answer and returning the question asked. This is a common natural language task that can be used to, for example, increase training datasets for QA models.

I want to compare the models' performances using both zero-shot and few-shot learning. I will be using two GPT-3 models as baselines: one zero-shot model and one few-shot model. The focus of my project will then be zero-shot and few-shot Llama2 models. I will use examples from the SQuAD dataset to provide demonstrations for few-shot learning.

I will evaluate the results of the models using BLUE and do a qualitative analysis and comparison of the questions generated by the models.

6 Summary of Progress

As of now, I have completed the necessary research on previous work and datasets. I have set up access to the OpenAI API so I can use GPT-3. I have explored the SQuAD dataset and loaded the training set for use in few-shot learning. I have tested baseline zero-shot and few-shot GPT-3 models.

Going forward, I need to set up Llama2 so that I can use it to run my experiments. I need to finish setting up the models, run the experiments and perform the analysis and comparison.

References

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.