Project 17 - Instructions Paradigm : An Alternative to Crowdsourcing

(CSE 576 FALL 2021 - Individual report)

1. Problem statement.

The objective of this project is to check the hypothesis - "Can the instruction paradigm help in replacing humans with machines for dataset creation?". Thinking of this hypothesis is only made possible by the SOTA language models like GPT-3 and T0_pp that we have today.

An example to prompt the T0_pp model:

"Generate a premise." - sentence1: A woman narrates the details of a recent conversation that she had with her boyfriend.
"Generate a hypothesis from the premise: {sentence1}" :The man she is talking to is her boyfriend.

2. Motivation.

Answering the hypothesis in the problem statement helps us to determine if the instruction paradigm help the NLP community by

- Eliminating bias by human intuition.
- Saving labor cost and overheads by making dataset creation computation based.
- Makes the dataset easier to clean, aggregate and maintain.

3. Prompt design.

I used discrete prompt mining and prompt scoring methods where I took references from the original dataset to design a prompt and paraphrased it until the results were satisfying. I have further experimented with prompt augmenting using few-shot learning. I mainly worked on QQP, MNLI, and CoLA datasets, where I was able to get a presentable result. I experimented on GPT3_davinci-instruct-beta-v3 for few-shot prompt learning and T0_pp for zero-shot learning and resorted to T0_pp because it had similar/better results for tasks at hand and was open source.

3.1 The Quora Question Pairs (QQP).

The QQP dataset has 3 fields question1, question2, is_duplicate - Indicates whether these 2 questions mean the same[1].

3.1.1 Dataset Creation:

I had to look at the original dataset for prompt engineering. One of the major findings is that a positive case is not always a pair of paraphrased questions. Positive pairs can also be relatively related questions. Example: [*334577, {My questions are fine. Why does Quora keep saying they need improvement? It's frustrating.}, {All my questions on Quora need improvement. What is the best way to ask a question on Quora?}*][1]

Points on experiments and dataset creation:

- I experimented on the ***GPT-3 davinci-instruct-beta-v3*** model with few-shot prompting and T0_pp for zero-shot prompting. GPT-3 tends to follow the example template very closely and gives too many duplicates so I went with T0_pp for better diversity.

Prompt for T0_pp was trained in steps. First I generated a random question, and then I paraphrased/ generated a related question from the first.

1. GPT-3 example: ***Generate a question pair. Example: How do you remove spray paint from a mirror?, What is the best way to remove paint from glass?*** *How do you remove spray paint from a mirror? What is the best way to remove paint from glass?*
2. T0_pp example: ***Generate a random long question.*** *What is the name of the fictional country in the Harry Potter stories?* ***Generate a question on: {q1}.*** *Where is Harry Potter located?*

We can see that T0_pp not just paraphrases the Q1, but generates a related question with the same answer. So I used T0_pp.

- For the Negative case I just generated 2 random questions.
- I ended up with ***13K*** Question Question pairs. Further to preprocess these questions I filtered the dataset to remove duplicate questions, sentences that didn't end in '?' , and sentences that got trimmed because of API character limitation. In the end, I got ***7.5K*** rows of data.

3.1.2 Evaluation:

The idea is to train 2 models, one on the original dataset and another on the prompted dataset. Then I evaluate them on a test dataset from original data.

- I used fine-tuning on pre-trained model ***all-MiniLM-L12-v2*** [4], I experimented on binary cross-entropy loss and (OnlineContrastiveLoss + MultipleNegativesRankingLoss) latter resulted in better results on both original and generated dataset.
- Hyper parameters: *lr: 2e-05, weight_decay: float = 0.01, batch_size: 64, epochs: 30*.
- I have used a sentence transformer evaluator to generate the evaluations using the evaluation dataset from the original dataset.

Dataset	Cossim acc	Cossim f1	Cossim recall	Cossim precision
Original	0.82	0.77	0.85	0.7
Prompted	0.77	0.73	0.85	0.63

3.1.3 Key FIndings:

From the results, we can confirm that the dataset created from the zero-shot prompting on T0_pp had similar distribution as the actual database and with further finetuning/ training we can further improve the quality of the dataset.

3.2. MultiNLI.

The MNLI is a dataset with [sentence1-parse_tree, sentence2-parse_tree, sentence1, sentence2, label - with value entails, neutral, contradiction indicating relation between sentence1, sentence2][2], we are interested in only 2 of the sentences and label. Example: [*Sentence1: There are obstacles for innovation that the participants recognized.* Sentence2: *However, the participants recognized that one of the big obstacles for innovation in*, label: entailment]

3.2.1 Dataset creation:

I used prompt mining and scoring to determine the best prompts. From the training examples, I could infer that i. Premise can be short, ii. The hypothesis can be paraphrased. iii. Further looking at the dataset, domain-specific dataset premises were belonging to normal article premises and conversational premises.

With the above findings I experimented by:

- I again found that T0_pp was better than GPT-3 in generating the premise and hypothesis. GPT-3 tends to follow the examples closely, even with changes in model prediction temperature.
- The prompts to T0_pp was zero-shot and had a combination of generating a long/short/conversational premise, generating a random text from the premise (Example: **She asks if the man in the bar was talking to someone in a bar. *very funny people who work at a bar. *neutral*), generating a hypothesis from the premise (Example: **A cyborg has been created by melding together the DNA of five different cyborgs. It has no memories and no awareness of its surroundings. Its creator believes that it is the perfect soldier because it can be trained to do anything. However, the cyborg has a stutter and does not understand what he is saying. The cyborg kills several of the human soldiers and uses the experience to improve his fighting skills. It is now up to his creator to decide if the cyborg should stay in the military or be sent back to his home world. *A cyborg has been created by melding together the DNA of five different cyborgs. *entailment*) and contradicting statements of a hypothesis from the premise (It was not possible to generate contradicting hypotheses directly).
- I generated around ***30k samples*** and ended up with ***21k samples*** after filtering out duplicates and short premises/hypotheses.

3.2.2 Evaluation:

The idea is to train 2 models, one on the original dataset and another on the prompted dataset, and evaluate them on the STS dataset (as I trained on multiple negative loss training functions).

- I used fine-tuning on pre-trained model ***all-MiniLM-L12-v2***[4], I experimented with softmax loss function and label accuracy evaluator which resulted in a very low evaluation metric for original accuracy. So instead I used multiple negative rankings and ignored neutral samples and augmented the training dataset with premise/hypothesis and hypothesis/premise pairs.
- The model was evaluated using a scaled STS dataset (0-1 scores) to evaluate the scoring of texts entailment in the model. This resulted in the following metrics.

Hyper parameters: *lr: 2e-05, weight_decay: float = 0.01, batch_size: 64, epochs: 30*.

Dataset	Cosine Pearson	Cosine Spearman
Original	.84	.839
Prompted	.83	.82

3.2.3 Key FIndings:

From the Pearson and Spearman correlation values we can see that the model has learned to score similar and dissimilar hypotheses/ premises.

Although the prompted dataset gave good results, examining the actual dataset reveals few characteristics of the model used for prompting. Generally, the T0_pp model failed to generate a premise when the hypothesis was presented/ended with a question, it rather tried to answer the question [Example: **Why is the premise of the argument that women have a harder time in gaining employment true?, *Women have a harder time in gaining employment than men. *entailment*], a few data rows also reveal that if the premise is short, the hypothesis is just an addition/ paraphrase of a few words. [Example: **A thigh high bar is my. *A thigh high bar is a bar in my.*]. I left these data rows as they represent the drawbacks of this paradigm.

3.3 The Corpus of Linguistic Acceptability (CoLA)

CoLA includes a dataset where each row has a sentence and 2 annotations indicating whether these sentences are acceptable grammatical sentences belonging to the English model or contain mistakes. The annotations are in front of the experts and the original author of the sentence/article[3].

This dataset was by far the hardest to prompt engineers, and I failed to come up with a mechanism to generate this kind of data from prompting SOTAmodels like GPT-3 and T0_pp.

3.3.1 Dataset creation experiments:

- The negative cases in the dataset where there are grammatical errors are more semantical grammar related albeit they were few syntactic examples. This is the major hurdle in prompt engineering, I couldn't identify/ aggregate all the semantic error types to try a few shot/one-shot learning.
- I tried one shot and a few shots prompting ***GPT3 'davinci-instruct-beta-v3***' and zero-shot learning with T0_pp. The experiment included, i. Trying to generate an ungrammatical sentence with examples (GPT3) and without examples (T0_pp), ii. Given a prompt to generate grammatically correct sentences and then induce grammatical mistakes with and without examples (multi-steps). In all the permutations of the prompts and models, I couldn't get conclusive evidence that a prompt could generate ungrammatical sentences.
- Experiment examples: [GPT3 - Few-shot learning: "***Generate an ungrammatical sentence. Example:Bill reading Shakespeare and Maureen singing Schubert satisfy me. Example:A pastor was executed, notwithstanding on many applications in favor of him. Example:What exploded when and I warned you it would?*** *Example:The one that is related to the sentiment of pity is pleasing to me. Example:A picture of the boy striking a viper, it was exhibited at the Royal Academy. Example:I have a feeling that I am an item of a series**.".*** One-shot learning – ***Generate an ungrammatical sentence. Example:George is having lived in Toledo for thirty years.*** *When this sentence is fed into the parser, it will generate a syntactic parse tree as a sentence structure. In the example of George is living in Toledo for thirty years, the parser generates a sentence structure as: S (S (NP (D George)) (VP (V is) (NP (NP..*]

- To emphasize how semantically skewed the dataset was, I create the script to introduce 15 grammatical errors (including noun replacement, pronoun mismatch, tense change, adjective change, random capitalization, wrong punctuation, etc.) randomly to create a dataset of ***31k samples.*** I created a pipeline for this task based on the *bert_base_uncased* encoder followed by a fully connected mlp trained on the same model first on the original dataset and second on the manually generated dataset and used the original test set, I got 78.97% and 39.94%.

3.3.2 Key findings:

I was not able to generate a dataset that had similar distribution as the CoLA dataset. Manually created datasets with syntactical errors did not belong to the actual distribution from the evaluation report.

4.0 Conclusion:

In this project, I have experimented with using an instruction-based paradigm to generate datasets belonging to QQP, MultiNLI, CoLA. I have found that the prompt paradigm works well when the data belongs to the general language model and ground truth can be extracted/labeled during dataset generation (SST, QQP, MNLI, etc,.).

I have used variants of GPT-3 and T0_pp models for these experiments. GPT3 works well to extract information from generated data with the help of few-shot learning, whereas T0_pp is good at zero-shot prompts to generate data belonging to a specific domain and manipulate the existing data(QQP) and I preferred T0_pp for this undertaking. Furthermore, I experimented with pipe to evaluate the prompt-based datasets generated for QQP, MNLI, and manual datasets for CoLA. Although QQP and MNLI showed promising results upon closer look at the actual dataset we can say they had their drawbacks (Ex: I had to filter out almost 30% of the data from QQP, as the paradigm failed to generate related questions instead it answered them, which could be due to T0_pp being trained on QA dataset and mistaking the prompt/task at hand.) I also noticed that this Instructional prompting approach does not work very well with datasets like CoLA as writing prompts for generating semantically wrong text are harder and often the result is left to interpretation, whereas it worked very well for QQP (where I could filter dataset easily).

To answer the hypothesis “Can the instruction paradigm help in replacing humans with machines for dataset creation?”. For simple tasks indeed it can, but for complicated tasks human inference of the language outweighs the LM’s ability to generate coherent sentences.

5. References

[1]. Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. First Quora Dataset Release: Question Pairs. https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

[2]. Williams, Adina, Nangia, Nikita, Bowman, and Samuel. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. 2018. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics. http://aclweb.org/anthology/N18-1101

[3]. Warstadt, Alex and Singh, Amanpreet and Bowman, Samuel R. 2018. Neural Network Acceptability Judgments (Corpus of Linguistic Acceptability). arXiv preprint arXiv:1805.12471, warstadt2018neural.

[4]. Reimers, Nils and Gurevych, Iryna. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. 2020. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. https://arxiv.org/abs/2004.09813.