MDDial - Developing Prompting Techniques to achieve Efficient Differential Diagnosis Dialogue System

CSE 576 - Natural Language Processing
Fall 2022

Project Final Report

Team Members
Sai Kousthubha Das Kalvakolanu - 1224121496
Rithish Kesav Saravanan - 1222678626
Shreyaa Dinakar - 1224103101
Priyadharshini Amudan - 1222307593

Mentors
Mihir Parmar
Neeraj Varshney

GitHub Link
https://github.com/nrjvarshney/CSE_576_Dialog

Introduction

Automatic Differential Diagnosis (ADD) aims to develop a system that can diagnose patients via interaction with them. It has many benefits, such as simplifying the diagnostic procedure, reducing the cost of collecting information from patients, and assisting a real doctor to make decisions efficiently. In order to enhance the experience, leveraging language models and developing end-to-end systems is essential. To this extent, in this project, we design various prompting techniques that help improve GPT-3 performance in the domain of Differential Diagnosis and also look at how models like GPT-2 and DialoGPT perform when trained on a dialogue dataset for Differential Diagnosis. We compare and contrast GPT-2, and DialoGPT performance with GPT-3 to demonstrate how traditional transfer-based learning approaches scale in comparison to prompting techniques. Finally, we see if GPT-3 can be used to generate more dialogues on which GPT2 and DialogGPT can be trained to improve their diagnosis' accuracy.

Methods, Datasets, and Models Used

In this project, we leveraged OpenAI's GPT-3 model extensively in order to apply prompting techniques (listed below in Experiments) to see if GPT-3 can do a good job in the domain of differential diagnosis. GPT-3 performed moderately when no prompting techniques were used and had quite a few issues including but not limited to an incorrect diagnosis, inconsistent diagnosis, unnecessary explanations, etc. To help resolve these issues, we devised eight prompting techniques that worked well and eliminated one or more issues that we faced earlier. The prompting techniques have used a variety of approaches with some using Few-Shot, Chain-of-Thought, Least-to-Most, and even a combination of these approaches.

Next, for evaluating traditional transformer-based model performance against GPT-3 based on prompting techniques, we chose GPT-2 and DialoGPT given their ability to predict the next word given the context. We trained these models using the MDDial training dataset and evaluated their performance using the test dataset. The results of this evaluation are provided below. The MDDial dataset contains 1879 patient and doctor dialogues in the training set and 235 dialogues in the test set. The dataset contains dialogues with diagnoses for twelve diseases namely: Esophagitis, Enteritis, Asthma, Coronary heart disease, Pneumonia, Rhinitis, Thyroiditis, Traumatic brain injury, Dermatitis, External otitis, Conjunctivitis, and Mastitis. The dataset has 155-165 dialogues for each of the diseases resulting in a total of 16,560 turns between patient and doctor.

Experiments

Experiment 1: Prompting Engineering for Differential Diagnosis on GPT-3
As described in the previous section, we began by leveraging GPT-3 to understand its diagnosis efficiency without any prompting techniques and identified issues that prevented GPT-3 from reaching the correct diagnosis. For this, we developed a series of eight prompting techniques as listed below.

1. *Few Shot Learning*
   In this technique, we tried passing a sample conversation between doctor and patient to GPT-3 as a prompt before asking GPT-3 to take the role of a doctor and perform the diagnosis. This helped GPT-3 understand the structure of the conversation and thereby GPT-3 started asking questions regarding additional symptoms that the patient might have.
2. *List of Diseases and corresponding Symptoms*
   In this technique, in addition to passing a sample conversation between doctor and patient, we also pass the list of diseases and their corresponding symptoms assuming that GPT-3 will be able to identify the symptoms and map them back to the corresponding disease resulting in a correct diagnosis in a majority of the cases.
3. *Using metaprompt as an expert generator*
   The expert generator will render a simulation of an expert to answer the question. Once the machine identifies an expert, the entire simulation of questions and answering between the patient and the doctor is based on the doctor's expertise. This increases the accuracy of the GPT-3 and also makes it more aware of the context. This method can also be task agnostic.
4. *Using metaprompt in generic metacognition*
   The metaprompt acts as a wrapper for specific questions. In this case, a short phrase is given as a prompt which will lead to a step by step questioning and relevant answering on the topic. This makes the GPT-3 more aware of the specific task and the steps that are needed to be followed to increase the accuracy of the diagnosis. Techniques 3 and 4 have made the GPT-3 ask more relevant questions and not just randomly guess the disease.
5. *Least-to-Most Prompting*
   This unique prompting approach, least-to-most prompting, addresses easy-to-hard generalization difficulties. It is accomplished in two stages: reducing a difficult issue into a list of subproblems and then solving these subproblems progressively, whereby solving a particular subproblem is aided by the answers to previously solved subproblems.
6. *Automatic Chain of Through Prompting*
   There are two basic concepts for CoT prompting. To stimulate step-by-step thinking before answering a question, one might use a simple suggestion like "Let's think step by step." The other employs a series of physical examples, each consisting of a question and a logical chain that leads to a solution. The second paradigm's higher performance is dependent on the hand-crafting of task-specific examples one by one.
7. *Multimodal reasoning via Thought chains*
   Ideology of chain of thought is used where after the initial prompt we give context that explains the approach to be taken to reach the expected result. An initial question is given which mentions what final output is expected. With this technique, it starts asking more specific questions and gives an output.
8. *Reinforcement learning with few-shot.*

This technique is the incorporation of two techniques, reinforcement learning, and few shot. First, the modification of reinforcement learning is implemented where the expected output is recollected so that the prediction results match the expected. Then we use few shot where there are a few examples of how to approach which is learned and followed by later. In this approach, we get more target-related questions and the decision is made based on the recollection and the answers given to the questions.

These prompting techniques helped us improve GPT-3 performance by eliminating the said issues and thereby ending up at the right diagnosis in the majority of the cases. The actual results of the evaluation are provided below in the Results & Analysis section.

Experiment 2: Transformer-based Model Evaluation for Differential Diagnosis
Once GPT-3 evaluation was done, we wanted to understand how a traditional transformer-based training approach would compare to GPT-3 performance. Therefore, we trained two models namely GPT-2 and DialoGPT using the hugging face Trainer API. The models were trained over the MDDial dataset for 10 epochs each and have shown considerable diagnosis accuracy. A sample input sentence given to the model includes both patient and doctor turns appended together with the [END] token. For example, consider the below dialogue between the patient and doctor and its corresponding encoding.

Dialogue:
Patient: Recently, I am experiencing Burning sensation behind the breastbone
Doctor: In that case, do you have any Nausea?
Patient: Well not in my knowledge
Doctor: Oh, do you have any Stomach ache?
Patient: Yes, sometimes
Doctor: In that case, you have Esophagitis.

Encoding:
```
Recently, I am experiencing Burning sensation behind the
breastbone[END]In that case, do you have any Nausea?[END]Well not in my
knowledge[END]Oh, do you have any Stomach ache?[END]Yes,
sometimes[END]In that case, you have Esophagitis.[END]
```

While training the whole dialogue is passed in as input to the model with attention masks hiding the part that needs to be predicted by the model.

Experiment 3: Data synthesis using GPT3 - Generating dialogues for training GPT-2 & DialogGPT.
In this experiment, we tried to use GPT-3 to generate more dialogues similar to the ones within the MDDial dataset to see if these artificially generated conversations can further increase the

diagnosis accuracy of GPT-2 and DialoGPT. We started out by generating 4-5 dialogues per disease resulting in a total of 50 more dialogues which we used to further train both GPT-2 and DialoGPT models. For generating the dialogues we provided GPT-3 with a prompt asking it to generate differential diagnosis dialogues and provided sample dialogues as reference. The GPT-2 and DialoGPT model performance trained using these dialogs are provided below in the Results and Analysis section.

Results & Analysis

The performance of GPT-3 has been most astonishing to us. GPT-3 showed performance that was way better than GPT-2 and DialoGPT models fine-tuned on differential diagnosis dataset with more than 16 thousand turns. The results of all experiments are formulated below in a tabular format.

GPT-3 Diagnosis Accuracy	**GPT-2 Diagnosis Accuracy**		**DialoGPT Diagnosis Accuracy**	
	MDDial Only	**MDDial + GPT-3 Generated**	**MDDial Only**	**MDDial + GPT-3 Generated**
88.20%	57.02%	61.27%	60.42%	56.59%

Both GPT-2 and DialoGPT have quite similar performance with DialoGPT leading with 3% higher accuracy. It is interesting to note that data generated using GPT-3 through prompting techniques helped increase the accuracy of GPT-2 while on the other hand, DialoGPT suffered when GPT-3 dialogues were introduced. That being said, these results are in no way representative of GPT-3's ability to provide well-synthesized datasets given the number of dialogues generated using GPT-3 was far less (50 dialogues) when compared to the dialogues present in the training set of MDDial (1827 dialogues).

Below you can see the disease wise diagnosis accuracies.

Disease	**GPT-3 Diagnosis Ratio (Positive/Total)**	**GPT-2 Diagnosis Ratio (Positive/Total)**		**DialoGPT Diagnosis Ratio (Positive/Total)**	
		MDDial Only	**MDDial + GPT3 Generated**	**MDDial Only**	**MDDial + GPT3 Generated**
Esophagitis	4/5	12/27	9/27	14/27	7/27
Enteritis	6/6	14/24	17/24	14/24	17/24
Asthma	4/5	9/19	9/19	11/19	9/19
Coronary heart disease	4/5	4/19	6/19	7/19	8/19
Pneumonia	3/5	2/20	5/20	9/20	6/20
Rhinitis	5/5	9/15	8/15	6/15	10/15
Thyroiditis	5/5	9/19	14/19	9/19	7/19
Traumatic brain injury	5/5	11/19	13/19	14/19	14/19
Dermatitis	4/5	20/20	20/20	20/20	18/20
External otitis	4/5	15/17	15/17	14/17	12/17
Conjunctivitis	6/6	17/21	18/21	16/21	15/21
Mastitis	6/6	12/15	11/15	11/15	7/15
Overall	54/61	134/235	145/235	145/235	130/235

Individual Contributions

Contributor	Description
Sai Kousthubha Das Kalvakolanu	- Worked on testing GPT-3 to identify possible issues that could limit GPT-3 in acting as a differential diagnosis agent. - Tested GPT-3 using patient-doctor conversations number 1-5, 21, 22 from the MDDial dataset. - Identified two techniques using a few-shots and chain-of-thought approach to help improve the performance of GPT-3 and resolve issues that hindered GPT-3 from providing the correct diagnosis. - Preprocessed the MDDial dataset to split dialogues between patient and doctor so that proper attention masks can be applied while training GPT2 and DialoGPT. - Trained GPT-2 small and medium models on the MDDial dataset to see how they perform. - Trained DialoGPT small model on the MDDial dataset to compare its performance to GPT-2. - Evaluated GPT-2 (small) and DialoGPT (small) models to see how they perform on unseen data and to derive a diagnosis accuracy for these models. - Created a combined dataset file for GPT-3 generated dialogues collected by Rithish and Shreyaa. - Link to work: Sai
Rithish Kesav Saravanan	- Worked with GPT3 giving different possible prompts, conditions, symptoms, questions, and so on for it to come up with the expected diagnosis. - Tested GPT-3 using patient-doctor conversations number 6-10 from the MDDial dataset. - Identified two prompting techniques that can be used to improve the output of GPT3's diagnosis. The two techniques were Least-to-Most Prompting and Automatic Chain of Thought Prompting. This resulted in an improvement in the diagnosis when compared to the unprompted GPT3's output. - Worked on testing the Reinforcement learning with fewshot, on 5 different diseases. - Each of the diseases are tested for at least 5 dialogs. - Contributed to the dataset creation by using the improvised prompting technique.

	- Link to work: Rithish
Shreyaa Dinakar	- Tested 5 dialogs from the given dataset in GPT3. - Dialogs worked:11 to 15 from the dataset. - Found issues after testing the 5 dialogs. - Implemented 2 techniques, Multimodal reasoning via thought chains and Reinforcement learning with few-shot, which resulted in improvement in the diagnosis. - Worked on testing the Reinforcement learning with few-shot, on 6 different diseases. - Each of the diseases is tested for at least 5 dialogs. - Contributed to the dataset creation by using the improvised prompting technique. - Link to work: Shreyaa Dinakar
Priyadharshini Amudan	- Tested the GPT3 on differential diagnosis by a doctor based on the patient's symptoms and identified the issues associated with the dialogs 16-20 from the dataset. - The two techniques that were identified were based on using meta prompt as an expert generator and using it in generic metacognition. - By identifying the techniques, the improvement achieved concerning the language model involves more context-aware questioning. - Evaluated GPT3 on the dataset created using the improved prompting technique and found the accuracy to be 88.2%. - Link to work: Priyadharshini

Future Works

- Increase the number of dialogues generated by GPT-3 to see if that increases the performance of GPT-2 and DialoGPT.
- Train both GPT-2 & DialoGPT for more epochs to see if that increases the performance.

Conclusions

Prompt Engineering proves to be a really good option to consider when designing Automatic Differential Diagnosis systems using GPT-3. The diagnosis accuracy is way better than traditional fine-tuning-based model training. Also, although not to a great extent, it is evident that GPT-3 can generate synthetic data sets which can help improve the accuracy of fine-tuned models. In addition, fine-tuned models like GPT-2 and DialoGPT can be used as a decent option

when building an end-to-end differential diagnosis system if the data of considerable amounts can be accumulated.