CS57700 Natural Language Processing
Project Report: Prompt-Enhanced Medical Question Summarization

Farhanaz Farheen (0034769559)
ffarheen@purdue.edu

Md Ajwad Akil (0036436992)
makil@purdue.edu

1 Introduction

Advances in natural language processing have propelled computers into realms of understanding, interpreting, and crafting text and dialogues with an unprecedented depth. These strides have naturally spilled over into the realm of medical data interpretation and comprehension. The emergence of formidable pre-trained language models such as T5, BERT, and GPT marks a significant milestone in NLP, as evidenced by their remarkable performance across various tasks (Raffel et al., 2020; Devlin et al., 2018; Radford et al., 2018). While the challenge of condensing vast texts within a given context is not new in NLP, exploring this endeavor through the lens of reasoning and prompting remains an active area of research ripe for exploration. In this investigation, we aim to delve into the realm of medical question summarization utilizing prompting techniques, probing how such approaches enhance the contextual understanding capabilities of language models.

The innovation in our approach primarily stems from utilizing prompts to elevate the quality of summarizing medical inquiries, employing tags or pertinent cues generated independently. While prompt-based experimentation isn't entirely new to us, the inquiry into whether we can generate more meaningful and precise summaries of medical queries using prompts crafted from NER or contextually relevant tags and co-occurrence is uncharted territory. Furthermore, the feasibility of this endeavor is underpinned by our access to a publicly available dataset and the computational resources necessary to execute our planned experiments. Comprehensive information regarding the dataset and resources can be found in subsequent sections.

Our main contributions are:

- Experimenting with Named Entity Recognition tags as part of generating prompts for enhancing medical question summarization
- Analyzing how changing prompts can impact the performance of language models
- Comparing performance of zero-shot and fine tuned models with and without enhanced prompts
- Analyzing the impact of including co-occurrence information in the prompts

In our project proposal, we planned to perform three main things: (1) Keyword extraction (2) Co-occurrence analysis (3) Adaptive prompting and experiment these using zero-shot and fine tuned models. We mentioned the dataset and the language model we would be using. We mentioned evaluation metrics being BLEU(Papineni et al., 2002) score, ROGUE(Lin, 2004)score and BERTscore(Zhang* et al., 2020). All of these remain the same in the final version of the project report, i.e. compared to the proposal, the final report is the same in terms of mentioned experiments and analysis. All our experiment codes and datasets are available at the link: `https://github.com/AJAkil/cs577_proj`. The project presentation slides are available at this link: Project Presentation Slides

2 Methods

We worked on summarizing medical questions given contexts or dialogues. The dataset was particularly challenging because whether the context are just general conversations or descriptions was not explicitly labelled. We first took the dataset and use the standard splits for train, validation and tests. Then we modified the datasts in two ways as suggested in the proposal. First we used off the shelf DeBERTA (He et al., 2020) model trained on PubMed dataset to obtain the tags from the sentences. We only took the general tags instead of the

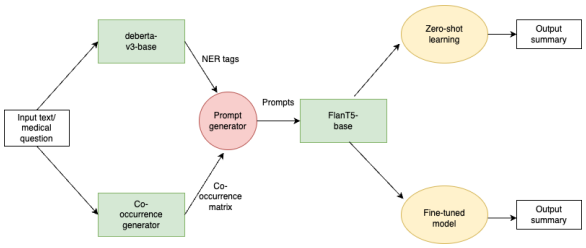

Figure 1: Architecture implemented in our study

standard BIO tags. Then we augmented our original dataset(all splits) with the tags by adding the tags at the start of the entities. For instance for the part of a sentence in the train set, the tags were set in such a way: **i have a few questions about <DISEASE_DISORDER> autonomic disorders combined with <DISEASE_DISORDER> arthritis**. After generating the modified datasets with the NER tags, we used co-occurrences of the words to create another modified dataset(for all splits). We used a sliding window based bi-gram approach to construct the co-occurrence dataset. Firstly a table of word co-occurrence was constructed from the entire corpus of train and validation set. Then using a window size 4 for each sentence and a threshold value of 5, if two words in the window has a co-occurrence value more than the threshold in the matrix, we formed the co-occurrence tags. The tags were appended to the end of the sentences. An example from the test set is: **co q10 how effective and side effects for blood pressure treatment? Cider vinegar? <effects-side> <blood-pressure>**. Then we used instruction fine-tuning to train the base version of Flan-T5(Chung et al., 2022) model. Our Prompt structure was straightforward. We provided a start prompt, followed by the question with context of the dataset and then finally a short end prompt which would tell the model to generate the summary of the question with context we provided. We used the same **"medical question summary:"** end prompt for experimenting and modified the start prompt dynamically in different schemes. We conducted summarization with zero-shot prompting, fine-tuning as a sequence to sequence model with normal prompts and also inclusion of NER tags and co-occurrence tags based dataset in combination with task specific prompts to observe how the model utilizes the tags and the given prompts. The details are laid out in the experiments and results section. The overall architecture we implemented in our study is shown in fig 1.

3 Experiments and Results

We conducted the following experiments:

- Zero-shot prompting (1) and fine-tuned prompting (3). The prompts we used were:
  (a) Summarize the following medical question.
  (b) Read through the whole context and summarize the medical question. (b)
- Zero-shot prompts with NER (2):
  (a) Read through the whole context and summarize the medical question focusing on tags supplied within <> brackets.
- Fine-tuning with NER (4):
  (a) Read through the whole context and summarize the medical question focusing on tags supplied within <> brackets. (a)
  (b) Reading the context, shortly summarize the medical question focusing on tags within <>. Focus on <MEDICATION>,<DIAGNOSTIC_PROCEDURE>, <BIOLOGICAL_ATTRIBUTE>, <SIGN_SYMPTOM>, <BIOLOGICAL_STRUCTURE>, <DISEASE_DISORDER> if present.(b)
- Zero-shot (5) and fine-tuning (6) with co-occurrence:
  (a) Read through the whole context and summarize the medical question focusing on co-occurrence of pairs of words in <> brackets separated by - appear together if <> is present after the sentence.

As it can be seen, we experimented with adaptive prompts, NER or keywords extraction and co-occurrence as planned in our proposal.

For training, we used 30 epochs with learning rate 0.0001, batch size of 16, gradient accumulation with step-size 4 and AdamW optimizer. We used Google Colab Pro with an L4 GPU with 22.5GB of VRam and Huggingface framework for all of our experiments.

As for dataset, we used the **MeQSum** (Ben Abacha and Demner-Fushman, 2019) dataset for conducting our experiments.

Our results are shown in figure 2. All the metrics were calculated upon the test set.

Exp	Name	BLEU	R1	R2	RL	RLsm	B-Pre	B-Rec	B-F1
1a	Z	1.997	16.376	5.057	14.754	14.749	86.802	85.297	85.991
1b	Z	3.096	21.298	7.400	18.349	18.346	87.442	87.675	87.511
2	Z + N	4.847	23.476	8.726	21.215	21.203	87.834	87.520	87.640
3a	F	5.245	24.282	8.978	22.558	22.571	89.258	87.861	88.533
3b	F	4.197	26.237	9.365	23.316	23.191	88.902	88.928	88.882
4a	F+N	4.236	24.214	8.028	22.214	22.307	88.817	87.839	88.298
4b	F+N	**7.061**	**27.601**	**10.742**	**26.235**	**26.231**	**90.670**	**89.117**	**89.865**
5	Z+C	4.011	22.334	8.106	19.548	19.533	87.462	87.551	87.464
6	F+C	4.024	24.900	9.353	21.082	21.064	86.692	87.589	87.108

Figure 2: Results obtained by different experiments. Here, Z = Zero Shot, F = Fine Tuning, N = NER, C = Co Occurrence. For numbers and prompt strategies, refer to the experiment list.

Here The metrics with R prefix indicates the Rogue scores. The metrics with B prefix indicates the bert score. We consider the zero shot results of experiment 1 and 2 as our baselines. We can see from the results that adjusting the prompts does improve the score. We observe the best performance in all the metric for 4b, that is Fine tuning with NER tags with custom prompt. This goes well with our expectation because we tell the model to focus on specific important tags in the dataset that adheres to biological and medication based phenomenon. This directly aligns with medical texts. Thus the model was able to pickup that certain tags were more important than others and used those effectively while summarizing the questions. We observe that for the zero shot case, as expected, the results are not that better, but show some improvements in the metrics in experiment 2 with the addition of NER tags. As for zero shot performance on co-occurrence tags dataset in experiment 5, the performance improves over the baseline zero shot but not better than experiment 2. For normal finetuning with regular prompts in experiment 3a and 3b, they occupy most of the second place in terms of metrics as seen by the underlined results. To our surprise, experiment 4a which was conducted with NER tagged datasets but with with simple prompt, performed worse than regular finetuning as seen in experiment 3a and 3b. This is more prominent in the Recall and F1 score of BertScore. Interestingly co-occurrence did not perform as well as expected. We observe even with finetuning in experiment 6, only in Rouge score-2 the model comes in close second to the top model. In terms of other metrics, the results are significantly lower from the best performing model and also lower than normal finetuned model of experiments 3a and 3b. Even with the prompt telling the model to utilize the co-occurrence pairs of words explicitly did not help the model to better summarize the questions. Also compared to zero shot performance on the dataset tagged with co-occurrence tags, the performance did not improve much with finetuning as seen between experiment 5 and 6. One of the core reasons could be due to sparsity issue of co-occurrences and thus not getting enough strong signal from these to rely upon such tags even with the explicit prompts. The validation loss plots in figure 3 and relevant discussion for the models in experiment 3b, 4b and 6 have been added in the appendix section for more clarity. As discussed in that section, despite having similar trends in evaluation loss, the test results differ significantly between these experiments, specially between models of experiments 6 and 4b.

4 Analysis

We learned a few interesting things from the project. We found that supplying meaningful and helpful prompts can allow a language model to understand the context significantly better. Our findings underscored the pivotal role of well-crafted prompts in facilitating a deeper understanding of the underlying context. Specifically, we observed that supplying prompts tailored to the specific domain or topic at hand markedly improved the language model's ability to grasp nuanced nuances and extract relevant information effectively. Moreover, the provision of contextually rich prompts not only aided in deciphering ambiguous or complex textual inputs but also contributed to generating more coherent and contextually accurate responses. This revelation underscores the significance of prompt customization in refining the performance of language models, offering promising avenues for further exploration and optimization in natural language processing tasks.

In terms of methodology and data utilization, our approach centered on leveraging the Flan-T5-Base model, a pre-trained language model renowned for its prowess in natural language processing tasks. For our dataset, we focused on medical question summarization, which provided a rich source of domain-specific text for analysis. Our methodology entailed employing Named Entity Recognition (NER) tags and co-occurrence matrices within our prompts. NER tags facilitated the identification

and extraction of key entities within the medical domain, enabling the creation of targeted prompts tailored to the specific context of each query. Simultaneously, co-occurrence matrices allowed us to capture the relationships between different entities, further enriching the contextual cues provided to the language model. By integrating these methodological elements, we aimed to enhance the model's understanding of medical queries and improve the quality of its summarization outputs.

What did not work as expected was that while we had anticipated that leveraging co-occurrence matrices would enhance the model's contextual understanding, the complexities of medical language and dataset characteristics likely contributed to the unexpected outcomes observed in our experiments. Several factors could contribute to this unexpected outcome. Firstly, the complexity and variability inherent in medical language might have posed challenges for accurately capturing relevant co-occurrences. Medical terminology often encompasses a wide array of specialized terms and concepts, which may exhibit intricate relationships not easily discernible through simple co-occurrence analysis. Additionally, the size and specificity of the dataset might have influenced the effectiveness of the co-occurrence approach. If the dataset lacked sufficient diversity or granularity, the co-occurrence matrix may not have adequately captured the nuanced associations between entities. Furthermore, the inherent limitations of co-occurrence analysis, such as its reliance on surface-level statistical patterns and inability to capture semantic nuances, could have hindered its effectiveness in this context.

Another interesting observation we learned from experiment 4a where we conducted finetuning on NER-tagged dataset but supplied simple prompt to rely on tags represented in <> brackets. This strategy did not come as fruitful. Simply telling the model to focus on these tags did not provide enough useful information to the model to take advantage from. Therefore, the model objectively performed worse than a regular finetuned model of experiment 3b where the dataset did not contain any additional tags. We learned that supplying the model with prompts focusing on important medical NER tags such as **<MEDICATION>, <BIOLOGICAL_STRUCTURE>, <BIOLOGICAL_ATTRIBUTE>** etc proved itself to be the most useful among all of our experiments. We can understand the fact that the model performed better in this case because the prompt provided useful information for the model to focus on. It knew explicitly what to look for during summarization. Albeit, the underlying transformer model in flan-t5 does use self attention to make sense of the structural dependencies, providing with additional cues such as which tags to focus more on can be of great help to the model as evident from the results.

We did a qualitative analysis from the inference results. They are as follows. Please refer to experiment list and results figure for Experiment number and codes (Z, F, C).

- Example 1: How can i get rid of a lower lip birthmark permanently? (Remarks: Best by fine-tune + NER, worst by the simple prompt 1a)
  - 1a(Z) - I have a lower lip birthmark. I have tried all the options out there and guess what still have it, continues to grow back.....any suggestions?
  - 1b(Z) - Is there a cure for a lip sagging?
  - 2(Z + N) - What is the best way to treat a lip birthmark?
  - 3a(F) - What are the treatments for lower lip birthmark?
  - 3b(F) - Is there a cure for lower lip birthmark?
  - 4a(F+N) - What are the treatments for lower lip birthmark?
  - 4b(F+N) - What are the treatments for lower lip birthmark?
  - 5(Z+C) - Is there a cure for a lip birthmark?
  - 6(F+C) - What are the treatments for a lower lip birthmark?
- Example 2: Can ertapenem cause seizures? (Remarks: Bad by all zero-shot, fine-tune performs better, but the simplistic prompt gives very lengthy and unnecessary summary as seen in 4a(F+N).)
  - 1a(Z) - Is ertanpenem safe?
  - 1b(Z) - Is ertanpenem safe for my father?
  - 2(Z + N) - Is ertanpenem safe for my father?
  - 3a(F) - Have your doctor diagnose Klebsiella pneumoniae.

- 3b(F) - Is ertanpenem a good or bad for my father's infection?
- 4a(F+N) - My father had hip surgery. About one month he was detected an infection with Klebsiella pneumoniae. They started to give him intravenous ertanpenem. After almost a month, he is having episodes of seizures. He never had before. The doctors say that ertanpenem has nothing to do with it and they continue to provide it and at the same time they are giving him anti-seizure medication. I would like to know if it is correct or is at risk for his life.
- 4b(F+N) - What are the treatments for Klebsiella pneumoniae?
- 5(Z+C) - Is ertanpenem safe for my father?
- 6(F+C) - If ertanpenem has nothing to do with Klebsiella pneumoniae, and the doctors continue to provide it, and at the same time they are giving him anti-seizure medication, I would like to know if it is correct or is at risk for his life.

We can see from the given examples that fine-tuning performs better than zero-shot and adjusting the prompts to give more meaningful information helps to achieve better performance with NER tags. Compared to them, co-occurrence does not perform that well.

As for future strategies, If we had one more month to work on the project, we would experiment with additional prompting methods and explore other complex techniques such as chain of thought and graph of thought prompting to add reasoning to the summarization task. We would explore how to augment the dataset such as simplifying the questions first or paraphrasing or changing the style etc. We would also conduct error analysis to reiterate our training loop. We would also consider other useful features rather than simple co-occurrence. And if we had six more months, we would experiment with more complicated architectures such as other transformers and try out different hybrid models. Also we would experiment with other task formulations such as extracting relations (open and close ended) from the texts and incorporate graph neural network with sequence models and also prompting strategies together to see how the model performs. We would also try to reframe the problem in other views such as a question answer type system where we frame the summarization in a question answering task where the answer will be the question summary. If we had 5 more years, we would focus on deeper studies alongside regular experiments such as what is the model actually learning while dealing with medical texts and also consider ethics and bias issues with medical nlp domain and design models based on those. We would collaborate with medical professional and push the current state of the art not on summaries but also other medical texts and bring these solution to practice.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

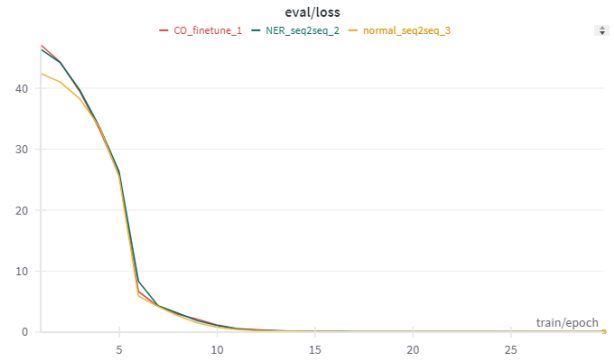

Figure 3: Evaluation Loss during finetuning of models in experiment 3b,4b and 6

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

The validation set loss for experiments 3b,4b and 6 have been added below:

The red curve indicates validation loss for experiment 6, the green one indicates for experiment 4b and yellow one for normal finetuning in experiment 3b. We can see some interesting results from the attached plot. All the evaluation losses seem to have similar trajectory and seem to converge in the same way. Though model in experiment 3b starts at a lower loss than others, by the time it reaches 30 epochs, the losses have stopped decreasing and turns out to be flat.