Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240497

Automation of Trainable Datasets Generation for Medical-Specific Language Model: Using MIMIC-IV Discharge Notes

Youngrong LEE[a], Chansik KIM[a,b] and Taehoon KO[a,b,1]
[a] *Department of Medical Informatics, College of Medicine, the Catholic University of Korea,* [b] *Department of Biomedicine & Health Sciences, College of Medicine, The Catholic University of Korea, Republic of Korea*
ORCiD ID: Youngrong Lee https://orcid.org/0000-0003-1367-4381, Chansik Kim https://orcid.org/0009-0003-9727-2183, Taehoon Ko https://orcid.org/0000-0002-4045-0036

Abstract. This study introduces a novel approach for generating machine-generated instruction datasets for fine-tuning medical-specialized language models using MIMIC-IV discharge records. The study created a large-scale text dataset comprising instructions, cropped discharge notes as inputs, and outputs in JSONL format. The dataset was generated through three main stages, generating instruction and output using seed tasks provided by medical experts, followed by invalid data filtering. The generated dataset consisted of 51,385 sets, with mean ROUGE between seed tasks of 0.185. Evaluation of the generated dataset were promising, with high validity rates determined by both GPT-3.5 and a human annotator (88.0% and 88.5% respectively). The study highlights the potential of automating dataset creation for NLP tasks in the medical domain.

Keywords. Large Language Models, Machine-generated Datasets, Fine-tuning

1. Introduction

Recent advancements in natural language processing (NLP) have primarily revolved around the development of pre-trained large language models (LLM) and fine-tuning using instruction datasets [1]. However, the manual creation of datasets for fine-tuning language models poses significant time and financial challenges. To address this, recent researches have explored automating the generation of machine-generated datasets using human-provided seed tasks, particularly exemplified by models like GPT [2]. Also, human-generated datasets often tend to focus on popular NLP tasks, lacking coverage across diverse tasks. Additionally, the collection of medical domain data presents further challenges due to constraints related to sensitive personal information and data curation. Therefore, this study aims to introduce the process of generating machine-generated instruction datasets for fine-tuning medical-specialized language models using MIMIC-IV discharge records.

[1] Corresponding Author: Taehoon Ko; E-mail: thko@catholic.ac.kr.

2. Methods

This study utilized a modified SELF-INSTRUCT framework [2] to create a text dataset based on MIMIC-IV discharge notes. The dataset included instructions, discharge notes as inputs, and outputs in JSONL format. A random sample (n=9,781) of the discharge notes (n=331,794) was selected and cropped randomly to less than 250 words to fit the token limit of the GPT3.5. The data generation involved three main stages: 1) generating task instructions based on seed tasks created by medical experts, 2) generating outputs using the instructions and input, and 3) filtering invalid data. This process was repeated until reaching the target instruction size of 55,000. The statistical distribution of the dataset and mean Rouge-L score [3] between seed tasks were evaluated. We also randomly sampled 200 sets and had an expert annotator and GPT3.5 determine their validity based on whether the instructions valid and output acceptable. Sets meeting both criteria were considered valid.

3. Results

The study utilized 101 seed tasks, covering 12 representative NLP tasks[2]. A total of 51,385 instruction sets were generated, with average lengths (in number of words) of 17.66, 251.46, and 35.36 for instructions, inputs, and outputs. The mean Rouge between seed tasks was 0.185. GPT-3.5 found 94% of instructions and 88% of outputs valid in 200 sets (overall, 88.0%), while the human annotator found slightly higher rates: 98.5% of instructions and 88.5% of outputs. Overall, the human annotator deemed 88.5% of the sets valid.

4. Discussion and Conclusions

This study proposes a novel approach for generating machine-generated large-scale text data using discharge records, increasing the potential for developing medical-specialized model through fine-tuning. The strengths of the study lie in the comprehensive generation of high-quality training data covering diverse medical NLP tasks. Both GPT and human annotators demonstrated excellent quality evaluation of the data. Future research should expand data size and evaluate fine-tuned model to evaluate the utility of the generated dataset.

References

[1] BACH SH, et al. Promptsource: An integrated development environment and repository for natural language prompts. arXiv preprint. 2022. DOI: 10.48550/arXiv.2202.01279.
[2] WANG Y, et al. Self-instruct: Aligning language models with self-generated instructions. arXiv preprint. 2022. DOI: 10.48550/arXiv.2212.10560.
[3] Lin CY. Rouge: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out, 2004.

[2] i.e., Co-reference Resolution, Question Answering, Natural Language Generation, Text Summarization, Text Classification, Temporal Information Extraction, Relation Extraction, Named Entity Recognition, Paraphrasing, Clinical Concept Normalization, Keyword Extraction, and Abbreviation Expansion