HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing

Sonish Sivarajkumar[1], Yanshan Wang, PhD, FAMIA[1,2,3*]
[1]Intelligent Systems Program, School of Computing and Information, University of Pittsburgh, PA; [2]Department of Health Information Management, University of Pittsburgh, PA; [3]Department of Biomedical Informatics, University of Pittsburgh, PA

Abstract

Developing clinical natural language systems based on machine learning and deep learning is dependent on the availability of large-scale annotated clinical text datasets, most of which are time-consuming to create and not publicly available. The lack of such annotated datasets is the biggest bottleneck for the development of clinical NLP systems. Zero-Shot Learning (ZSL) refers to the use of deep learning models to classify instances from new classes of which no training data have been seen before. Prompt-based learning is an emerging ZSL technique in NLP where we define task-based templates for different tasks. In this study, we developed a novel prompt-based clinical NLP framework called HealthPrompt and applied the paradigm of prompt-based learning on clinical texts. In this technique, rather than fine-tuning a Pre-trained Language Model (PLM), the task definitions are tuned by defining a prompt template. We performed an in-depth analysis of HealthPrompt on six different PLMs in a no-training-data setting. Our experiments show that HealthPrompt could effectively capture the context of clinical texts and perform well for clinical NLP tasks without any training data.

Introduction

Machine learning techniques have achieved the state-of-the-art performance in many clinical NLP tasks. However, the necessity for vast amounts of annotated text is a fundamental bottleneck of applying these techniques in clinical NLP systems. Despite some good quality and publicly available clinical corpora, such as i2b2[1] datasets, MIMIC-III[2] datasets, and BioNLP[3] datasets, more annotated clinical text datasets from real-world Electronic Health Records (EHRs) are needed to train machine learning models. Most of the existing machine learning-based clinical NLP systems are based on the supervised approach, requiring abundant data for reasonable accuracy. Supervised learning has been used in the clinical NLP for different applications, including disease identification[4], cohort selection[5], drug regimen selection[6], etc. These systems have proved to be accurate and effective in scenarios with abundant training data availability. However, the reality is that most clinical NLP projects have only a small number of annotated clinical documents due to the expensive and time-consuming manual annotation process. Therefore, building clinical NLP systems with few or no annotated documents remains a top focus in informatics research.

A Pre-trained Language Model (PLM)[7] is a neural network trained on large amount of unannotated data (such as Wikipedia data or PubMed data) in an unsupervised way. This process is called pre-training, which allows the model to learn the general structure of a language (or domain), including the usage of vocabulary in that domain-specific context. The state-of-the-art biomedical and clinical PLMs, such as *BioBERT*[8] and *Clinical BERT* [9], have been trained using millions of clinical texts, including EHRs and radiology reports. The PLM model is then transferred for a downstream NLP task, where a smaller task-specific labelled dataset is used to tune the PLM, to construct the final model capable of executing the downstream task. This process is called fine-tuning a PLM. The entire paradigm is called transfer learning, as the model learns a general context during pre-training, and this knowledge is transferred for a specific NLP task by fine-tuning the model in a low-data regime. For example, *Clinical BERT* has been pre-trained largely on an unlabeled clinical text datasets, i.e., MIMIC, during which it has learned clinical linguistic characteristics from clinical narratives such as physician's notes. Later, this general clinical knowledge could be transferred for specific tasks, such as adverse event detection[10] or clinical Named Entity Recognition (NER)[11] tasks, by fine-tuning the PLM with less amount of task-specific annotated data.

The approach described above with few annotated documents is usually called few-shot learning (FSL). In the real-world scenario when there may not be any annotated data, one needs zero-shot learning (ZSL)[12] which doesn't require any laborious data annotation by experts. In ZSL, models pretrained from unlabeled data can be used for making predictions on unseen data. This technique gained popularity after PLMs have been successfully applied on cross-

* Corresponding author: Yanshan Wang, PhD, FAMIA; yanshan.wang@pitt.edu

domain adaptation tasks[13]. Studies have shown that PLMs could be effectively used for sentiment analysis tasks in a zero-shot cross-lingual setting[14]. In clinical NLP, there are numerous medical concepts that need to be extracted from clinical notes and meanwhile there are a limited number of available annotated datasets. However, ZSL has been rarely investigated in clinical NLP. Therefore, there is an urgent need to investigate the use of ZSL systems in clinical NLP, which doesn't require annotated data for downstream applications.

In this study, we investigate the use of the ZSL technique in addressing the issue of the lack of annotated datasets for clinical NLP tasks. We show that clinical text classification can be executed in a no-training-data setting by tuning a PLM to predict the probability of classifying clinical text. We use prompt-based learning[15], one of the latest ZSL methods, which involves creating templates for a given NLP task. The process of building prompting functions or templates for the input text data is called prompt engineering. We can define a mapping function that maps this set of tokens to the expected prediction for the downstream task. In short, prompt-based learning can be defined as a two-step process, where we apply a template with masked slots, denoted as [MASK], and use the knowledge learned by a PLM to predict the token that can be filled in the slot [MASK].

An example of the prompt-based learning process is illustrated in Table 1. We use the PLM to fill in the masked tokens in sequences *"The patient has cough and expanded chest that does not deflate when he exhales. This is the symptom of [MASK] disorder"*. Then the PLM constructs an explicit word-context or word co-occurrence matrix from the pre-trained embeddings. Using prompt-based learning, we try to tune the PLMs to make predictions by defining these sequences or prompts. Till the date of writing this paper, prompt-based learning has not been used for NLP systems in the clinical domain to the best of our knowledge.

Table 1. An example of the prompt-based learning process. We define prompting functions or templates, which transform the original input text into forms that can be filled by a PLM that predicts the probabilities of texts that can be served in the slot.

Description	Input
Input text which needs to be classified	"The patient has cough and expanded chest that does not deflate when he exhales."
We can define a Prompting Function	[text]. This is a symptom of [MASK] disorder.
Now the transformed prompt becomes	"The patient has a cough and expanded chest that does not deflate when he exhales. This is a symptom of [MASK] disorder"
The language models predict a token, phrase or sentence that fills the slot [MASK]	"lung", "asthma", or "respiratory".

Background

With the development of PLMs, in other words Neural Language Models, various architectures have been proposed to solve NLP tasks. The first-generation PLMs, such as Skip-Gram[16] and GloVe[17], were based on word embeddings. A major drawback of these models was that they are not capable of understanding complex linguistic concepts and the underlying contexts behind these embeddings. Nevertheless, they were effective in comprehending the meaning of a word/sentence.

The second-generation PLMs are based on contextual embeddings. These models can differentiate the meaning of words in different contexts. *BERT*[18], *GPT-2*[19], *RoBERTa*[20], *XLNet*[21], *T5*[22] are examples of the second-generation PLMs that can be fine-tuned for various NLP tasks. These models have achieved state-of-the-art performance for various NLP tasks. Table 2 summarizes six PLMs used in this study.

Prompt-based learning is one of the latest developments in the field of NLP and has not been effectively explored for many Natural Language Inference (NLI) applications. There have been multiple attempts to use PLMs are unsupervised learners. Radford et al.[23] demonstrated that PLMs could be used for many NLP tasks by "task descriptions", without the need for fine-tuning. They have used this method for reading comprehension, translation,

summarization, question answering tasks, etc. and found that GPT-2 could produce promising results in most of these tasks without any explicit supervision.

Researchers have also tried to use semi-supervised training procedures with cloze-style templates for text classification and NLI tasks[24]. This few-shot approach outperformed supervised learning in a low-resource setting by a large margin. GPT-3[25], with its 125-175 Billion parameters, achieved good results in the zero-shot and one-shot settings and outperformed state-of-the-art models in the few-shot scenario[25]. GPT-3 shows 86.4% accuracy in the few-shot setting of a word prediction task, an increase of 18% over previous models.

Table 2. A summary of PLMs used in this study.

PLM	Description
BERT[18]	Bidirectional Encoder Representations from Transformers (*BERT*) is a pre-trained language model developed by Google, which was trained using 800 million words from the BookCorpus dataset and 2.5 billion words from Wikipedia. It is based on Transformer architecture[26], a neural network architecture for language interpretation built on a self-attention mechanism.
RoBERTa[20]	Robustly Optimized BERT Pre-training Approach (*RoBERTa*) is based on *BERT* architecture. It is trained with a bigger batch size and longer text sequences without Next Sentence Prediction (NSP). It contains more parameters than BERT (123 million for RoBERTa base and 354 million for RoBERTa large). It is trained using dynamic masking patterns, in which the masked token is altered with each iteration of the input.
BioBERT[8]	Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (*BioBERT*) has the same architecture as BERT but is mainly pre-trained on Biomedical texts. It is trained on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC), apart from the Wikipedia and BookCourpus data.
Clinical BERT[9]	This BERT model is primarily trained on clinical texts, EHRs and discharge summaries and thus produces better performance than other models on EHR text corpus.
GPT-2[19]	This is essentially a decoder-only transformer. The model is built by stacking up the transformer decoder blocks. Unlike the self-attention used by transformers, GPT-2 uses masked self-attention. It is an autoregressive language model.
T5[22]	Text-to-Text Transfer Transformer (*T5*) is one of the latest PLMs released by Google, which outputs a text string instead of a label or a span of the input to the input sentence. Thus, it is a seq2seq model. Instead of employing task-specific architecture for different sorts of language tasks, it was built to have one standard architecture to learn many tasks

Based on unlabeled corpora and a pre-defined label schema, researchers have shown that the prompt-based ZSL approach can automatically learn and summarize entity kinds[27]. This approach could outperform even 32-shot fine-tuning. It has been shown that text classification tasks can be formulated as a prompt -based learning problem, given by:

$$p(y \in \mathcal{Y} \mid x) = p\big([\text{MASK}] = w \in \mathcal{V}_y \mid T(x)\big)$$

where $y \in \mathcal{Y}$ is a label, and $\mathcal{V}_y = [l_1, l_2, .. l_n]$ is the label word set with labels $l_1, l_2, .. l_n$. A prompt template T modifies the input text x and creates a prompt function $T(x)$, with additional tokens than x. [MASK] can be considered the slot, which needs to be predicted by the PLM from the label set $\mathcal{V}$.

Methods

HealthPrompt Paradigm

In this paper, we propose a novel clinical NLP framework using the prompt-based ZSL approach named HealthPrompt. The architecture of the HealthPrompt framework is shown in Figure 1. HealthPrompt comprises the following components:

1. EHR Chunk Encoder: Split each EHR document into segments of texts or chunks. This forms a chunk-level representation of an EHR. Details will be discussed in the following sections.
2. Label Definition: A label set is created, containing the labels to be predicted by the zero-shot clinical text classification model.
3. PLM Selection: Choose a PLM to predict the prompt class. Different PLMs have different properties, and hence, they perform differently in each task. For example, *BioBERT* may work better on biomedical texts, while *Clinical BERT* may have better performance on clinical texts.
4. Template Definition: Define a prompt template. Prompt templates are the text templates that modify the input text before being fed into a PLM. These templates need to be carefully engineered for the downstream task. This component is the most crucial part of the framework.
5. Inference: Infer the most appropriate label that fits the prompt template. This is done based on an answer search technique, which will be discussed in the following sections.

The key feature of the HealthPrompt framework is that it does not require any training data. We can directly feed an unlabeled clinical note document to the framework after defining the prompt template and label set. Thus, HealthPrompt is a ZSL framework for clinical texts. The input clinical note document is first split into chunks of sentences. These chunks or segments of sentences are then passed into the prompt-based model, which infers the label of the clinical texts.

HealthPrompt can be used for any clinical NLP task that can be formulated as a classification problem, including clinical Named Entity Recognition (NER), clinical text classification, and Adverse Drug Event (ADE) detection. In our implementation, the Openprompt[24] python library was used to build a prompt-based zero-shot model. This unified python library gives the flexibility to define the prompt templates and verbalizers (i.e., label sets).

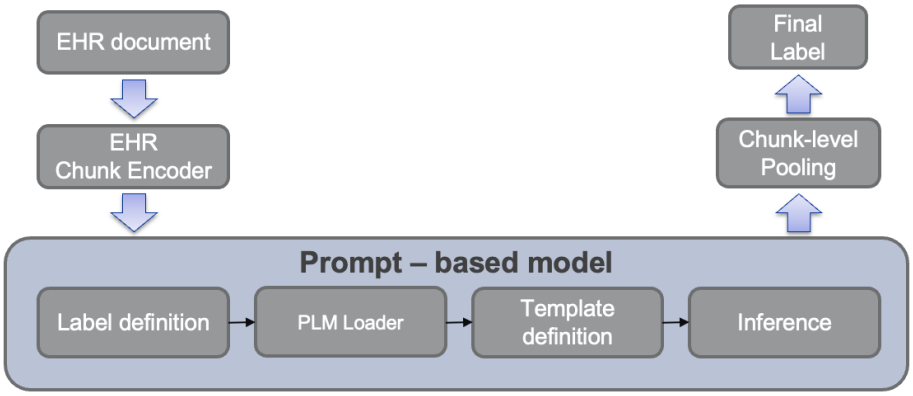

Figure 1. The architecture of the HealthPrompt framework.

Chunk Encoder and Pooling

Most PLMs are limited by the maximum length of input text due to hardware limitations. BERT, for instance, supports only 512 tokens at a time. But clinical note documents usually contain thousands of tokens. For instance, each document in MIMIC-III contains 8,131 tokens on average.

Chunk-level pooling[28] is a solution to this problem. In this technique, Chunk Encoder splits each clinical note document into chunks(segments) of equal lengths. The hierarchical encoding and pooling framework was originally designed to overcome the limitation of the PLMs to be trained on long documents. We adopted this chunk representation in the HealthPrompt framework, such that each chunk is passed to the prompt-based model, and the predicted label of all the chunks of a document are collected as a label set or collection set of the document, as illustrated in Figure 2. Chunk Pooler aggregates the predicted labels from the label collection. We use maximum

pooling to combine the chunk labels into document labels, and thus perform a document-level classification of long clinical text documents. Short documents or clinical texts can also be classified using HealthPrompt, where they will be split into a smaller number of chunks. HealthPrompt is flexible for sentence-level clinical text classification, where the chunk encoder considers the sentence as one chunk, and hence no splitting is performed on the clinical text.

The labels corresponding to each chunk of the document are then normalized with respect to their number of occurrences in the label collection set of the document, and the label with maximum probability is chosen as the final label of the entire document. Let L be the label set with $L_1, L_2, .., L_n$ labels corresponding to chunk $C_1, C_2, ..., C_n$. The final label L_f is given by:

$$L_f = Max[P(L_n / L)]$$

The type of encoder and the pooling mechanism used by HealthPrompt determines how a document is represented. Though we have used maximum pooling, HealthPrompt has the flexibility to use other pooling mechanisms like average pooling, transformer encoder, Long Short-Term Memory (LSTM), and Convoluational Neural Network (CNN) based pooling.

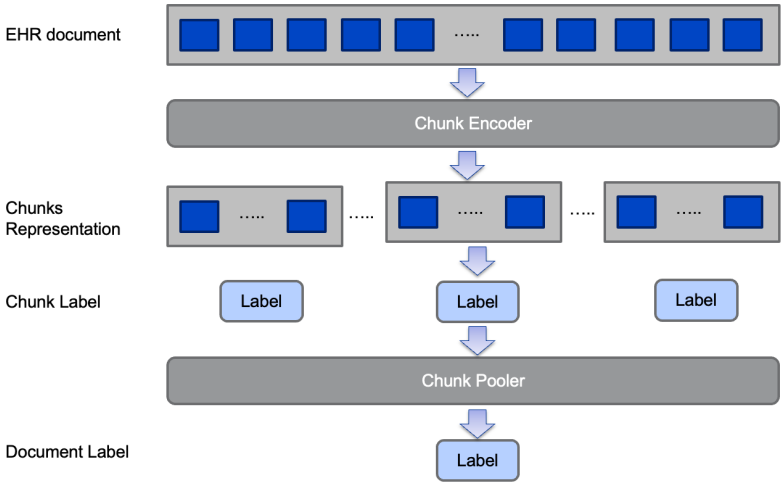

Figure 2. The architecture of Chunk-level pooling in HealthPrompt.

Template Definition: A Prompt Engineering Step

Prompt Engineering is related to the template definition of the HealthPrompt framework. Different templates can be used to express the same concept. Hence it is essential to carefully design the templates for exploiting the zero-shot capability of a PLM. Thus, the prompt template designing requires domain expertise with the task on which zero-shot learning is performed. In this study, we explore two types of prompt templates, namely cloze prompt and prefix prompt.

Cloze Prompt. Cloze prompts are templates with slots in a continuous textual string. This can be viewed as a fill-in-the-blank query on the PLM. Cloze prompts have been effective in an open-domain question answering system[29]. Researchers have also shown that cloze templates using BART achieve competitive results on the NER benchmark and outperform traditional sequence labelling methods[30]. In general, Cloze prompts are used for text classification and Natural Language Understanding (NLU) tasks. "[text]. This is [MASK] disease." is an example of cloze prompts.

Prefix Prompt. In prefix prompts[31], the tokens predicted by a PLM will fill in the subsequent masked tokens. These are commonly used for text generation and sequence prediction tasks. Li et al.[32] experimented with prefix tuning and showed that

Prompt Template	Type
'[text]. Disease : [MASK]'	Prefix Prompt
'[text] : This effects [MASK]'	Prefix Prompt
'[text] : [MASK] disorder'	Cloze Prompt
'[text] : [MASK] type of disease'	Cloze Prompt

Table 3. Four different prompt templates used in the experiment. [text] represents the input clinical text and [MASK] the token to be filled by PLMs.

these prompts could improve generalization on Natural Language Generation (NLG) tasks. The prompt template "[text]. Disease: [MASK]" is an example of prefix prompts.

In this study, we created four different prompts for a clinical text classification task, whereby each patient note is classified into different disease labels using prompt-based learning without training the model. Among these prompts, two were cloze prompts, and two were prefix prompts, which are given in Table 3. Till the date of writing this paper, there is no guidelines on how to design prompt templates, particularly for clinical NLP tasks. The four prompts we considered for this study were designed with the clinical expertise on the phenotype extraction task. But there could be other ways by which these prompts could be defined. We considered four templates in this paper as the purpose of this study is to show that ZSL can be applied for clinical NLP tasks using prompt-based learning.

Pretrained Language Models (PLM)

The proposed HealthPrompt framework could embed any PLMs. In this study, we experimented with six PLMs, namely *BERT, BioBERT, Clinical BERT, RoBERTa, GPT-2* and *T5*. We evaluated these language models for the clinical text classification tasks with the four prompt templates. We would like to validate whether careful prompt engineering is effective for clinical NLP tasks and which PLM has the best performance in the HealthPrompt framework.

Inference using Answer search

A label set is a list of labels on which the input text needs to be classified. Once the label set is defined, PLMs will generate multiple tokens that can fill the masked tokens in a prompt template. For instance, "Patient has severe headache and nausea. This is a symptom of [MASK]". In this case, a PLM can fill the [MASK] with tokens like malaria, brain injury, brain aneurism, etc.

Answer search is the process by which a PLM selects the token with the highest probability for a task definition. It is performed over the set of potential answers based on the probability of each value in completing the prompt. A verbalizer[33] function is usually defined for the PLM to predict the probability distribution over the label set. This set of logits is then mapped to the corresponding labels. The label with the highest probability is selected as the predicted token.

Dataset

We used the phenotype annotation dataset for patient notes in the MIMIC-III database[34] as the testing dataset. This dataset identifies whether a patient has a given phenotype based on their patient note. The patient notes were retrieved from MIMIC-III, a dataset collected from Intensive Care Units of a large tertiary care hospital in Boston. Those notes were manually annotated for the presence of several high-context phenotypes relevant to treatment and risk of re-hospitalization. We randomly sampled a subset of 347 patient notes and the corresponding phenotype annotations to evaluate the performance of our zero-shot HealthPrompt framework. These long patient notes had around 7651 tokens on average. This subset consisted of 10 classes of different diseases. The distribution of the dataset and its labels is shown in Table 4.

These documents are split by the chunk-level encoding mechanism of the HealthPrompt to get chunk level labels, which is aggregated to by the pooling mechanism to fetch the final document label.

Table 4. Distribution of different phenotype categories in the MIMIC-III phenotype subset used in this study.

Phenotype Category	Number of documents
Chronic Neurological Dystrophies	54
Advanced Heart Disease	49
Depression	41
Advanced Cancer	38
Advanced Lung Disease	38
Chronic Pain Fibromyalgia	35
Obesity	33
Non-Adherence	24
Alcohol Abuse	24
Dementia	11

Evaluation

We used the entire dataset with 347 patient notes as the testing set to evaluate the HealthPrompt framework. All four prompts were evaluated on the testing set, with each of the six PLMs we considered for this study. Accuracy, precision, recall, and F1 score were calculated for each experiment. The following are the definitions of these metrics:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Nagative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Nagative}$$

$$F1\ score = \frac{2 \cdot True\ Positive}{2 \cdot True\ Positive + False\ Positive + False\ Nagative}$$

Results

We report all models' performance in Table 5 for four prompt templates. For the MIMIC-III phenotype dataset, we report the average accuracy, precision, recall, and Macro F1 score across the ten classification labels. Cloze prompt "[text]. [MASK] type of disease" was the best performing prompt for the phenotype classification task with the best F1 score of 0.86 and an accuracy of 0.85. The prefix prompt also has promising performance, with the best F1 score of 0.81 and accuracy of 0.81. The other two prompts received comparatively low performance with all the models, with accuracy scores in the range of 0.50-0.75. Hence, it is clear that cloze prompts produce the best results comparable to the state-of-the-art clinical text classification models.

Table 5. Results on the MIMIC-III phenotype subdataset with six PLMs.

Prompt Template	Model	Accuracy	Precision	Recall	F1-score
'[text]. Disease : [MASK]'	BERT	0.61	0.68	0.69	0.68
	BioBERT	0.78	0.78	0.71	0.74
	Clinical BERT	**0.81**	**0.79**	**0.83**	**0.81**
	RoBERTa	0.69	0.71	0.71	0.71
	GPT-2	0.68	0.67	0.57	0.62
	T5	0.63	0.64	0.66	0.65
'[text] : This effects [MASK]'	BERT	0.55	0.45	0.41	0.43
	BioBERT	0.71	0.70	0.71	0.70
	Clinical BERT	**0.73**	**0.72**	**0.72**	**0.72**
	RoBERTa	0.61	0.60	0.57	0.58
	GPT-2	0.51	0.50	0.61	0.55
	T5	0.65	0.69	0.61	0.65
'[text] : [MASK] disorder'	BERT	0.61	0.68	0.69	0.68
	BioBERT	0.71	0.69	0.70	0.69
	Clinical BERT	**0.75**	**0.73**	**0.72**	**0.72**
	RoBERTa	0.58	0.51	0.48	0.49
	GPT-2	0.69	0.71	0.71	0.71
	T5	0.68	0.69	0.70	0.69
'[text] : [MASK] type of disease'	BERT	0.70	0.74	0.73	0.73
	BioBERT	0.77	0.73	0.73	0.73
	Clinical BERT	**0.85**	**0.86**	**0.86**	**0.86**
	RoBERTa	0.71	0.71	0.69	0.70
	GPT-2	0.73	0.73	0.75	0.74
	T5	0.71	0.71	0.72	0.71

Clinical BERT showed the best performance in all evaluation metrics in all the prompts, including F1-score and accuracy. The reason might be that Clinical BERT is pre-trained on the MIMIC dataset that contains clinical language chracteristics. *BioBERT* produced comparable results to *Clinical BERT*. Despite having a complex model and significantly a greater number of parameters, GPT-2 and T5 showed poor performance compared to others. The

potential reason is that *GPT-2* and *T5* models are autoregressive and seq2seq models, respectively. Our prompt approach mainly relied on masking a token in the prompt template. Hence, the Masked Language Models (MLM) performed better than *GPT-2* and *T5*. Within Masked Language Models, *Clinical BERT* and *BioBERT* performed better since these models were explicitly trained on biomedical and clinical texts.

These results show that prompt-based learning with the *Clinical BERT* model can be effectively applied to clinical text classification tasks in a no-training-data setting. Using HealthPrompt, we directly utilized the PLMs and tuned the task definitions with prompts. This alignment of clinical task definition to PLM produced promising performance, on par with the text classification models in the general domain.

Discussion

Recently prompt-based ZSL has become a hot research area in NLP. Most existing prompt-based NLP research focused on simple sentence-level classification tasks. However, clinical texts from EHRs are mostly long text documents that differ from these sentence-level data in the general domain. This is one of the biggest challenges when applying prompt-based ZSL and PLM to clinical NLP. In the HealthPrompt framework, we incorporated chunk-level encoding and pooling to tackle this problem.

HealthPrompt can serve as a general NLP framework for any clinical NLP tasks, when we do not have any training data at the initial stages of a project. Particularly, at the beginning of the outbreak of COVID-19 pandemic, the HealthPrompt could be rapidly applied to classify new disease and phenotypes based on the clinical texts. In addition, most of the clinical text datasets are not publicly available and may require many agreements and attestations to access the data, which is one of the biggest impediments in clinical NLP research. The proposed HealthPrompt framework can be effectively used in most of the clinical NLP tasks by carefully designing different prompt templates, regardless of the data availability and NLP expertise.

Limitations and Future Work

There are three major limitations in this study. First, only six PLMs were tested in the HealthPrompt framework. Among the six PLMs, four were general domain PLMs, namely *BERT, RoBERTa, GPT-2,* and *T5*, and two were clinical domain PLMs, namely *BioBERT* and *Clinical BERT*. This was due to the lack of availability of clinical domain-specific PLMs. Nevertheless, we were able to produce results close to SOTA clinical NLP systems with *Clinical BERT* and *BioBERT*, without using any training data. Specific target-domain PLMs have been developed for domain-specific NLP tasks. For example, *COVID-Twitter-BERT*[35] is a such attempt, where a transformer-based PLM was developed by pretraining on a vast corpus of COVID-19 related tweets. Researchers have also released fine-tuned PLMs that can be used for specific applications, as *BioBERT* finetuned on COVID-19 datasets. These finetuned domain-specific PLMs could produce better results on the HealthPrompt framework for the corresponding downstream task, which will be tested in our future work.

Second, only four prompt templates were designed and utilized in the HealthPrompt framework. The reason is that testing different prompt templates is out of scope of this study as our primary goal is to develop a prompt-based ZSL clinical NLP framework. More evaluations with the new prompt templates and with the development of new domain-specific PLMs are subject to a future work.

Third, the testing dataset used in this paper is relatively small. Only 347 clinical documents from the MIMIC-III phenotype dataset were used to test the HealthPrompt framework. The reason is that most documents are lengthy documents with hundreds of sentences and the chunk encoding and pooling component in HealthPrompt is computationally expensive to process these documents. Neverthless, Healthprompt can extract texts from shorter documents in considerably less amount of time. In future work, we would like to investigate more efficient way of processing lengthy documents and test HealthPrompt on larger datasets.

Conclusion

The lack of publicly available clinical text datasets is the biggest bottleneck for the development and widespread adoption of deep learning techniques in clinical NLP systems. In this study, we developed a novel prompt-based clinical NLP framework called HealthPrompt and applied the paradigm of prompt-based learning on clinical texts. In this technique, rather than fine-tuning or re-training a PLM, the task definitions are tuned by defining a prompt template. We find that, by carefully designing the prompt templates, PLMs can be effectively used for clinical text classification tasks. We performed an in-depth analysis of HealthPrompt on six different pre-trained language models in a no-data setting. Our experiments indicate that prompts could effectively capture the context of clinical texts in a zero-shot setting and perform well on clinical text classification without any training data. To the best of our

knowledge, prompt-based ZSL has not been effectively applied to clinical NLP tasks, and hence the proposed HealthPrompt framework may be viewed as the foundation for a new paradigm for Clinical NLP, especially for NLU and NLI tasks.

Acknowledgment

This project was partially supported by the University of Pittsburgh Momentum Seed Funds, Clinical and Translational Science Institute Exploring Existing Data Resources Pilot Awards, and the School of Health and Rehabilitation Sciences Dean's Research and Development Award, and the National Institutes of Health through Grant Number UL1TR001857 funds. The funders had no role in the design of the study, collection, analysis, and interpretation of data and in preparation of the manuscript. The views presented in this report are not necessarily representative of the funder's views and belong solely to the authors.

References

1. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*. 2011;18(5):552-556.
2. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Scientific data*. 2016;3(1):1-9.
3. Nédellec C, Bossy R, Kim JD, et al. Overview of BioNLP shared task 2013. In: *Proceedings of the BioNLP Shared Task 2013 Workshop*. ; 2013:1-7.
4. Savova GK, Fan J, Ye Z, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. In: *AMIA Annual Symposium Proceedings*. Vol 2010. American Medical Informatics Association; 2010:722.
5. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*. 2019;26(11):1163-1171.
6. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*. 2009;16(3):328-337.
7. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*. 2020;63(10):1872-1897.
8. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.
9. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:190403323*. Published online 2019.
10. Fan B, Fan W, Smith C. Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management*. 2020;57(1):102131.
11. Sun C, Yang Z. Transfer learning in biomedical named entity recognition: an evaluation of BERT in the PharmaCoNER task. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. ; 2019:100-104.
12. Xian Y, Lampert CH, Schiele B, Akata Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*. 2018;41(9):2251-2265.
13. Socher R, Ganjoo M, Manning CD, Ng A. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*. 2013;26.
14. Pelicon A, Pranjić M, Miljković D, Škrlj B, Pollak S. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*. 2020;10(17):5993.
15. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:210713586*. Published online 2021.
16. Guthrie D, Allison B, Liu W, Guthrie L, Wilks Y. A closer look at skip-gram modelling. In: *LREC*. Vol 6. Citeseer; 2006:1222-1225.
17. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ; 2014:1532-1543.
18. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*. Published online 2018.
19. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog*. 2019;1(8):9.

20. Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:190711692*. Published online 2019.
21. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le Q v. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*. 2019;32.
22. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:191010683*. Published online 2019.
23. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog*. 2019;1(8):9.
24. Schick T, Schütze H. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:200107676*. Published online 2020.
25. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in neural information processing systems*. 2020;33:1877-1901.
26. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems*. 2017;30.
27. Ding N, Chen Y, Han X, et al. Prompt-learning for fine-grained entity typing. *arXiv preprint arXiv:210810604*. Published online 2021.
28. Su X, Miller T, Ding X, Afshar M, Dligach D. Classifying Long Clinical Documents with Pre-trained Transformers. *arXiv preprint arXiv:210506752*. Published online 2021.
29. Petroni F, Rocktäschel T, Lewis P, et al. Language models as knowledge bases? *arXiv preprint arXiv:190901066*. Published online 2019.
30. Cui L, Wu Y, Liu J, Yang S, Zhang Y. Template-based named entity recognition using BART. *arXiv preprint arXiv:210601760*. Published online 2021.
31. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:210408691*. Published online 2021.
32. Li XL, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:210100190*. Published online 2021.
33. Hu S, Ding N, Wang H, Liu Z, Li J, Sun M. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:210802035*. Published online 2021.
34. Moseley E, Celi LA, Wu J, Dernoncourt F. Phenotype annotations for patient notes in the MIMIC-III database. *PhysioNet*. Published online 2020.
35. Müller M, Salathé M, Kummervold PE. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:200507503*. Published online 2020.