Navigating Data Scarcity: Pretraining for Medical Utterance Classification

Do June Min, Verónica Pérez-Rosas, Rada Mihalcea
Department of Electrical Engineering and Computer Science
(dojmin, vrncapr, mihalcea)@umich.edu

Abstract

Pretrained language models leverage self-supervised learning to use large amounts of unlabeled text for learning contextual representations of sequences. However, in the domain of medical conversations, the availability of large, public datasets is limited due to issues of privacy and data management. In this paper, we study the effectiveness of dialog-aware pretraining objectives and multiphase training in using unlabeled data to improve LMs training for medical utterance classification. The objectives of pretraining for dialog awareness involve tasks that take into account the structure of conversations, including features such as turn-taking and the roles of speakers. The multiphase training process uses unannotated data in a sequence that prioritizes similarities and connections between different domains. We empirically evaluate these methods on conversational dialog classification tasks in the medical and counseling domains, and find that multiphase training can help achieve higher performance than standard pretraining or finetuning.

1 Introduction

Current language technologies have enabled the analysis of large amounts of medical conversations to gain insights into important aspects of provider-patient interactions such as patient experience, response to treatment, time allocation for health issues, or quality assurance (Zhou et al., 2021). However, many challenges remain open for this growing field of research on natural language processing (NLP) for healthcare. Among them, there is a need for efficient training frameworks that address the lack of large-scale, publicly available medical dialog datasets.

The recent success of large transformer-based models (Vaswani et al., 2017) in many Natural Language Processing (NLP) tasks related to dialog has motivated their application in the healthcare domain, mainly because of their adaptability capabilities. Work in this area has shown that large pretrained language models (PLMs) are effective for tasks such as assessing and analyzing the quality of counseling conversations or building chatbots for mental health care (Flemotomos et al., 2021).

Previous work on dialog-oriented pretraining approaches has used discourse-aware (intersentential) learning tasks to learn "rich and robust context representations and interactive relationships of dialog utterances" (Zhang and Zhao, 2021). In addition, the composition and order of pretraining corpora have also been studied as crucial factors for downstream task performance (Gururangan et al., 2020), with a multiphase pretraining regimen consisting of general, domain-adaptive, and task-adaptive shown to be effective. While these methods have been found useful for general-purpose dialog, it is still unclear how they perform in medical dialog. Different from other types of conversational dialog, medical conversations are domain-specific, and classification models require not only capturing the discourse relations between dialog utterances and turns but also being aware of speaker dynamics and medical terminology.

In this work, we seek to empirically study the effectiveness of dialog-aware pretraining and multiphase pretraining in medical utterance classification tasks. We focus on pretraining tasks that allow the model to leverage conversational properties, such as turn-shift, speaker role, and intersentential dependencies. We evaluate these methods on datasets that are limited in size, especially compared to large corpora that are typically used to pretrain language models. Thus, our goal is to confirm if the pretraining approaches lead to improvements over just finetuning with a small dataset.

The contributions of this work are threefold. (1) We design and implement simple dialog-aware pretraining tasks for medical conversations. (2) we evaluate dialog-aware pretraining and multiphase pretraining and show that while the former does

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 59–68

not fare better than dialog-agnostic approaches, the latter can effectively leverage unannotated corpora of varying task relevance. (3) we draw lessons for practitioners from our experiments.

2 Related Work

Pretrained Language Models. Previous works have explored pretraining objectives and strategies that use large amounts of unannotated data to optimize large neural networks (Kim et al., 2020). These methods can be broadly categorized into autoregressive models (e.g., GPT (Radford and Narasimhan, 2018)) and autoencoding models (BERT) (Devlin et al., 2019). The strategies used in this work belong to the autoencoding class since they rely on reconstructing the original input from its corrupted version. In particular, we explore domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT), which have been shown effective while incorporating both, domain-relevant and task-relevant, unlabeled data (Gururangan et al., 2020). In addition, given recent advances in large language models (LLMs), Lehman et al. (2023) show that smaller language models carefully trained on clinical data outperform much larger models trained on general domain data, motivating our work on leveraging domain-specific pretraining on health-related conversational data.

Pretraining Methods for Conversations. Recent work in this area has used intersentential learning objectives to infer properties associated with the relationship of sentences in the input. For instance, Mehri et al. (2019) used masked-utterance retrieval to guess the replaced (masked) utterance based on the context utterance, and Zhang and Zhao (2021) explored intersentential coherence through utterance order restoration and contrastive loss. Other approaches have explicitly incorporated the dialog structure and information as part of the pretraining task. For example, MPC-BERT (Gu et al., 2021a) focused on multi-party conversations and used dialog-unique information such as utterance speakers and receivers.

Conversation Utterance Classification. Within medical utterance classification, work has been done to either categorize or forecast utterances that describe behaviors from conversation participants (Cao et al., 2019). For categorization, Pérez-Rosas et al. compiled a dataset of motivational interviewing (MI) conversations, where each counselor utterance is annotated with a predefined MI behavior code for the counseling strategy employed in the utterance (Pérez-Rosas et al., 2016). In this work, we adopt the framework joint sentence representation (JSR) (Cohan et al., 2019). That is, instead of encoding a single sentence at a time and using its embedding for classification, we jointly encode multiple sentences in a window with specified context size and use their final embeddings to classify multiple sentences at a time.

3 Pretraining Approaches

We explore dialog-aware pretraining and multi-phase training to leverage unlabeled data in medical conversations.

3.1 Discourse Structure Objectives

We focus on two pretraining tasks that incorporate information about turn shift behavior and the speakers' role. We believe that these play a more significant role in clinical dialog than in everyday conversations since the expected role and behavior of each participant are fixed and understood by the speakers. Thus, we hypothesize that models with improved awareness of these structures will lead to higher performance in medical dialog downstream tasks.

Turn-shift Prediction. An important aspect of conversation dynamics is turn-shifting behavior i.e., points in the conversation where a speaker starts a new turn, which can provide information on power balance and rapport between participants. In clinical conversations, turn-taking behavior helps to move the conversation forward and facilitates patient-provider communication. We incorporate turn-taking information into contextual embeddings by designing a pretraining task in which a model is evaluated and trained on its ability to correctly identify the start of a new turn. We define a turn as a contiguous span of utterances spoken by one speaker. In our model, each utterance is separated by `[SEP]` tokens and we predict whether the current `[SEP]` is the start of a new turn. We use a simple feed-forward neural network and a final sigmoid activation for binary prediction on `[SEP]` tokens with binary cross entropy loss for scoring. Note that since the model receives speaker role information through speaker embedding, it is possible that information leakage may make this task trivial. However, during our experiments, we observe that the model does not converge quickly,

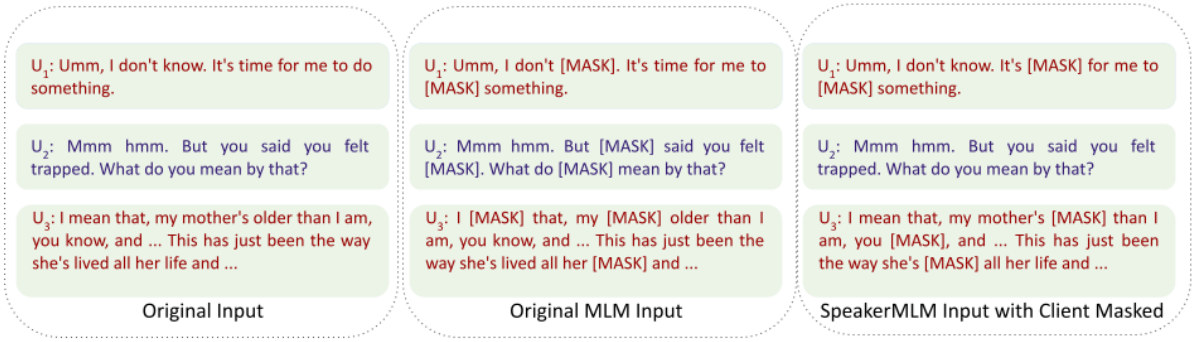

Figure 1: An example dialog snippet and the generated pretraining examples for masked language modeling (MLM) and SpeakerMLM pretraining objectives. Counselor utterances are in red, while client utterances are in purple. Note that here actual token representation of the input sequence is skipped for legibility.

thus, suggesting it is not exclusively relying on the speaker embeddings while predicting a turn-shift.

SpeakerMLM. Although dialog-agnostic MLM has been shown to be effective for many domains and tasks (Devlin et al., 2019), it often fails to leverage potentially helpful domain information, since each non-special token at any position has an equal chance of being masked. In order to augment an MLM with dialog-specific information, we design a masking strategy where masked tokens are selected based on their dialog-specific information. We hypothesize that by forcing the transformer model to infer one speaker's masked tokens from the other speaker's unmasked tokens, the model will learn intersentential and inter-speaker dependencies. More concretely, we start by randomly selecting a speaker with uniform probability and then randomly mask tokens from their utterances with a specific `mask probability`, which is a hyperparameter. The loss used for this task is negative log-likelihood, identical to the original MLM.

3.2 Discourse Coherence Objectives

In addition to the discourse structure pretraining objectives described above, we experiment with Order Recovery and Intruder Detection, two existing pretrained objectives related to discourse coherence that have been found useful in dialog-related tasks (Mehri et al., 2019; Gu et al., 2021b).

Order Recovery. We hypothesize that since the order of utterances is crucial in determining its overall meaning, learning to recover their original permutation may lead to a better contextual representation of a medical conversation. Our modeling approach is similar to Gu et al's (Gu et al., 2021b), but instead of training the model to minimize the KL-divergence between the approximated rank-1 probability and the permutation probability, we minimize the cross entropy between them, following ListNet (Cao et al., 2007). We pass the `[SEP]` embeddings to a feed-forward network with a final sigmoid layer to derive a relevance score for the utterance ranking with respect to the text order.

Intruder Detection. Intruder detection, also known as inconsistency identification (Mehri et al., 2019), seeks to model the coherence of utterances. The goal of this pretraining task is to identify the "intruder" i.e., an utterance that does not belong to the original dialog snippet. We generate intruder detection examples by randomly selecting an utterance i from $[1, 2, \cdots, k]$, and replacing it Utt_i with a negative sample randomly chosen from the pool of all utterances spoken by the same speaker. Note that the negative example cannot be Utt_i itself. As in order recovery, the `[SEP]` embeddings corresponding to each sentence in the dialog snippet are obtained using a feed-forward network that uses cross-entropy loss.

3.3 Multiphase Adaptive Pretraining

An important design aspect of pretraining strategies is the way unlabeled dialog corpora is used during pretraining. While pretraining language models on large, general-topic corpora such as Wikipedia articles have been found useful for general-purpose dialog, pretraining on in-domain or unlabeled target-domain presents an additional opportunity for leveraging unlabeled corpora. This is particularly relevant for the clinical and psychotherapy domains, where large collections of domain data (or annotated data) are often not readily available. To address this, we experiment with two main strategies while using unlabeled data: domain-adaptive

pretraining (DAPT) and task-adaptive pretraining (TAPT) (Gururangan et al., 2020). The first allows the model to access a set of unlabeled texts semantically and stylistically similar to the target domain so additional performance gain can be achieved on the in-domain corpus. The second enables pretraining on unlabeled target corpora (Gururangan et al., 2020).

During our experiments we use BERT (Devlin et al., 2019), a popular choice for the contextual embedding of text sequences, as our backbone architecture. Below, we describe important elements of the architecture that are adapted to conduct our experiments while incorporating the pretraining objectives.

Input Representation. Instead of independently encoding each dialog utterance, we opt for jointly encoding multiple sentences in a local neighborhood. Hence, we directly encode contextual information from surrounding utterances as opposed to a single encoding approach, which requires an extra step to contextualize single representations. Previous works such as (Cohan et al., 2019) have shown that context-augmented representations can improve the model performance on sequential sentence classification (SSC) tasks for conversational domains. The main advantage of this strategy is that a separate contextualization step is not necessary and the resulting representation can be directly fed into a feed-forward network for classification. Thus, it allows classifying multiple sentences at a time. Specifically, we set a context window of fixed size k and concatenate all the utterances in the window, separated by special tokens. Thus, a sample sequence given k consecutive utterances spoken in a dialog snippet will be represented as shown below:

$$\texttt{[CLS]}\ \text{Utt}_1\ \texttt{[SEP]}\ \text{Utt}_2\ \texttt{[SEP]}\ \cdots \text{Utt}_k\ \texttt{[SEP]}$$

where Utt_i refers to the sequence of tokens for utterance i, and [CLS] and [SEP], respectively denote the special tokens for classification and utterance separation.

Speaker Embeddings. While the original BERT uses segment embedding to distinguish multiple sentences, we choose to use speaker embeddings to focus on dyadic conversations only. The speaker embedding layer maps two speakers to a learnable embedding in the hidden dimension space. This approach is similar to (Gu et al., 2020), but instead of directly modeling end of turns with an additional token, we provide the relevant turn information through the speaker embedding.

Adaptation to Different Pretraining Objectives. To adapt the model to different pretraining objectives, we add a task-specific feed-forward layer and an activation layer if necessary. Likewise, during training on downstream tasks, a feed-forward layer is added after the last layer of the encoding model so that the model can be fined-tuned to classify an utterance label from [CLS] embedding for forecasting, and from [SEP] embeddings for jointly categorizing.

4 Datasets

We evaluate our pretraining approach using two datasets portraying clinical interactions between patients and their care-providers (Min et al., 2020; Pérez-Rosas et al., 2016) and also a general-purpose chit-chat dataset.

GRADE Clinical Conversations. This dataset consists of clinical conversations from a large diabetes study (Nathan et al., 2013). The conversations are conducted in English and portray interactions between a diabetic patient and a care provider during the patient's regular check-up for diabetes management. The dataset is annotated at the utterance-level with eight diabetes-related codes, covering a range of medical and diabetic-specific topics including "Not Applicable" code for any other topic (Min et al., 2020).

Motivational Interviewing Dataset. This dataset consists of 277 motivational interviewing (MI) counseling sessions also in English, compiled by (Pérez-Rosas et al., 2016). It includes utterance-level annotations for ten behavioral codes from the Motivational Interviewing Treatment Integrity (MITI) coding scheme, the current standard for evaluating MI counseling fidelity and quality (Moyers et al., 2016). The behavior codes indicate the counseling strategy employed in each counselor utterance. In addition to the 10 behavioral codes, we include two generic codes, one for all client utterances and another for counselor utterances with no counseling strategy assigned, resulting in a total of 12 codes.

DailyDialog. In addition to medical corpora, we use the DailyDialog dataset, a corpus of human-written dialogues covering general-domain and

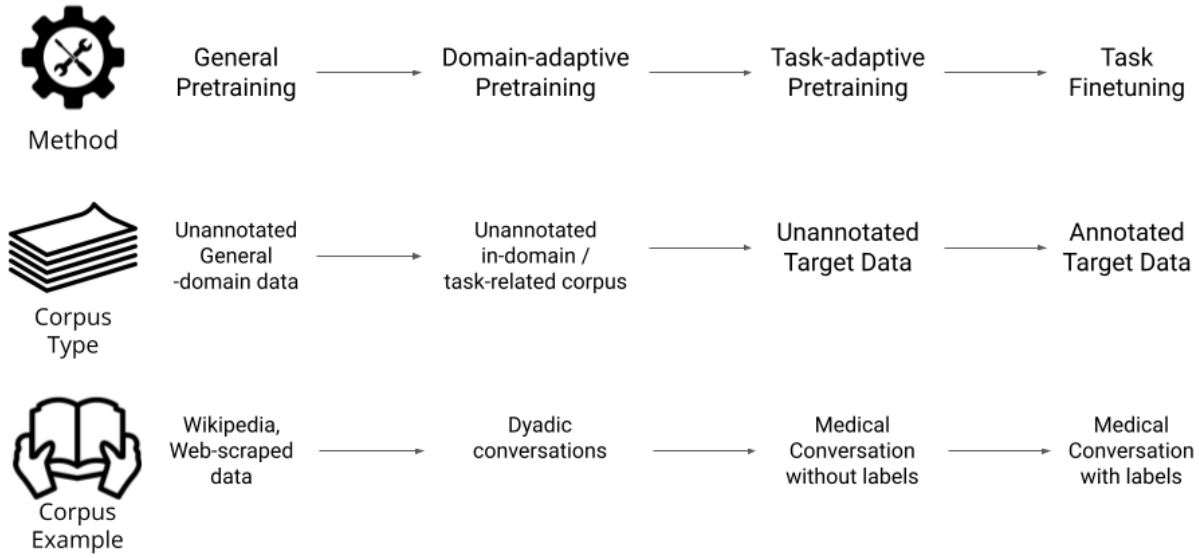

Figure 2: An overview of our multiphase pretraining framework for medical conversations. We start with an pretrained model previously trained on large amounts of general-domain data such as Wikipedia or BookCorpus (Zhu et al., 2015). We then apply domain-adaptive pretraining (DAPT), followed by task-adaptive pretraining (TAPT). Finally, the model is finetuned on the training set of the target task.

chitchat topics such as school life or personal finances (Li et al., 2017). We use DailyDialog as an outer-domain corpus related to the target task in terms of data format or genre (dialogs), but domain distant i.e., general daily life vs medical conversations.

Since all datasets are originally in long full-session length form, we use a sliding window of five conversational turns to segment the sessions into smaller units. Table 1 shows overall statistics for the three datasets.

	GRADE	MI	DailyDialog
# conversations	56	277	13118
# examples	14195	48157	49486
# Finetune Labels	8	12	NA

Table 1: Dataset statistics of GRADE, MI, and DailyDialog datasets

5 Experiments

We focus on two utterance classification tasks: categorization and forecasting, formulated as shown in Table 2. For the categorization task, we seek to label utterances in a medical conversation, where the set of labels is a predefined set of speaker behaviors or conversation topics. In the forecasting task, we use the same set of labels but seek to forecast the label for an upcoming utterance based on previous utterances. Our choice of these tasks is motivated by the hypothesis that if our dialog-aware pretraining objective leads to models with a better contextual representation of neighboring utterances, that improvement will translate into a higher performance boost for forecasting tasks than in categorizing.

Utterance Classification Tasks
Categorizing
Input: $u_1, u_2, \cdots, u_n \xrightarrow{\text{predict}}$ Target: $c_1, c_2 \cdots, c_n$
Forecasting
Input: $u_1, u_2, \cdots, u_n \xrightarrow{\text{predict}}$ Target: c_{n+1}

Table 2: Comparison of categorizing and forecasting tasks. u_i denotes an utterance i in an example. c_i denotes the target label for u_i.

5.1 Experimental Setup

The experiments are run on a GeForce RTX 2080 Ti. For MLM and SpeakerMLM, we set the masking probability of each token to 0.15 and 0.30 respectively. We set the masking probability for SpeakerMLM as 0.30, since by selecting only one of the two speakers to mask we are asymptotically masking only half of the total utterances, in contrast to MLM. Our evaluations are conducted with

5-fold cross-validation. For training in both pretraining and fine-tuning, we use a sliding window with stride $= 1$ to maximize the model learning opportunities. During test time, we partition the dataset with a window of the same size.

We chose BERT as our base architecture since pretrained parameters fine-tuned on large natural language corpora are readily available, and also because due to its design the additional context input could easily be supplied through the use of separate token type ids. We used the `bert-base-uncased` model implemented in (Wolf et al., 2020) with a learning rate of 2e-5. The input to the model is a sequence of token-level embeddings of each utterance in the conversation and the predicted label is assigned using a multilayer perceptron.

5.2 Fine-tuning Strategy

While DailyDialog is a conversational dataset, its topic, and semantic content is generally not domain specific like GRADE or MI. Thus, we use this dataset as a precursor pretraing corpus to DAPT and TAPT, hoping to maximize the gains from domain-adaptive pretraining by creating a conceptually "closer" stepping stone from the Wikipedia-trained weights of `bert-base-uncased`.

We process each conversation in the datasets to transform the long sequence into a set of smaller dialog snippets of size k measured in the number of utterances.

5.3 Utterance Classification Experiment

We evaluate the pretrained models on both categorizing and forecasting tasks. We experiment with two baselines: (1) the same transformer model described above with no pretraining (No) and (2) standard MLM pretraining (MLM). We conduct a set of experiments where we fix the pretraining method but vary (1) the composition of pretraining corpus (2) the pretraining strategy, and whether to use multi-phase adaptive pretraining or not. Results are shown in Table 3 and Table 4. In the tables, "Small" indicates that models are trained using only the target corpus, whereas under the "Mixed" setting models are trained on the shuffled and combined corpus of all the three corpora, and intended as a control against the "Multi" setup which uses the multi-phase pretraining on all corpora. For the "Multi" evaluation, the pretraining order is DailyDialog $\rightarrow$ Non-target Corpus $\rightarrow$ Target Corpus.

From these results, we note that overall, multi-phase adaptive training ("Multi") achieves the highest performances, but does not always lead to performance gains. For instance, as seen in the performance degradation in MI categorizing tasks, the domain and target adaptive strategies actually lead to lower performance, especially when the same setup with "Mixed" pretraining strategy resulted in performance degradation or stagnation. We hypothesize that multi-phase pretraining amplifies the effect of pretraining objectives hence leading to a higher performance boost than when using the "Mixed" setting only. This indicates that choosing the right pretraining schedule/strategy is important but doesn't provide the full recipe for successful pretraining.

We also observe that across datasets and tasks, MLM and SpeakerMLM perform consistently higher than other pretraining methods. However, we see a clear difference in task performance for the GRADE and MI datasets. Particularly, discourse-aware objectives (Order Recovery, Intruder Detection) in MI tasks achieve comparable or higher performances in both categorizing and forecasting. In one notable instance, intruder detection achieves the highest score with multi-phase training.

This trend is in line with existing discourse-aware pretraining work (Mehri et al., 2019; Santra et al., 2021) suggesting that pretraining tasks that require the model to infer how local utterances are related to each other can benefit from explicit intersentential pretraining approaches. In our case, MI tasks focus on counselor strategy and verbal behavior rather than the semantic content of the utterance, whereas the GRADE task is about the utterance topic. In the former, correct classification relies not only on the content of the target utterance but also on the surrounding context.

Moreover, another comparison can be made along the categorizing vs forecasting axis for both datasets. While MLM outperforms SpeakerMLM in categorizing tasks, SpeakerMLM performs best in forecasting tasks. We believe that forecasting represents a more intersentential task since the model has no access to the target utterance and has to rely only on the context utterances for classification. This may explain why SpeakerMLM outperforms MLM in forecasting despite employing a similar principle and using a similar amount of compute (15% of tokens).

	Categorizing Acc			Forecasting Acc		
No Pretraining	65.57			60.03		
Objective / Corpus	Small	Mixed	Multi	Small	Mixed	Multi
MLM	65.67	66.14	**66.72**	61.03	61.76	62.09
Turn-shift Prediction	52.83	51.45	51.46	51.53	51.63	51.63
Order Recovery	65.00	55.53	55.21	57.79	53.83	52.76
Intruder Detection	62.18	62.67	61.56	53.26	58.73	57.40
SpeakerMLM	64.76	66.11	66.41	61.09	61.03	**62.50**

Table 3: Performance of pretrained BERT contextual embeddings on the GRADE topic classification task

	Categorizing Acc			Forecasting Acc		
No Pretraining	76.87			69.58		
Objective / Method	Small	Mixed	Multi	Small	Mixed	Multi
MLM	77.17	77.25	77.07	69.81	69.61	69.91
Turn-shift Prediction	76.85	73.92	73.92	69.57	69.60	69.63
Order Recovery	77.12	76.94	76.78	69.80	70.11	**70.22**
Intruder Detection	57.07	50.54	**83.31**	69.59	69.57	70.18
SpeakerMLM	77.18	76.82	76.39	69.50	69.47	70.18

Table 4: Performance of pretrained BERT contextual embeddings on the MI behavioral coding task

5.4 Evaluation Under Low Resource Settings

We also conduct experiments to evaluate the proposed pretraining strategies on downstream task performance in low-resource settings, where available supervised learning data is limited in quantity.

We measure the performance of pretrained models when using incremental amounts of finetuning data: 0.01, 0.1, and 0.5. We limit our experiments to the multi-phase adaptive setting ("Multi") and No Pretraining, MLM, and SpeakerMLM objectives.

Results are shown in Table 5 and Table 6. Overall, results indicate a similar trend to experiments conducted with all available data, with SpeakerMLM showing a better performance in the forecasting task. In addition, we find that the No Pretraining model has similar or better performances as MLM methods in lower resource settings (0.01, 0.1, 0.25), which is in contrast to the full-resource setting result. This suggests that domain-specific pretraining does not always lead to robust performances under lower resource settings, especially when finetuning is required to improve the model performance.

6 Is Our Finding Still Relevant in the Era of LLMs?

Recently, there have been significant advances in large language models (LLMs), which contain more than several hundred billion parameters and exhibit state-of-the-art performances on several natural language benchmarks, or even academic and professional tasks (Chowdhery et al., 2022; OpenAI, 2023). Given this development, the relevance and need for NLP research that focuses on and optimizes smaller-scale models, such as this work, may be questioned.

We believe that research on the optimization and development of smaller-scale models will still play an important role in NLP research and application. First, the ownership and control of a language model can be important, especially to organizations that handle and process medical data, which is a focus of this work. Such organizations curate patient data with sensitive information, and as such, feeding the data to LLMs, often only available through APIs, may cause legal, ethical, or security violations. Moreover, because LLMs are often trained with large amounts of labeled data, they often underperform task-specific finetuned models that use fewer parameters (Lehman et al., 2023). Thus, leveraging small to mid-size datasets for finetuning remains a viable option.

7 Conclusion & Lessons Learned

In this work, we studied the performance of pretraining strategies on utterance classification in the medical field, a domain that often suffers from a lack of large, publicly available datasets. We evaluated existing and novel dialog-aware and inter-

	Classification Acc				Forecasting Acc			
Objective / Data Fraction	0.01	0.1	0.25	0.5	0.01	0.1	0.25	0.5
No Pretraining	12.04	51.46	51.44	55.23	7.40	51.66	51.56	51.60
MLM	4.60	51.46	52.04	57.99	8.30	51.63	51.56	52.53
SpeakerMLM	12.30	51.46	52.54	**59.22**	8.70	51.63	51.56	**52.56**

Table 5: GRADE Low-resource evaluation of Multi-setting pretrained models using incremental amounts of fine-tuning data.

	Classification Acc				Forecasting Acc			
Objective / Data Fraction	0.01	0.1	0.25	0.5	0.01	0.1	0.25	0.5
No Pretraining	54.86	70.96	70.99	72.98	36.78	53.44	69.38	69.54
MLM	49.76	70.96	70.99	**74.74**	31.66	49.53	51.44	69.17
SpeakerMLM	53.10	70.82	70.99	71.20	33.60	49.53	51.98	**69.56**

Table 6: MI Low-resource evaluation of Multi-setting pretrained models using incremental amounts of finetuning data.

sentential pretraining objectives, and we showed that multi-phase adaptive training can effectively harness unlabeled data based on task similarity and relevance. We derive several lessons and further directions from our findings:

Pretraining is often beneficial but also has the potential to amplify the negative effects of ill-matched pretraining tasks. Our experimental results confirmed previous works' findings that pretraining strategies to incorporate unlabeled data can be helpful in classification tasks (Devlin et al., 2019; Gururangan et al., 2020). However, we found that using dialog-aware pretraining tasks in medical utterance classification can also lead to poor performance when they are not compatible with the target task.

Pretraining with unlabeled non-target corpora is a useful strategy when the availability of fine-tuning data is limited. Our experimental results showed that pretraining with similar non-target data can boost performance. This is in line with previous findings by Gururangan et al. (2020), showing that after general-domain pretraining on large corpora, additional, domain or target-related training can lead to performance gains. Moreover, we recommend using a multiphase pretraining schedule that uses pretraining corpus on increasing order of task similarity. However, one caveat we observed during our low-resource experiments is that in settings where the amount of fine-tuning data is below a certain threshold, the advantage of pretraining can be limited.

Pretraining with domain-specific data does not result in better performance when compared to domain-agnostic objectives. We implemented several dialog-aware objectives and adapted MLM so that the masking procedure can utilize speaker information assigned to each utterance in conversation. However, we did not see conclusive evidence that these task-specific adaptations led to a significant improvement. Furthermore, in some cases, pretraining with dialog-aware objectives led to a degradation in performance.

Limitations

Our work does not cover the full range of domain-agnostic pretraining objectives, including denoising objectives such as ELECTRA (Clark et al. (2020)), or contrastive objectives, such as SimCSE (Gao et al., 2021; Rethmeier and Augenstein, 2023). This paper focused on comparing the masked language modeling (MLM) objective with specially designed dialog-aware objectives. It is our expectation that, given the empirical findings of this project, task-agnostic general objectives like ELECTRA, or SimCSE, will also outperform dialog-aware methods. In addition, due to the lack of task-related datasets, the set of corpora used during our experiments is limited.

Ethics Statement

The data used for this study was cleaned and anonymized to remove any personal and sensitive information before conducting the reported experiments.

References

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *EMNLP*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Torrey A. Creed, David C. Atkins, and Shrikanth Narayanan. 2021. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *CoRR*, abs/2102.11573.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, page 2041–2044, New York, NY, USA. Association for Computing Machinery.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021a. Mpc-bert: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Online. Association for Computational Linguistics.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021b. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12911–12919.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kang-Min Kim, Bumsu Hyeon, Yeachan Kim, Jun-Hyung Park, and SangKeun Lee. 2020. Multi-pretraining for large-scale text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2041–2050, Online. Association for Computational Linguistics.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models?

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining

methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Do June Min, Veronica Perez-Rosas, Shihchen Kuo, William H. Herman, and Rada Mihalcea. 2020. Upstage: Unsupervised context augmentation for utterance classification in patient-provider communication. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, pages 895–912.

Theresa Moyers, Lauren Rowell, Jennifer Manuel, Denise Ernst, and Jon Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65.

David M. Nathan, John B. Buse, Steven E. Kahn, Heidi Krause-Steinrauf, Mary E. Larkin, Myrlene Staten, Deborah Wexler, John M. Lachin, and the GRADE research group. 2013. Rationale and design of the glycemia reduction approaches in diabetes: A comparative effectiveness study (GRADE). *Diabetes Care*, 36(8):2254–2261.

OpenAI. 2023. Gpt-4 technical report.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining.

Nils Rethmeier and Isabelle Augenstein. 2023. A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives. *ACM Comput. Surv.*, 55(10).

Bishal Santra, Sumegh Roychowdhury, Aishik Mandal, Vasu Gurram, Atharva Naik, Manish Gupta, and Pawan Goyal. 2021. Representation learning for conversational data using discourse mutual information maximization. In *North American Chapter of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhuosheng Zhang and Hai Zhao. 2021. Structural pretraining for dialogue comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5134–5145, Online. Association for Computational Linguistics.

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2021. Natural language processing for smart healthcare. *CoRR*, abs/2110.15803.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

A Appendix

Hyperparameter	Value
Batch Size	32
Optimizer	Adam (`betas`=0.9,0.999)
Learning Rate	2e-5
Weight Decay	0.01
Training Epochs	5
MLM Probability	0.15
Speaker MLM Probability	0.30

Table 7: Training Hyperparameters