Task-specific Objectives of Pre-trained Language Models for Dialogue Adaptation

Junlong Li^{1,2,3}, Zhuosheng Zhang^{1,2,3}, Hai Zhao^{1,2,3,*}, Xi Zhou⁴, Xiang Zhou⁴

Department of Computer Science and Engineering, Shanghai Jiao Tong University
Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
CloudWalk Technology, Shanghai, China

{lockonn,zhangzs}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn, {zhouxi,zhouxiang}@cloudwalk.cn

Abstract

Pre-trained Language Models (PrLMs) have been widely used as backbones in lots of Natural Language Processing (NLP) tasks. The common process of utilizing PrLMs is first pre-training on large-scale general corpora with taskindependent LM training objectives, then fine-tuning on task datasets with task-specific training objectives. Pre-training in a task-independent way enables the models to learn language representations, which is universal to some extent, but fails to capture crucial task-specific features in the meantime. This will lead to an incompatibility between pre-training and finetuning. To address this issue, we introduce task-specific pretraining on in-domain task-related corpora with task-specific objectives. This procedure is placed between the original two stages to enhance the model understanding capacity of specific tasks. In this work, we focus on Dialogue-related Natural Language Processing (DrNLP) tasks and design a Dialogue-Adaptive Pre-training Objective (DAPO) based on some important qualities for assessing dialogues which are usually ignored by general LM pre-training objectives. PrLMs with DAPO on a large in-domain dialogue corpus are then finetuned for downstream DrNLP tasks. Experimental results show that models with DAPO surpass those with general LM pre-training objectives and other strong baselines on downstream DrNLP tasks.

1 Introduction

Recently, Pre-trained Language Models (PrLMs) have shown effective and achieved great performance in a series of Natural Language Processing (NLP) tasks. Some prominent examples are BERT (Devlin et al. 2019), GPT (Radford et al. 2018), ERNIE (Sun et al. 2019b,c), RoBERTa (Liu et al. 2019), ALBERT (Lan et al. 2019) and ELECTRA (Clark et al. 2020). Utilizing them usually follows a *Pre-training then Fine-tuning* strategy. They are first pre-trained on large-scale unlabeled task-independent corpora like WikiTest, WikiEn, and BookCorpus with task-independent training objectives like Masked Language Modeling (MLM)

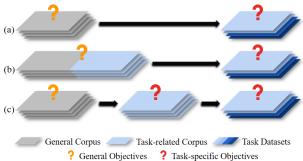


Figure 1: (a) Original PrLM workflow. (b) Existing task-specific PrLM workflow. (c) Our task-specific PrLM workflow with task-specific objectives.

(Taylor 1953) or Next Sentence Prediction (NSP) (Devlin et al. 2019), then fine-tuned on labeled task datasets with task-specific objectives.

Although pre-training on general corpora enables PrLMs to learn universal language representations, it still has limitations if task datasets are too focused on a certain domain (Whang et al. 2019), which cannot be sufficiently and accurately covered by the learned universal language representation. Thus, letting the task-specific factors involved in pretraining in advance has been tried for better downstream task performance. BioBERT (Lee et al. 2020), SciBERT (Beltagy, Lo, and Cohan 2019), and Clinical-BERT (Huang, Altosaar, and Ranganath 2019) pre-train BERT further on texts of Biomedicine, Science and Clinical-Medicine respectively and obtain good performance on the corresponding downstream tasks. DialoGPT (Zhang et al. 2020b) pre-trains GPT further on a large in-domain dialogue corpus, Reddit, and is able to generate responses close to humans on various metrics. A recent study further demonstrates the importance of task-specific pre-training. Gururangan et al. (2020) propose Domain-Adaptive Pre-training (DAPT) and Task-Adaptive Pre-training (TAPT) and conduct experiments on eight classification tasks of four fields.

Despite the success of the above previous studies simply putting task-independent and in-domain task-related corpora together for pre-training, the guideline of task-specific training objective is not well exploited. It is obvious that texts of

^{*}Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011), Huawei-SJTU long term AI project, Cutting-edge Machine reading comprehension and language model.

different domains and forms have different emphases. For example, *Specificity* and *Diversity* are important qualities for assessing dialogues while *Preciseness* and *Logicality* are crucial for academic papers. General language model (LM) pre-training objectives like MLM and NSP focus largely on the *Readability* and *Rationality* of texts based on semantics and syntax, which is destined for ignoring any task-specific features. As this work lays emphasis on Dialogue-related Natural Language Processing (DrNLP) tasks, so dialogue adaptation should be paid special attention and we thus introduce a task-specific objective, which is called Dialogue-Adaptive Pre-training Objective (DAPO) for this purpose. The overall workflow of our method and comparison between existing ones are shown in Figure 1.

Previous research (Mehri and Eskénazi 2020) shows that treating dialogue as a whole is properer and more reliable than treating it as several individual utterances when assessing it, so DAPO is designed on dialogue-level. A new indomain dialogue corpus is first constructed based on four existing dialogue datasets, which are manually proofread with high quality. For each dialogue in the corpus, three negative examples lacking Coherence are generated by Utterance Ordering (UO), Utterance Insertion (UI) and Utterance Replacement (UR) (Barzilay and Lapata 2005; Cervone, Stepanov, and Riccardi 2018; Mesgar, Bucker, and Gurevych 2020) respectively. These negative examples are scored 0, while the original dialogues (i.e., positive examples) are scored 1. The scores of the positive examples are then multiplied by a coefficient, which is called Token Specificity and measured by n-gram Normalized Inverse Document Frequency (n-NIDF). Through this step, scores of all examples range from 0 to 1 and can additionally reflect the Specificity and Diversity of dialogues while retaining the measurement of Readability and Coherence. After getting the final scores, we input the examples into PrLMs, and the outputs are mapped to 0-1 to get the prediction values. Models are trained with a loss function for the regression task.

Models with DAPO are then fine-tuned on several down-stream DrNLP tasks, including Dialogue-based Question Answering (DbQA), Response Selection (RS), and Dialogue Quality Evaluation (DQE). DREAM (Sun et al. 2019a) is chosen for DbQA and MuTual (Cui et al. 2020) for RS. For DQE, we adopt PERSONA-CHAT (Zhang et al. 2018) and DailyDialog (Li et al. 2017) annotated in a previous study (Zhao, Lala, and Kawahara 2020) and a newly released FED dataset (Mehri and Eskénazi 2020). Experimental results show that DAPO helps models achieve new state-of-the-art performance on some of the tasks and gain significant improvements on the others compared with models pretrained with general LM pre-training objectives on the same in-domain dialogue corpus and other strong baselines.

2 Background and Related Works

2.1 Pre-training Objectives

BERT (Devlin et al. 2019) adopts Masked Language Modeling (MLM) as its pre-training objective. MLM is also referred as a *Cloze* task. It first masks out some tokens from the input sentences and then trains the model to predict them

by the rest of the tokens. There are also some similar derivatives of MLM like Permuted Language Modeling (PLM) in XLNet (Yang et al. 2019) and Sequence-to-Sequence MLM (Seq2Seq MLM) in MASS (Song et al. 2019) and T5 (Raffel et al. 2019). Next Sentence Prediction (NSP) is another widely used pre-training objective. It trains the model to distinguish whether two input sentences are continuous segments from the training corpus. Sentence Order Prediction (SOP) is one of the replacements of NSP. It requires models to tell whether two consecutive sentences are swapped or not and is first used in ALBERT (Lan et al. 2019). Replaced Token Detection (RTD) is also used by recent PrLMs like ELECTRA (Clark et al. 2020) with a similar idea used in Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), which requires models to predicts whether a token is replaced given its surrounding context.

2.2 Task-specific Pre-training

Recently, some task-specific PrLMs have been proposed, such as BioBERT (Lee et al. 2020) for biomedical texts, SciBERT (Beltagy, Lo, and Cohan 2019) for scientific texts and Clinical-BERT (Huang, Altosaar, and Ranganath 2019) for clinical texts. These models are pre-trained further on the basis of BERT with large in-domain task-related corpora. DialoGPT (Zhang et al. 2020b) chooses to pre-train models on the basis of the GPT-2 (Radford et al. 2018) architecture with a large in-domain dialogue corpus, and achieve great performance on generating dialogue responses. There are also researchers who directly pre-train PrLMs further on the downstream task datasets. Whang et al. pre-train BERT further on Ubuntu (Lowe et al. 2015) and then fine-tune it on the same dataset for response selection. All these works show positive results for task-specific pre-training. Gururangan et al. (2020) recently summarized the current task-specific pre-training methods and divided them into two subclasses: Task-Adaptive Pre-training (TAPT) and Domain-Adaptive Pre-training (DAPT), and verify their effectiveness through lots of experiments.

Though with the same general LM pre-training objectives, most of the existing task-specific pre-training methods have shown the capability of model enhancement, pre-training with task-specific objectives customized for the concerned task may still better capture the features in the in-domain task-related corpora. In this work, we focus on Dialoguerelated Natural Language Processing (DrNLP) tasks, and thus propose a task-specific pre-training objective for dialogue adaption, which is called Dialogue-Adaptive Pretraining Objective (DAPO). DAPO is designed to measure qualities of dialogues from multiple important aspects, like Readability, Consistency and Fluency which have already been focused on by general LM pre-training objectives, and those also significant for assessing dialogues but ignored by general LM pre-training objectives, like Diversity and Specificity (See et al. 2019; Ghandeharioun et al. 2019; Adiwardana et al. 2020; Mehri and Eskénazi 2020).

	Train	Dev
# of all examples	1045K	116K
# of positive examples	261K	29K
# of negative examples	784K	87K
avg. # utter. per exmaple	9.84	9.84
avg. # tokens per exmaple	177.09	177.30

Table 1: Data statistics of our in-domain dialogue corpus.

3 Dialogue-adaptive Pre-training Objective

3.1 Pre-training Corpus Construction

Existing large in-domain dialogue corpora such as Reddit (Zhang et al. 2020b) or Ubuntu (Lowe et al. 2015) are directly crawled from the Internet forums without further processing. A considerable proportion of expressions in these large corpora do not follow grammatical standards or even have syntactic errors and spelling mistakes. As a result, we avoid using them and choose to construct a new in-domain dialogue corpus based on four manuallyproofread, medium-size datasets: DailyDialog (Li et al. 2017), PERSONA-CHAT (Zhang et al. 2018), Topical-Chat (Gopalakrishnan et al. 2019) and BlendedSkillTalk (Smith et al. 2020). The total number of dialogues from these datasets is 49,930. Dialogues extracted from these datasets with more than 10 utterances are split into several consecutive, overlapping dialogue segments, while the others stay intact. As a result, all dialogues in our corpus have no more than 10 utterances, and they are regarded as positive examples.

Negative examples are generated through Utterance Ordering (UO) and Utterance Insertion (UI) and Utterance Replacement (UR) (Barzilay and Lapata 2005; Cervone, Stepanov, and Riccardi 2018; Mesgar, Bucker, and Gurevych 2020), which aim at obtaining dialogue examples lacking Readability, Fluency and Coherence. For UO, the order of utterances in dialogues are permuted randomly. For UI, each utterance of dialogue is removed and then reinserted in any possible position except the original one in the dialogue. For UR, one of the utterances in a dialogue is randomly replaced with another utterance that is also randomly selected from another dialogue. By these operations, we have three negative examples for each positive example. The corpus is further split into train and dev sets with a ratio of 0.9:0.1. More detailed statistics of our corpus can be found in Table 1.

3.2 Scoring Examples

We then score all the examples in our in-domain dialogue corpus. All the positive examples have score 1 while the negative ones are given score 0 to distinguish whether a example is right or wrong. The scores of positive examples are additionally multiplied by a *Token Specificity* coefficient to judge their *Specificity*. This coefficient is measured by *n*-gram Normalized Inverse Document Frequency (*n*-NIDF), which is extented from Normalized Inverse Document Frequency (NIDF) (See et al. 2019) and shown effective for reflecting word rareness. The Inverse Document Frequency of

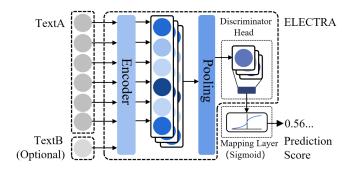


Figure 2: Overall model structure.

an n-gram ng is ${\rm IDF}(ng)=\log(D/c_{ng}^D)$ where D is the number of the original dialogues from 5 dialogue datasets (i.e. D = 49,930), and c_{ng}^D is the number of those dialogues that contain ng. Then Normalized IDF (NIDF) for ng is as follows:

$$NIDF(ng) = \frac{IDF(ng) - min-idf}{max-idf - min-idf}$$
(1)

where **min-idf** and **max-idf** are the minimum and maximum of all IDFs. The n-NIDF of an example e is the weighted mean for all NIDF of n-grams in this example:

$$n - \text{NIDF}(e) = \sum_{ng \in \{e_{ng}\}} \text{NIDF}(ng) * \frac{c_{ng}^e}{|e_{ng}|}$$
 (2)

where e_{ng} denotes all the *n*-grams in this example, and c_{ng}^e denotes the times ng appears in e.

For each positive example, we calculate n-NIDF of it with n = 3, and use it as the *Token Specificity* coefficient. The final scores for examples in our corpus is as follows:

$$\mathsf{score}(e) = \begin{cases} 0 & e \ is \ negative \\ 1*3\text{-NIDF}(e) \in [0,1] & e \ is \ positive \end{cases}$$

This method is able to measure *Coherence* and *Specificity* simultaneously, while does not require any time- and cost-intensive human labeling, allowing us to take full advantage of the large-scale unlabeled corpus.

3.3 Model Implementation

The discriminator of ELECTRA_{large} (Clark et al. 2020) is the PrLM adopted in our work, and referred as ELECTRA for brief in the following statements. Loosely specking, ELECTRA requires a *textA* and an optional *textB* as the inputs. The pooled representations of inputs are then fed into a discriminator head to get the final output. For DAPO, we regard each of our examples as a long text sequence and input it into ELECTRA as *textA* while leave *textB* as blank. To match with the range of scores, a mapping layer is added on top of ELECTRA. It consists of a sigmoid function, which maps the original output of PrLMs to a real number ranging from 0 to 1. The overall structure is shown in Figure 2. During pre-training on our in-domain dialogue corpus, the parameters are updated by mean-square error (MSE) loss:

$$MSE = \frac{1}{b} \sum_{i=1}^{b} (s_i - \hat{s_i})^2$$
 (3)

where b is the batch size, s_i and $\hat{s_i}$ denote the real score and the prediction score of an example respectively.

4 Downstream DrNLP Tasks

4.1 Task Description

We evaluate our method on three kinds of common downstream DrNLP tasks: Dialogue-based Question Answering (DbQA), Response Selection (RS) and Dialogue Quality Evaluation (QDE).

Dialogue-based Question Answering (DbQA) This task requires models to choose the correct answer from some candidate options given a question and a corresponding dialogue. Accuracy is selected as the evaluation metric.

Response Selection (RS) This task requires models to select the best response from some candidates with a given dialogue history. Recall at position n in candidates (R@n) and Mean Reciprocal Rank (MRR) (Voorhees 1999) are set to be the evaluation metrics.

Dialogue Quality Evaluation (DQE) This task has two subdivisions: turn-level and dialogue-level. Each example in the task datasets has one or more qualities with human judgment scores. For turn-level evaluation, models need to evaluate a turn given the previous dialogue history and yield a prediction score; for dialogue-level, each dialogue is evaluated as a whole. Following previous studies, we use Pearson and Spearman correlation to examine whether the prediction scores are correlated with human judgments.

4.2 Task-specific Fine-tuning

The model architecture for fine-tuning on downstream DrNLP tasks is the same as the one for DAPO pre-training shown in Figure 2, with MSE loss still used for parameter updating. Here, we show how to adapt it to DbQA, RS, and DQE tasks.

Dialogue-based Question Answering (DbQA) We input each dialogue into the model as *textA*, and combine the question and an option answer together as *textB*. For each question, inputs corresponding to the right answer have a real score 1, while others are 0. When evaluating, the option with the highest prediction score among the candidates is chosen.

Response Selection (RS) RS tasks follows similar procedures with DbQA, except that the dialogue history and candidate response are regarded as *textA* and *textB* respectively. For each example, inputs corresponding to the best response has real score 1, while others are 0, and candidate responses are sorted by prediction scores from large to small.

Dialogue Quality Evaluation (**DQE**) For the turn-level DQE tasks, the inputs are the same as the RS task, but the prediction score for each example is directly used for evaluation. Dialogue-level DQE tasks leave *textB* unfilled, and input the whole dialogue as *textA* to get the prediction scores.

5 Experiments

5.1 Datasets

We list the datasets used in our experiments in this section. Because of the limited space for paper writing, detailed statistics are elaborated in the Appendix.

Dialogue-based Question Answering (DbQA) We use DREAM (Sun et al. 2019a) as the task dataset. It has 6,444 dialogues and 10,197 questions collected from English exams. Given a dialogue example, there is at least one question, and each question has three candidate options. The most important feature of this dataset is that most of the questions are non-extractive. As a result, the dataset is small but quite challenging.

Response Selection (RS) MuTual (Cui et al. 2020) is selected as the dataset. It consists of 8,860 manually annotated dialogues based on Chinese student English listening comprehension exams and requires models to handle various reasoning problems. Experiments are also conducted on the advanced version of it, MuTual plus, where one of the candidate responses is replaced by a safe response (e.g., Could you repeat that? or I'm really sorry, I didn't catch that.) for each example. If the original right answer is replaced, then the safe response becomes the best one; otherwise, the original positive response is still the best one. The introduction of safe responses makes MuTual plus more challenging than MuTual.

Dialogue Quality Evaluation (DQE) 900 examples from DailyDialog (Li et al. 2017) and PERSONA-CHAT (Zhang et al. 2018) respectively annotated in a previous study (Zhao, Lala, and Kawahara 2020) are chosen for turn-level evaluation. These two datasets provide the human judgments of the *overall quality* for each example. FED (Mehri and Eskénazi 2020) with 125 examples is used for dialogue-level evaluation. Different from DailyDialog and PERSONA-CHAT, FED provides more fine-grained qualities with human judgments, like *Coherent, Consistent, Diverse, and Flexible*. It is noteworthy that FED has no train/dev/test split; therefore, we need to apply our models on it without fine-tuning. Since human judgment scores may have range unmatched with 0-1 (e.g., 1-5 or 0-3), they are uniformly mapped into the range of 0-1 to match the prediction scores.

5.2 Baseline Models

Pre-trained ELECTRA without any further task-specific pretraining is used as one of our baselines. To show the effectiveness of our proposed DAPO, we follow the same steps of our method by MLM and NSP pre-training objectives. The two models (ELECTRA-NSP and ELECTRA-MLM) obtained from these procedures are also used as our baselines. We generate a negative example by replacing the following sentence by a randomly selected one from all the sentences for each utterance in our original corpus with 49,930 dialogues to get the pre-training corpus used for NSP. The pre-training corpus of MLM is directly the original corpus.

Model	MuTual		\mathbf{MuTual}^{plus}		us	Model	DREAM		
	R@1	R@2	MRR	R@1	R@2	MRR		Dev	Test
In Paper (Cui et al. 2020)							In LeaderBoard		
Dual LSTM	0.266	0.528	0.538	0.266	0.528	0.538	BERT	66.0	66.8
SMN	0.274	0.524	0.575	0.274	0.524	0.575	XLNet	-	72.0
DAM	0.239	0.463	0.575	0.239	0.463	0.575	RoBERTa	85.4	85.0
BERT	0.657	0.867	0.803	0.657	0.867	0.803	MMM	88.0	88.9
RoBERTa	0.695	0.878	0.824	0.695	0.878	0.824	ALBERT	89.2	88.5
BERT-MC	0.661	0.871	0.806	0.661	0.871	0.806	DUMA	89.3	90.4
RoBERTa-MC	0.693	0.887	0.825	0.693	0.887	0.825	DUMA+Multi-Task Learning	91.9	91.8
Our Implementation									
ELECTRA	0.887	0.969	0.938	0.826	0.949	0.903	ELECTRA	87.4	87.4
ELECTRA-DAPO	0.907	0.976	0.949	0.827	0.962	0.907	ELECTRA-DAPO	88.0	87.7

Table 2: Results on MuTual, MuTual p^{lus} , and DREAM datasets. Scores in bold are the current state-of-the-art. The results of MuTual and MuTual p^{lus} are for dev set since there is no answer label provided in the test set, we will report the test results after obtaining the numbers from the leaderboard holder.

Model		Daily1	Dialog		PERSONA-CHAT				
	Dev		Test		Dev		Test		
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	
Our Re-running									
BLEU	0.32	0.14^{\dagger}	0.31	0.25	0.35	0.31	0.36	0.35	
ROUGE	0.34	0.22	0.33	0.26	0.36	0.40	0.32	0.43	
METEOR	0.37	0.33	0.33	0.27	0.37	0.48	0.34	0.49	
BERTScore	0.38	0.31	0.37	0.39	0.40	0.49	0.41	0.42	
ADEM	0.28	0.28	0.42	0.45	0.26	0.24	0.25	0.28	
RUBER	0.18^{\dagger}	0.15^{\dagger}	0.36	0.30	0.33	0.34	0.38	0.35	
RoBERTa-eval	0.68	0.71	0.62	0.63	0.72	0.75	0.76	0.77	
Our Implementation									
ELECTRA	0.47	0.50	0.45	0.46	0.44	0.46	0.52	0.52	
ELECTRA-DAPO	0.73	0.71	0.71	0.72	0.74	0.70	0.71	0.74	

Table 3: Pearson and Spearman correlation with human judgements of *overall quality* on DailyDialog and PERSONA-CHAT datasets. All values that are not statistically significant (p-value > 0.05) are marked by \dagger . Scores in bold are the current state-of-the-art. Following (Zhao, Lala, and Kawahara 2020), we divide the two datasets into train/dev/test set randomly with the ratio 0.8:0.1:0.1, and re-run baselines.

The corpora for NSP and MLM pre-training are also split into train/dev set with a ratio 0.9:0.1. We also combine MLM and NSP together (i.e., ELECTRA-MLM+NSP) like BERT (Devlin et al. 2019), and use it as a baseline model. Besides our implementation, our baselines also include the following works. Some of the results are from corresponding leader-boards.

Dialogue-based Question Answering (DbQA) PrLMs: BERT (Devlin et al. 2019), XLNet (Yang et al. 2019), RoBERTa (Liu et al. 2019), ALBERT (Lan et al. 2019). The pooled outputs of the PrLMs are directly used to predict the answer; Matching networks specially designed for multichoice: MMM (Jin et al. 2020), DUMA (Zhu, Zhao, and Li 2020) and DUMA+Multi-Task Learning (Wan 2020), which have complex matching networks for predicting answers.

Response Selection (RS) Individual scoring methods: Dual LSTM (Lowe et al. 2015), SMN(Wu et al. 2017), DAM

(Zhou et al. 2018), BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019). These models scores each response in an example individually; Multi-choice method: including BERT-MC (Devlin et al. 2019) and RoBERTa-MC (Liu et al. 2019), which are multi-choice models that handle all the responses in an example at the same time.

Dialogue Quality Evaluation (DQE) Reference-based metrics: BLEU (Papineni et al. 2002) (we use the best result among BLEU-1,2,3 and 4), ROUGE (Lin 2004), METEOR (Banerjee and Lavie 2005), BERTScore (Zhang et al. 2020a), ADEM (Lowe et al. 2017), RUBER (Tao et al. 2018). These methods evaluate the dialogues with a reference response; Reference-free metric: RoBERTa-eval (Zhao, Lala, and Kawahara 2020), which relies on the powerful PrLM RoBERTa and evaluates dialogues with no reference responses. For FED dataset, we only compare with DialoGPT-eval (Mehri and Eskénazi 2020) since there are no other existing methods that can evaluate so many differ-

Model	MuTual	\mathbf{MuTual}^{plus}	DREAM	DailyDialog Pearson/		PERSONA-CHAT /Spearman	
	R@1/R@2/MRR	R@1/R@2/MRR	Dev/Test	Dev	Test	Dev	Test
ELECTRA	0.887/0.969/0.938	0.826/0.949/0.903	87.4/87.4	.47/.50	.45/.46	.44/.46	.52/.52
ELECTRA-DAPO	0.907/0.976/0.949	0.827/0.962/0.907	88.0/87.7	.73/.71	.71/.72	.74/.70	.71/.74
ELECTRA-DAPO w/o TS	0.898/0.975/0.944	0.819/0.945/0.899	87.4/87.5	.66/.67	.69/.70	.63/.65	.66/.72

Table 4: The results of ablation study for DAPO. w/o TS refers to without Token Specificity.

Model	MuTual	\mathbf{MuTual}^{plus}	DREAM	1 DailyDialog PERSO Pearson/Spearman		NA-CHAT	
	R@1/R@2/MRR	R@1/R@2/MRR	Dev/Test	Dev	Test	Dev	Test
ELECTRA-DAPO (1-NIDF)	0.904/0.980/0.949	0.831/0.940/0.904	87.6/87.5	.73/.65	.55/.60	.63/.60	.65/.69
ELECTRA-DAPO (2-NIDF) ELECTRA-DAPO (3-NIDF)	0.903/0.973/0.947 0.907/0.976/0.949	0.819/0.958/0.902 0.827/0.962/0.907	88.1/87.8 88.0/87.7	.77/.73 .73/.71	.65/.69 .71/.72	.65/.65 .74/.70	.71/.73 .71/.74

Table 5: The results for DAPO with different *n*-NIDF as *Token Specificity*.

ent qualities to the best of our knowledge.

5.3 Implementation Details

Our code is written based on Transformers¹. Some baseline models used in DQE tasks are from Zhao, Lala, and Kawahara(2020), Zhang et al.(2020a), and Sharma et al. (2017).

We use Adam (Kingma and Ba 2015) for parameter updating with ϵ = 1e-8 and no weight decay. The learning rate of our task-specific pre-training (DAPO, MLM and NSP) is 1e-5, batch size per GPU is 10, warmup rate is 0.1, and the max sequence length is 512. We train 5 epochs on our indomain dialogue corpus to get the pre-trained models. Then they are fine-tuned for 8 epochs on each downstream DrNLP tasks with learning rate, batch size per GPU, and warmup rate the same as pre-training. All our experiments are conducted on 6 NVIDIA V100 GPUs.

5.4 Main Results

Tables 2-3 show the results on DREAM, MuTual, MuTual^{plus}, DailyDialog and PERSONA-CHAT. We see that ELECTRA-DAPO gives substantial gains over the strong baseline ELECTRA, which demonstrates the effectiveness of designing a task-specific objective for task-specific pre-training.

6 Analysis

6.1 Ablation Study

As mentioned in Section 3.2, we score positive examples in our in-domain dialogue corpus with a *Token Specificity* coefficient. To evaluate the contributions of this factor, we conduct an ablation study by removing it from our method and re-run all the downstream DrNLP tasks except FED. Specifically, an example *e* is scored as follows:

$$score(e) = \begin{cases} 0 & e \ is \ negative \\ 1 & e \ is \ positive \end{cases}$$

The results are shown in Table 4. Since the weakened scoring strategy still takes some qualities such as *Coherence* and *Fluency* into account, it makes models to boost ELECTRA on various DrNLP tasks. However, without *Token Specificity* which indicates the *Specificity* and *Diversity* of a dialogue example, ELECTRA pre-trained with DAPO becomes less powerful compared with the full model. It also holds intuitively because the complete DAPO leverages more significant qualities of dialogues simultaneously to pre-train.

6.2 The influence of n in n-NIDF

Results in Section 6.1 show the importance of Token Specificity in DAPO, thus it is reasonable to investigate the influence of n when calculating n-NIDF. We further score the examples in our in-domain dialogue corpus with 1 and 2-NIDF as the Token Specificity coefficient, while keeping all the other steps the same, including task-specific pre-training and fine-tuning. Experiments are conducted on all downstream tasks except FED. The results are shown in Table 5. There is no obvious difference between ELECTRA-DAPO with different Token Specificity on DREAM and MuTual datasets, while ELECTRA-DAPO with 1-NIDF as Token Specificity is clearly weaker than the ones with 2-NIDF and 3-NIDF on DailyDialog and PERASONA-CHAT. We further explore the distribution of *n*-NIDF scores. Figure 3 shows the results. It is clear that the distribution of the 1-NIDF score is more concentrated than the one of the 2 and 3-NIDF scores, which leads to a weaker separating capacity of it. This may be a potential explanation for the above observation. It is also found that all the *n*-NIDF scores generally consistent with the normal distribution. To some extent, we believe this reflects the general pattern of human dialogues.

6.3 Explanation of *Token Specificity*

Besides the qualitative description given in Section 3.2, we do further experiments to find a quantitative interpretation for the usefulness of *Token Specificity*. For each FED dialogue example, the 3-NIDF of it is calculated as the *Token Specificity*. Then we evaluate the correlation between this

¹https://github.com/huggingface/transformers.

Quality	DialoGPT-eval	ELECTRA	ELECTRA-DAPO	ELECTRA-MLM	ELECTRA-NSP	ELECTRA-MLM+NSP
Coherent	0.251	-0.213	0.446	-0.382	0.091^{\dagger}	0.146^{\dagger}
Error Recovery	0.165^{\dagger}	-0.221	0.329	-0.277	0.080^\dagger	0.066^{\dagger}
Consistent	0.116^{\dagger}	-0.088^{\dagger}	0.294	-0.332	0.005^{\dagger}	0.258
Diverse	0.449	-0.373	0.312	-0.132^{\dagger}	0.176	-0.108^{\dagger}
Topic Depth	0.522	-0.436	0.334	-0.148^{\dagger}	0.218	-0.080^{\dagger}
Likeable	0.262	-0.315	0.328	-0.255	0.034^{\dagger}	0.048^\dagger
Understanding	0.306	-0.277	0.365	-0.398	0.034^{\dagger}	0.074^{\dagger}
Flexible	0.408	-0.330	0.337	-0.317	0.145^{\dagger}	0.049^{\dagger}
Informative	0.337	-0.395	0.386	-0.192	0.133^{\dagger}	-0.019^{\dagger}
Inquisitive	0.298	-0.200	0.144	-0.216	0.125^{\dagger}	0.079^\dagger
Overall	0.443	-0.361	0.480	-0.318	0.130^{\dagger}	0.080^{\dagger}

Table 6: Spearman correlation with the human judgments of several qualities on the FED dataset. Results of **DialoGPT-eval** are from (Mehri and Eskénazi 2020). All values that are not statistically significant (p-value > 0.05) are marked by \dagger . Scores in bold are the best results.

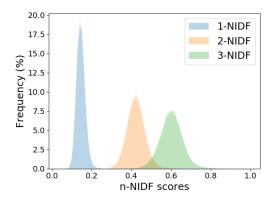


Figure 3: Distribution of *n*-NIDF scores.

Model		MuTual			DREAM		
	R@1	R@2	MRR	Dev	Test		
Baseline	0.887	0.969	0.938	87.4	87.4		
DAPO	0.907	0.976	0.949	87.7	87.7		
MLM	$-\bar{0}.\bar{8}4\bar{7}$	0.955	0.915	86.1	86.9		
NSP	0.902	0.972	0.946	87.6	87.5		
MLM + NSP	0.891	0.964	0.939	86.8	86.2		

Table 7: Results of different pre-training objectives.

value and two human-labeled qualities, *Informative* and *Diverse*. The values of Spearman correlation are 0.129 and 0.179 respectively, indicating this automatically computed coefficient does reflect some vital qualities of dialogues.

6.4 Comparison with Different Pre-training Objectives

We compare DAPO with other general pre-training objectives mentioned in Section 5.2. Tables 6-7 show the results on the downstream DrNLP tasks and quality judgments on FED. According to the results, ELECTRA-MLM, although task-specific pre-trained, has much worse performance than ELECTRA on almost all these downstream

DrNLP tasks. Therefore we argue that MLM is not a suitable pre-training objective for dialogue-based texts. The performance of ELECTRA-NSP is between ELECTRA and ELECTRA-DAPO, which indicates that NSP is a feasible pre-training objective for dialogues. For ELECTRA-MLM+NSP, since it combines a proper objective and an improper objective, it is reasonably that it has performance between ELECTRA-MLM and ELECTRA-NSP.

In addition, we wonder if this kind of difference is caused by an insufficiency of pre-training. The models pre-trained on the train set of pre-training corpus is then evaluated on the dev set of it. The Pearson and Spearman correlation of ELECTRA-DAPO is 0.810 and 0.690 respectively. The accuracy of ELECTRA-NSP and ELECTRA-MLM+NSP are 94.7% and 92.8% respectively. The perplexity of ELECTRA-MLM is 4.96. These values show that models with distinct pre-training objectives are all fully pre-trained. This gives another evidence that the performance of models mainly depends on whether its pre-training objective is suitable for a in-domain task-related corpus.

6.5 Discussions

ELECTRA-DAPO has less improvement in the DbQA task compared with other tasks. We infer that the dialogue and the question-option pair are not syntax-coherent; thus, a processing method merely combining them together is not the optimal choice. According to the FED results in (Mehri and Eskénazi 2020), the inter-annotator agreement is high for all of the dialogue qualities (with Spearman correlation in 0.75-0.85), indicating that these qualities are highly correlated. This partly explains why ELECTRA-DAPO shows good performance on FED dataset even though it uses only one prediction score to measure all the qualities.

7 Conclusion

This paper proposes an effective workflow for utilizing PrLMs with task-specific pre-training between the original two stages: task-independent pre-training and task-specific fine-tuning. We apply this method for dialogue adaption and design a Dialogue-Adaptive Pre-training Objective (DAPO).

We find that that strong pre-trained language models like ELECTRA further pre-trained with DAPO show the superiority over the baseline PrLMs pre-trained with general LM pre-training objectives and other strong baselines on several downstream DrNLP tasks, including dialogue-based question answering, response selection, and dialogue quality evaluation. This work discloses the effectiveness of task-specific pre-training objectives and the potential of further enhancing strong PrLMs with deep customized pre-training settings.

References

- Adiwardana, D.; Luong, M.-T.; So, D.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; and Le, Q. V. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv*:2001.09977.
- Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *IEEvaluation@ACL*.
- Barzilay, R.; and Lapata, M. 2005. Modeling Local Coherence: An Entity-Based Approach. In *ACL* 2005.
- Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP/IJCNLP 2019*.
- Cervone, A.; Stepanov, E.; and Riccardi, G. 2018. Coherence Models for Dialogue. In *INTERSPEECH* 2018.
- Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR* 2020.
- Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; and Zhou, M. 2020. Mu-Tual: A Dataset for Multi-Turn Dialogue Reasoning. In *ACL* 2020.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL* 2019.
- Ghandeharioun, A.; Shen, J. H.; Jaques, N.; Ferguson, C.; Jones, N. J.; Lapedriza, À.; and Picard, R. W. 2019. Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. In *NIPS 2019*.
- Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS 2014*.
- Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; and Hakkani-Tur, D. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *INTERSPEECH 2019*.
- Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *ACL* 2020.
- Huang, K.; Altosaar, J.; and Ranganath, R. 2019. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342*.

- Jin, D.; Gao, S.; Kao, J.-Y.; Chung, T.; and Hakkani-tur, D. 2020. MMM: Multi-stage Multi-task Learning for Multi-choice Reading Comprehension. In *AAAI 2020*.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *CoRR*.
- Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A lite bert for self-supervised learning of language representations. *arXiv:1909.11942*.
- Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.
- Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP 2017*.
- Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL* 2004.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.
- Lowe, R.; Noseworthy, M.; Serban, I.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *ACL* 2017.
- Lowe, R.; Pow, N.; Serban, I. V.; and Pineau, J. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL* 2015.
- Mehri, S.; and Eskénazi, M. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *SIGDIAL 2020*.
- Mesgar, M.; Bucker, S.; and Gurevych, I. 2020. Dialogue Coherence Assessment Without Explicit Dialogue Act Labels. In *ACL* 2020.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL* 2002.
- Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI*.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *JMLR* 2019.
- See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *NAACL 2019*.
- Sharma, S.; Asri, L. E.; Schulz, H.; and Zumer, J. 2017. Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation. *CoRR* .

- Smith, E. M.; Williamson, M.; Shuster, K.; Weston, J.; and Boureau, Y.-L. 2020. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. In *ACL* 2020.
- Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *ICML* 2019.
- Sun, K.; Yu, D.; Chen, J.; Yu, D.; Choi, Y.; and Cardie, C. 2019a. DREAM: A challenge data set and models for dialogue-based reading comprehension. *TCAL*.
- Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019b. ERNIE: Enhanced representation through knowledge integration. *arXiv:1904.09223*.
- Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; and Wang, H. 2019c. ERNIE 2.0: A Continual Pre-training Framework for Language Understanding. In *AAAI 2020*.
- Tao, C.; Mou, L.; Zhao, D.; and Yan, R. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *AAAI 2018*.
- Taylor, W. L. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism & Mass Communication Quarterly*.
- Voorhees, E. 1999. The TREC-8 Question Answering Track Report. In *TREC 1999*.
- Wan, H. 2020. Multi-task Learning with Multi-head Attention for Multi-choice Reading Comprehension. *arXiv*:2003.04992.
- Whang, T.; Lee, D.; Lee, C.; Yang, K.; Oh, D.; and Lim, H. 2019. An Effective Domain Adaptive Post-Training Method for BERT in Response Selection. In *INTERSPEECH* 2020.
- Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; and Li, Z. 2017. Sequential Matching Network: A New Architecture for Multiturn Response Selection in Retrieval-Based Chatbots. In *ACL* 2017.
- Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *NIPS* 2019.
- Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv:1801.07243*.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020a. BERTScore: Evaluating Text Generation with BERT. In *ICLR* 2020.
- Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, W. 2020b. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL* 2020.
- Zhao, T.; Lala, D.; and Kawahara, T. 2020. Designing Precise and Robust Dialogue Response Evaluators. In *ACL* 2020.
- Zhou, X.; Li, L.; Dong, D.; Liu, Y.; Chen, Y.; Zhao, W. X.; Yu, D.; and Wu, H. 2018. Multi-turn response selection

- for chatbots with deep attention matching network. In ACL 2018.
- Zhu, P.; Zhao, H.; and Li, X. 2020. Dual Multi-head Co-attention for Multi-choice Reading Comprehension. *arXiv:2001.09415*.