Representation Learning for Conversational Data using Discourse Mutual Information Maximization

Bishal Santra[♠] IIT Kharagpur
Sumegh Roychowdhury IIT Kharagpur
Aishik Mandal IIT Kharagpur
Vasu Gurram IIT Kharagpur
Atharva Naik IIT Kharagpur
Manish Gupta[♣] Microsoft, India
Pawan Goyal[♦] IIT Kharagpur

[♠]bsantraigi ✉ gmail.com, ♣ gmanish ✉ microsoft.com, ♦ pawang ✉ cse.iitkgp.ac.in

Abstract

Although many pretrained models exist for text or images, there have been relatively fewer attempts to train representations specifically for dialog understanding. Prior works usually relied on finetuned representations based on generic text representation models like BERT or GPT-2. But such language modeling pretraining objectives do not take the structural information of conversational text into consideration. Although generative dialog models can learn structural features too, we argue that the structure-unaware word-by-word generation is not suitable for effective conversation modeling. We empirically demonstrate that such representations do not perform consistently across various dialog understanding tasks. Hence, we propose a structure-aware Mutual Information based loss-function DMI (Discourse Mutual Information) for training dialog-representation models, that additionally captures the inherent uncertainty in response prediction. Extensive evaluation on nine diverse dialog modeling tasks shows that our proposed DMI-based models outperform strong baselines by significant margins.

1 Introduction

Representation learning has transformed how we can apply machine learning to solve real-world problems. However, despite a vast body of research on pretrained language representations, there have been relatively fewer attempts to train representations specifically for dialog understanding. Prior works mostly relied on finetuned representations based on generic models like BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019). In our experiments, we demonstrate that such representations do not perform uniformly across various dialog understanding tasks such as dialog-act classification, intent detection or dialog evaluation.

On the other hand, prior works on pretraining large-scale dialog models focused mainly on open-domain generation. These works evaluated their models only on dialog generation (Zhang et al., 2020; Roller et al., 2021; Adiwardana et al., 2020) or tasks related directly to the pretraining objective (Henderson et al., 2020; Gao et al., 2020). Their effectiveness on other dialog understanding tasks like act classification or intent detection remains unexplored. So we ask the following research question: *Can we learn enriched representations directly at the pretraining phase that are specifically helpful for dialog understanding?*

Existing language modeling (causal or masked) pretraining objectives unfortunately are not the best to model dialogs for these reasons: (1) The model is not directly trained to learn the content discourse structure (e.g., context-response in dialogs). (2) Such models are trained to generate the response word-by-word rather than predicting a larger unit. (3) The inherent one-to-many nature of dialog generation implies that the encoding model should be able to capture uncertainty in the response prediction task, that such models ignore.

Hence, in this paper, we propose pretraining objectives for improved dialog modeling that turn the discourse-level organizational structure of texts from natural sources (e.g., documents, dialogs, or monologues) into a learnable objective. We call this objective the Discourse Mutual Information (DMI). The key insight towards the design of our pretraining objective is to capture representations that can account for a meaningful conversation out of a specific ordered sequences of utterances. We hope that a discourse-level pretraining objective with conversational data would guide the model to learn complex context-level features. For example, in Fig. 1, we illustrate the differences between standard language modeling (causal or masked) based pretraining objectives and a discourse-level reasoning task.

The second research question that we ask is *whether discourse-level features learned using self-supervised pretraining outperform word-level pre-*

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1718 - 1734
July 10-15, 2022 ©2022 Association for Computational Linguistics

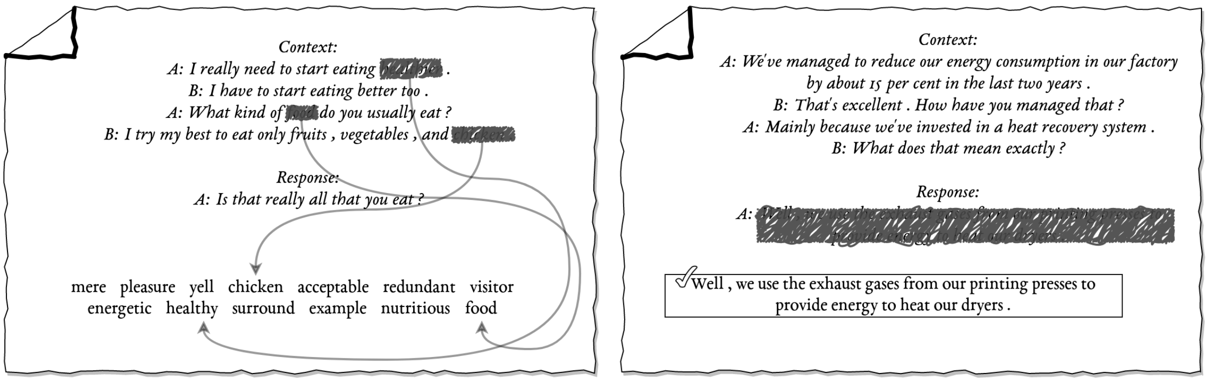

Figure 1: Possible reasoning involved in two types of pretraining: Word-level (left), Discourse-level (right). In a discourse-level reasoning task, the immediately preceding utterance may not be enough for understanding the full context. To predict the correct response, the model will need to capture both the larger context, in this case the topic of discussion, and the intent (e.g., asking for details) of the preceding utterance. In comparison, word-level reasoning is often easier and can be solved using local reasoning. Each of the three masked-words, in the left image, could have been predicted with reasonable confidence without any more information than the utterance itself.

training objectives for downstream dialog understanding tasks. Experimentally, we show that representations learned using the proposed objective function are highly effective compared to both existing discriminative as well as generative dialog models. In terms of various dialog understanding tasks, our models achieve state-of-the-art performances in several tasks (absolute improvements up to 8.5% and 3.5% in task accuracies in probing and finetuning setups, resp.) and perform consistently well across a variety of dialog understanding tasks, whereas baseline models usually have a rather imbalanced performance across tasks.

Overall, our main contributions are as follows.

- We propose DMI, a novel information-theoretic objective function for pretraining dialog representation.

- We release pretrained dialog-representation models in three different sizes (small, medium and base) based on our proposed self-supervised learning objectives[1].

- We extensively evaluate our DMI based representations on multiple open-domain downstream tasks like intent detection, dialog-act classification, response retrieval, dialog reasoning, and response-generation evaluation, and beat state-of-the-art across nine tasks in both probe as well as finetune setups.

[1]To access the pretrained dialog representation models and the source codes, please visit https://bsantraigi.github.io/DMI

2 Literature Review

2.1 Dialog System Pretraining

There have been quite a few efforts towards utilizing existing representations or developing new pretrained models for dialog systems. While BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), GPT-2 (Radford et al., 2019) and other general purpose large-scale pretrained networks are not specific to dialogs, transfer learning from such models could be reasonable. Basic language understanding capability available through these representations helps to get decent performance on many dialog-understanding tasks (Hosseini-Asl et al., 2020).

On the other hand, there have been various works on pretraining dialog specific representations or large-scale generation models. We summarize the properties of various previously proposed dialog-representation learning models in Table 1. DialoGPT (Zhang et al., 2020), Meena (Adiwardana et al., 2020) and Blenderbot (Roller et al., 2021) are large-scale Transformer-based language models, which are trained to generate the gold-response (as per the dataset) given a dialog context. ContextPretrain (Mehri et al., 2019), ConveRT (Henderson et al., 2020) and ConvFiT (Vulić et al., 2021) are trained on the response retrieval task using MultiWoz or Reddit conversations. DEB or Dialog Evaluation using BERT (Sai et al., 2020) is a model based on extended pretraining of the BERT architecture using Reddit data. DialogRPT (Gao et al., 2020), on the other hand, is pretrained to predict human-feedback (e.g., upvotes and downvotes) on comments to Reddit threads. This model is initial-

Model	Training Data Size	Pretraining Obj.	Architecture	Param	Downstream Task
DialoGPT_Small	147M Dialogs	CE	GPT-2	125M	Generation /w MMI
DialogRPT	133M CR pairs	Response Ranking	DialoGPT	345M	Human Feedback Prediction
Blenderbot_Small	1.5B comments	CE	Tr. S2S	90M	Generation
Meena ‡	40B words, 341 GB text	CE	Evolved Tr. S2S	2.4B	Generation
ContextPretrain ‡	10k Dialogs, MultiWoz	NUR, NUG, MUR, I2	HRED	-	Multiwoz (DST, Act, NUG, NUR)
DEB	727M Dialogs	MLM, NSP	BERT	110M	Adv/Random Dialog Evaluation
ConveRT †	727M Dialogs	Response Selection	Tr. Encoder	29M	Response Selection
ConvFiT ‡	8% of 727M Dialogs + Intent data	Response Selection	BERT	110M	Intent Detection
DMI_Base	7.5-10% of 727M Dialogs	InfoNCE-S	Tr. Encoder	124M	9 Dialog-NLU tasks

Table 1: Survey of Pretrained Dialog Models. NUR: next utterance retrieval, NUG: next utterance generation, MUR: masked utterance retrieval, I2: inconsistency identification, CR: Context-response, S2S: Seq2Seq, Tr.: Transformer, CE: Cross-entropy, HRED: Hierarchical RNN Encoder-Decoder. † Pretrained checkpoints available but only for inference. ‡ Both source-code and checkpoints are not available.

ized using the weights of DialoGPT model. Wu et al. (2020) thoroughly investigate these existing pretrained representations, both generic and dialog specific, for understanding their effectiveness on various goal-oriented dialog-understanding tasks.

2.2 Self-supervised Representation Learning with InfoMax

Mutual Information maximization (InfoMax) is one of the popular approaches for self-supervised learning, first used by Oord et al. (2018) and Belghazi et al. (2018). Oord et al. (2018) proposed InfoNCE loss which is an estimator for lower bound to mutual information (MI) between two continuous-valued random variables. InfoNCE has also been used for other NLP applications like training sentence embeddings (SIMCSE (Gao et al., 2021)), question answering (QA-InfoMax (Yeh and Chen, 2019)), etc. Other estimators for mutual information have also been proposed like MINE (Mutual Information Neural Estimator) (Belghazi et al., 2018) and SMILE (Song and Ermon, 2020). In general, these estimators are also broadly studied in contrastive Learning (CL) literature for training both self-supervised (Mikolov et al., 2013; Devlin et al., 2019; Liu et al., 2019; Gao et al., 2021; Henderson et al., 2020; Vulić et al., 2021) and supervised models (Schroff et al., 2015; Gunel et al., 2020). Some prior works in the dialog generation domain have used the concept of mutual information to design loss functions or scoring mechanisms to improve specificity of the generated responses (Li et al., 2016; Yoo et al., 2020). These works predominantly used MI either as a regularizer, along with cross entropy loss, or as a scoring function for ranking generated responses in a post-processing step. In the next section, we derive our pretraining loss function DMI for conversational texts from an information-theoretic perspective.

3 Discourse Mutual Information

We define Discourse Mutual Information (DMI) as the mutual information[2] between two random variables representing two different segments within the same discourse. This is a general concept that can be applied to any form of discourse, no matter the domain or type of signal. In this paper, we focus on dialog type discourses and representation learning for conversational texts. We define two random variables for the contexts (C) and responses (R) that jointly construct a valid conversation. Conversations between humans represent samples from the joint distribution P_{CR} of C and R. We pose the following learning problem, *“learn continuous representations for the textual random variables C and R such that the true mutual information between C and R can be closely estimated.”*

In the remainder of this section we show that, if the lower bound on MI estimated by some representations of context and response is close to the true value, the representation of the context would be as predictive of the response as the natural language form itself. Existing generative training objectives as used in DialoGPT or Blenderbot are extremely focused on predicting target response only. Per-word cross-entropy loss, used for training these models, fails to take into account the inherent uncertainty in the context-to-response generation function. Adapting context representations so as to predict the target responses optimally, helps our proposed DMI-based models learn better dialog representations applicable to a versatile set of dialog understanding tasks.

[2]Mutual Information between two random variables is defined as the reduction in uncertainty/entropy of one of the random variables by having knowledge about the value of the other random variable. Mathematically, this is written as $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

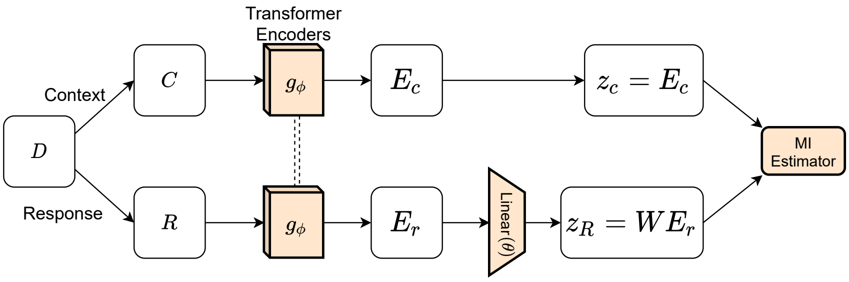

Figure 2: Base Pretraining Architecture for DMI. In our implementations of the model, f_ϕ denotes the transformer (Vaswani et al., 2017) based encoders. Context and response encoders share all parameters for efficient learning. d denotes sample dialogs from the training dataset.

Objective Function Formulation Let E_c and E_r be the representations[3] for C and R based on some encoder. Using the data processing inequality from Information theory (Cover, 1999), we have

$$I(C;R) \geq I(E_c;E_r) \tag{1}$$

This tells us that MI between any encoded version of C and R will always be less or equal than the true mutual information. The equality will hold if E_c and E_r are both fully-invertible encoding processes (as opposed to representations which are lossy or compressive, and inversion is thus not possible). However, neural networks generally embed the data points in a low dimensional manifold by learning robust features that can represent the data points efficiently. Because of this, neural representations are usually not invertible.[4] Now, the exact computation of MI is not possible for continuous-valued random variables. In recent years, various variational lower bounds have been proposed for estimating MI between continuous-valued random variables. Including the MI estimator ($\hat{I}_\theta$), the overall relation becomes

$$I(C;R) \geq I(E_c;E_r) \geq \hat{I}_\theta(E_c;E_r) \tag{2}$$

This leads us to the proposed learning objective DMI:

$$\max_{\theta,\phi} \hat{I}_\theta(E_c^{(\phi)};E_r^{(\phi)}) \tag{3}$$

where $\hat{I}_\theta(E_c;E_r)$ is a variational lower bound estimate of $I(E_c;E_r)$ (Equation 1) parametrized by θ and ϕ denotes the parameters of the encoder used for encoding C (or R) to E_c (or E_r).

[3] C, R, E_c, E_r in caps denote the random variables, whereas the lowercased versions c, r, e_c, e_r denote samples.

[4] One general exception to this is a neural model/representation overfitted on some training data. In such cases, the model may exactly memorize the input/output pairs.

Loss function For training our models, we minimize a loss function depending on the estimator being used.

$$\min_{\theta,\phi} \left[L_{\theta,\phi}(C,R) = -\hat{I}_{\theta,estimator}(E_c^{(\phi)}, E_r^{(\phi)}) \right]$$

We experimented with various MI estimators from literature, namely, MINE (Belghazi et al., 2018), InfoNCE (Oord et al., 2018), JSD (Hjelm et al., 2019) and SMILE (Song and Ermon, 2020). These MI estimators generally compute samples of E_c and E_r using C and R drawn from the joint distribution P_{CR}. Based on our preliminary experiments, we found that InfoNCE estimator produces better representations. The InfoNCE MI-estimate is computed as,

$$I(C;R) \geq \log N - L_N \tag{4}$$
$$L_N = -\frac{1}{2}\mathbb{E}_{P_{CR}} \left[\log \frac{e^{f(c,r)}}{\sum_{r' \in R} e^{f(c,r')}} \right]$$

where N denotes the batch size, and $f(c,r)$ is a scoring function for the $\langle c,r \rangle$ pair.

InfoNCE-S: The original InfoNCE formulation only considers negative samples for one of the random variables, but does not pose any constraint on which of the variables should be considered for negative sampling. As identifying the true response, from a pool of negative samples, would require different reasoning than identifying the true context out of a pool, we consider both these cases and create a symmetric version of the InfoNCE loss function. The final expression of this loss is given in Equation 5 and we refer to it as InfoNCE-S. This considerably improves the speed of training and convergence, and also gives a boost to downstream task performance.

$$L_N = -\frac{1}{2}\mathbb{E}_{P_{CR}}\left[\log \frac{e^{f(c,r)}}{\sum_{r' \in R} e^{f(c,r')}}\right] - \frac{1}{2}\mathbb{E}_{P_{CR}}\left[\log \frac{e^{f(c,r)}}{\sum_{c' \in C} e^{f(c',r)}}\right] \quad (5)$$

For other loss functions, more detailed discussion can be found in the Appendix.

Comparison with ConveRT (Henderson et al., 2020): There are a couple of differences between ConveRT's contrastive loss and our DMI objective. ConveRT models the problem as a response selection task and focuses on modeling cosine similarity between the context and the response. On the other hand, we propose a generic similarity computation function $f(c, r)$ in Eqn. 4 and 5. Another difference is in encoding the input. ConveRT splits the context into previous turns and current query, and encodes them independently. Our model encodes the entire context jointly and hence is capable of better learning the correlations between previous turns and current query.

4 DMI vs. Language Modeling Objectives

In this work, we focus on utilizing DMI for pretraining dialog representations incorporating strong discourse-level features. *But why should the DMI objective learn better discourse-level features than models trained on conversational data using MLM or LM objectives?* We can find the answer by looking at various LM-based objectives through the lens of InfoMax, as shown by Kong et al. (2020). They connected various pretraining objectives for natural language representations, including the ones used for training Skipgram, BERT and XLNet, to the InfoMax learning principle.

If we consider an input text T and a masking function M that returns a masked text $\tilde{T}$ and the masked word w, the MLM objective is equivalent to $\mathcal{L}_{MLM} = -\hat{I}_\theta(E^{(\phi)}(\tilde{T}), e_w)$ where, $E^{(\phi)}$ is the language encoder (e.g., a Transformer encoder) and e_w is the embedding of the token w. Similarly, in the case of auto-regressive LMs like GPT-2, the InfoMax objective equivalent to the loss is $\mathcal{L}_{autoLM} = -\hat{I}_\theta(E^{(\phi)}(T_{1:t-1}), e_{T_t})$, where $T_{1:t-1}$ is the input sequence till $t-1^{th}$ token and T_t is the t^{th} token.

Compared to these LM objectives, DMI focuses on optimizing $I(E_c, E_r)$, where c and r are two structural components from the discourse with designated roles. This enables DMI to discover more important features at the discourse level.

5 Experiments

5.1 Architecture

The exact encoder architecture and the pretraining pipeline has been shown in Figure 2. We use a dual encoder architecture for encoding the contexts and responses separately. We observe that sharing parameters between the two encoders leads to a more efficient learning process and faster convergence. We use vanilla transformer-based encoders[5] (Vaswani et al., 2017) for encoding the natural language inputs. The first tokens for both context and response sequences are the special [CLS] tokens whose contextual embeddings from the encoder are used as the context or response representations. The utterances in the context are delimited by another special token [EOU] (for end-of-utterance). We construct the context using as many utterances from the dialog history as possible up to a maximum of 300 subword tokens. We use the WordPiece tokenizer from BERT for tokenizing the input texts, with a vocabulary size of 30,522.

The scoring function $f(c, r)$ in Eqs. 4 and 5 is implemented using a Bilinear dot product between the context and response representations: $f(c, r) = e_c^T W e_r$ where, W is a square weight matrix trained along with other parameters in the model. This function can take any real value, positive or negative, thus allowing the $\hat{I}_\theta(E_C^{(\phi)}; E_R^{(\phi)})$ function to take any positive real value. While any complicated function with that range could be chosen, we chose this as a simple formulation satisfying the range constraint and left most of the learning to the transformer and the projection matrix W.

5.2 Model Variants

We train three different scales of the DMI model: **DMI_Small** with 6 layers, **DMI_Medium** with 8 layers, and **DMI_Base** with 12 layers. All configurations use 12 attention heads and 768-dimensional embeddings. DMI_Small is initialized with *"google/bert_uncased_L-6_H-768_A-12"*[6], DMI_Medium is initialized with *"google/bert_uncased_L-8_H-768_A-12"*

[5] We implemented all models and experiments using the PyTorch and Huggingface libraries.

[6] These pretrained model tags are from the Huggingface model repository.

and DMI_Base is initialized by weights from *"RoBERTa-base"* pretrained checkpoint, and further pretrained on the pretraining dataset (see §5.4). All of these models are trained using the InfoNCE-S estimator, unless specified otherwise.

5.3 Hyper-parameter Settings

We use Adam optimizer with a linear learning rate schedule for training both the models. Learning rate is first linearly increased to a max value of 5×10^{-5} during the warm-up phase (first 1000 steps). Following this, in the remaining training period, learning rate is linearly decayed down back to zero. Before training DMI_Base, we reset the parameters of the 12th self-attention layer, and it is trained again from scratch along with the weight matrix W using our DMI objective. The embedding layer and initial 11 self-attention layers of the RoBERTa-base encoder are finetuned at a slower learning rate (5×10^{-6}) during our pretraining phase.

As the mutual information value obtained by the InfoNCE loss is upper bounded by $\log(N)$, N being the batch size, we try to keep the value of N as large as possible. Both 8 and 12-layer models are trained on 4-GPU (4x32 GB V100s) systems with overall batch size of 480 and 384, respectively[7]. All the trained models will be publicly shared upon publication.

5.4 Pretraining Dataset

We pretrained all our models using the Reddit corpus (**Reddit-727M conversational-data**) released by Henderson et al., 2019. We ran the scripts released by the authors to recreate the dataset of 727M English conversations. Out of these 727M conversations, we utilize around 7.5% to 10% of the dataset to train our models, after which the validation loss generally saturates. In the rest of this paper, we will refer to this dataset as **rMax**, in short.

Dialog Unrolling for Pretraining For training our models, we need samples of context-response (CR) pairs. Each dialog is unrolled to create context-response pairs with each utterance in the dialog as a response, except the first one. Hence, for each dialog $D = \{U_1, U_2, \ldots, U_T\}$, we generate the following set of samples $S = \{(\mathcal{C}_t : U_1, \ldots, U_{t-1}; \mathcal{R}_t : U_t) : t \in [2, T]\}$. If we process the full rMax dataset, this leads to, approximately, 2.7B CR pairs.

[7]Training time: A maximum of 2 weeks of training time was allowed for 8-layer and 12-layer models. Though, the training process saturates long before the maximum allowed time, and we evaluate our models based on checkpoints when the best validation scores are first obtained.

5.5 MI Estimation

During pretraining, we compare the checkpoints from different epochs and across hyperparameter settings in terms of the bits of mutual information extracted by the trained representation on an unseen set of dialogs. This is calculated as $MI_{valid} = \log(N) - L_N$ (see §3 for more details). As per the Information Bottleneck theory (Tishby et al., 2000), the mutual information learned between the two observed random variables can be factorized into two components, namely, predictive and redundant information. Predictive information generally identifies whether the features learned by the representation are useful for a downstream task. The redundant information is caused by features that do not help in any downstream tasks. Such features can exist due to noise or spurious correlation in the dataset, or even overfitting. Hence, we train our final models on a fraction of the rMax dataset but only for one epoch (i.e., we never repeat the samples) which removes any possibility of overfitting.

Predictive features identified based on a fixed set of downstream tasks (Tishby et al., 2000; Alemi et al., 2017) may not be a sufficient to assess other features learned in the training process. Since, ideally, we want to maximize the amount of predictive information in the representation, we compare the bits of MI on the training set against the bits of MI on an unseen validation set, as captured by the learned representation. To make sure that we do not assume anything about the domain or the conversation topics, we use the validation set of dialogs from the open-domain Daily Dialog dataset (Li et al., 2017).

5.6 Downstream Tasks

Instead of focusing on a single downstream task like many previous works on dialog representation learning, we consider a more versatile range of tasks to evaluate the learned representations from DMI or the baseline models. To find out whether a certain representation is effective for some downstream task, we evaluate in two setups: probe and finetune. In both cases, the pretrained model is used along with an MLP classifier of fixed complexity (Pimentel et al., 2020). In probing setup, we only train the parameters of the MLP classifier.

Task	Description	Train	Valid	Test	Metric
Banking77	Intent 77-class Classification	8,002	2,001	3,080	Accuracy
SWDA	Dialog Act 41-class Classification	213,543	56,729	4,514	Accuracy
MuTual	Reasoning as Response Selection	25,516	2,836	3,544	R@1, R@2, MRR
MuTual Plus	MuTual + Safe response candidate	25,516	2,836	3,544	R@1, R@2, MRR
DD++	Dialog Evaluation	92,590	10,280	11,420	Accuracy
DD++/Adv	Train: Adv. neg., Test: Adv neg. samples	92,590	10,280	11,420	Accuracy
DD++/Cross	Train: Random neg., Test: Adv neg. samples	92,590	10,280	11,420	Accuracy
DD++/Full	Train: All samples, Test: Adv. neg. samples	138,885	10,280	11,420	Accuracy
Empathetic Intent	Emotion and Intent 44-class Classification	25,023	3,544	3,225	Accuracy

Table 2: Downstream task details. Adv.: Adversarial, Neg.: Negative

In finetuning setup, we also train the pretrained model parameters along with the MLP classifier parameters. We use the context and response representations from our models as the input to the MLP classifier.

For downstream tasks, we have two reasoning tasks based on the MuTual dataset (Cui et al., 2020), three classification tasks based on conversational intent detection (Casanueva et al., 2020), emotion detection (Welivita and Pu, 2020) and act classification (Stolcke et al., 1998), and four dialog evaluation tasks based on the DailyDialog++ dataset (DD++, Sai et al., 2020)[8]. Table 2 shows dataset details and metrics for these nine tasks. Both MuTual and DailyDialog++ datasets have an adversarial configuration for the respective tasks, which allows us to assess each of the evaluated models in adversarial settings also.

6 Results and Discussions

6.1 Pretraining DMI based Representations

During pretraining, we used "Validation MI" to evaluate model checkpoints. As the goal of our models is to learn a representation that captures maximum MI between the context and the response texts, this metric tracks how well the learned representation captures the mutual information between contexts and responses of unseen dialogs.

We use the validation split from Daily Dialog dataset as our validation set for evaluation the model during pretraining. It is not specific to a domain and, hence, covers a versatile range of topics. This set comprises 1,000 full conversations between two persons which on unrolling leads to 7,069 context-response (CR) pairs. We illustrate the variation in validation-MI metric against training steps in Fig. 3 in the Appendix.

[8]Note that DailyDialog++ is different from DailyDialog.

6.2 Comparison of Representations on Downstream Task Performance

In this set of experiments, we probe/finetune the DMI models with various downstream tasks that require knowledge of many different types of dialog-understanding features. The results of our probing and finetuning experiments are shown in Table 3.

We have used two types of models as our baselines: generic pretrained models and dialog-specific pretrained models. RoBERTa, BERT, T5 (Raffel et al., 2019), GPT-2 are all trained on large corpora of generic web-crawled English text. But, since these models were not specifically pretrained on any dialog corpus, they may suffer from poor performance on certain dialog understanding tasks. Hence, we consider DialoGPT[9], DialogRPT, DEB and ConveRT models, which were trained on conversational data. For DialogRPT, we used "human-vs-rand" checkpoint released by authors. All models are 12-layer except Blender_Small (8 layers), ConveRT (6 layers), DialogRPT (24 layers) and DMI_Medium (8 layers). We used the publicly available model checkpoints for all baselines, wherever possible. The ConveRT model's checkpoint has been removed from Github[10] by its authors. Hence, it was only possible for us to MLP-probe the representations, without finetuning of the model, based on a cached version released by another user under a valid license[11]. Pretrained checkpoints for Meena, ContextPretrain and ConvFiT are not available, and hence we do not compare with them.

6.2.1 Results in Probing Setup

We observe that, on average, DEB and ConveRT have good performance among the baselines. How-

[9]DialoGPT and DEB are based on GPT-2 and BERT models and were further pretrained on conversations from Reddit. They use the original loss functions of GPT-2/BERT.

[10]https://github.com/PolyAI-LDN/polyai-models

[11]https://github.com/davidalami/ConveRT

		B77	SWDA	E-Intent	MuTual			MuTual Plus			DD++	DD++/adv	DD++/cross	DD++/full
	Model	Acc.	Acc.	Acc.	R@1	R@2	MRR	R@1	R@2	MRR	Acc.	Acc.	Acc.	Acc.
Probing	RoBERTa_Base	72.84	67.18	50.45	**49.70**	**75.20**	**70.00**	43.60	66.60	65.10	55.75	84.20	65.11	68.76
	BERT_Base	72.74	67.99	46.84	45.40	72.80	67.30	42.60	67.70	64.90	60.39	86.56	65.25	72.50
	T5_Base	60.82	68.79	44.50	43.20	69.40	65.60	38.30	65.70	62.20	57.46	84.14	61.23	63.35
	GPT-2_Small	76.64	69.17	49.94	44.92	70.54	66.60	40.75	66.70	63.46	67.37	82.06	67.53	73.93
	DialoGPT_Small	53.00	65.10	43.42	29.80	53.50	55.15	25.51	57.56	54.05	63.63	78.02	**70.61**	70.77
	Blender_Small	70.39	70.11	48.52	41.42	68.06	64.29	42.89	68.85	65.18	60.07	65.14	57.76	68.20
	ConveRT	**89.88**	**71.36**	**55.47**	45.30	72.00	67.00	40.90	**69.00**	64.30	**79.14**	**88.67**	69.59	**80.86**
	DialogRPT	81.54	67.92	50.74	39.50	66.80	63.00	34.20	61.50	59.20	74.11	81.29	68.49	67.20
	DEB	79.18	68.50	45.31	45.10	74.00	67.50	**45.00**	67.70	**66.00**	70.66	86.07	67.25	67.77
	DMI_Small	89.81	72.33	55.72	51.24	74.94	70.76	46.39	70.09	67.14	85.01	91.51	75.75	86.34
	DMI_Medium	90.42	72.49	57.33	52.48	75.62	71.47	46.61	72.46	67.79	85.80	91.38	76.73	86.94
	DMI_Base	91.43	72.73	60.00	52.48	76.41	71.65	48.98	71.33	68.73	86.91	91.98	79.15	88.32
	Δ	1.55	1.37	4.53	2.78	1.21	1.65	3.98	2.33	2.73	7.77	3.31	8.54	7.46
Finetuning	RoBERTa_Base	**92.75**	73.61	**62.81**	48.42	**77.20**	69.70	**49.55**	**73.70**	**69.50**	90.00	95.70	**73.76**	91.09
	BERT_Base	92.27	72.29	60.12	47.86	73.93	68.80	49.10	72.35	69.00	87.05	94.33	67.70	88.82
	T5	89.11	**73.77**	60.66	49.77	73.93	69.80	43.00	66.93	64.90	82.03	90.89	65.85	85.63
	GPT-2_Small	92.49	72.62	58.44	48.42	72.69	68.90	45.71	70.99	67.10	85.69	93.60	68.43	87.83
	DialoGPT_Small	92.59	73.48	59.33	49.32	75.17	69.80	47.86	73.02	68.44	83.68	91.99	64.06	85.54
	Blender_Small	91.59	71.10	58.31	**52.93**	75.85	**71.80**	47.97	70.99	68.30	86.83	92.29	66.39	87.82
	DialogRPT	92.70	72.02	62.13	52.14	76.19	71.40	46.95	70.54	67.66	**90.26**	**95.81**	73.34	**91.25**
	DEB	92.53	72.14	59.69	48.19	74.49	69.00	46.95	70.65	67.80	85.74	94.05	64.42	89.02
	DMI_Small	92.44	71.29	61.05	55.42	75.28	72.92	47.63	72.01	68.19	87.57	94.99	77.33	88.96
	DMI_Medium	92.76	71.53	62.88	55.76	77.88	73.56	50.68	73.25	70.04	89.12	95.63	78.26	90.80
	DMI_Base	93.93	74.50	64.62	56.43	79.91	74.27	52.14	75.06	71.09	91.03	96.39	81.69	92.61
	Δ	1.18	0.73	1.81	3.50	2.71	2.47	2.59	1.36	1.59	0.77	0.59	7.93	1.35

Table 3: Results from probing (top) and finetuning (bottom) setups on 9 downstream tasks for assessing dialog understanding. (DD++: DailyDialog++, B77: Banking77 task, R@k: Recall at k, MRR: Mean reciprocal rank). Our model consistently performs better than SOTA on all the tasks in both probing as well as finetuning setups.

	B77	SWDA	E-Intent	MuTual			MuTual Plus			DD++	DD++/adv	DD++/cross	DD++/full
Model	Acc.	Acc.	Acc.	R@1	R@2	MRR	R@1	R@2	MRR	Acc.	Acc.	Acc.	Acc.
DMI_Base	**93.93**	**74.50**	64.62	56.43	**79.91**	74.27	**52.14**	**75.06**	**71.09**	**91.03**	96.39	76.01	92.61
DMI_Base - Sym	93.28	72.69	**65.18**	**57.34**	77.88	**74.32**	48.08	72.69	68.60	90.94	**96.65**	**76.45**	**93.13**
DMI_Base - RoB	92.34	74.10	60.96	53.84	77.31	72.47	50.34	72.80	69.81	87.23	92.95	73.53	87.85
DMI_Base - Sym - RoB	91.59	73.55	60.71	54.06	75.40	72.24	47.97	71.45	68.24	86.79	92.96	70.29	87.13

Table 4: Ablation study results for the finetune setup for our base model on 9 downstream tasks. "-RoB" → No RoBERTa initialization. "-Sym" → Training with non-symmetric version of InfoNCE. (DD++: DailyDialog++, B77: Banking77 task, R@k: Recall at k, MRR: Mean reciprocal rank).

ever, the RoBERTa model outperforms all other baselines on the MuTual task by a significant margin. In the MuTual Plus task, the DEB model outperforms other models in the R@1 and MRR metrics. ConveRT performs the best among all baselines on the other tasks. ConveRT's loss function is also contrastive in nature and is similar to ours. This explains the model's generally high strength across the tasks among all the baselines.

Our DMI_Base beats ConveRT on all the tasks, and DMI_Medium beats the baseline on 7 out of 9 tasks. We believe DD++ tasks to be the most demanding ones with respect to context-level understanding. Here, all non-dialog baselines have a weaker performance, with DEB and ConveRT being the best of the bunch. These are also the tasks where our models excel the most, with both DMI_Medium and DMI_Base beating all baselines with strong margins. DD++/cross is the most difficult among all four DD++ tasks. Here, the model is trained on random negative samples and tested on a dataset with human-curated adversarial negatives. Our DMI_Base beats the best baseline on DD++/cross by 8.54 points. This shows the superior quality of context representations from our models.

6.2.2 Results in Finetuning Setup

In the finetuning setup, on average, RoBERTa and DialogRPT have good performance among the baselines. DialogRPT performs well for DD++ tasks while Blender works well for the MuTual task. For all other tasks, RoBERTa is the best baseline, even outperforming models especially trained for dialog tasks (like DialoGPT).

Similar to the probe setup, DMI_Base beats baseline methods by significant margins. In general, finetune results are better than probe results across all models, as expected.

Our large-scale RoBERTa-initialized DMI_Base model outperforms the best baseline for all tasks, by a considerable margin. Additionally, our DMI-based models are able to perform well uniformly across all tasks, unlike even baselines like DialoGPT, DialogRPT and Blenderbot models which are explicitly trained on dialog data. This makes DMI the best overall model for dialog related tasks. Across multiple tasks, we show qualitative examples where our proposed DMI-based models provide accurate results, in the Appendix.

6.3 Ablations

We evaluate the importance of using RoBERTa based pretraining as well as the symmetric version of the InfoNCE loss in Table 4. We observe that RoBERTa based pretraining helps significantly across all tasks. The symmetric InfoNCE improves performance for SWDA and MuTual Plus tasks.

7 Conclusions and Future work

In this paper, we proposed the concept of Discourse Mutual Information (DMI) which is better suited for learning dialog-specific features in a self-supervised manner. Using the InfoMax principle we formulated a pretraining method for dialog-specific representation learning. Across 9 downstream dialog understanding tasks, our 12-layer model outperforms state-of-the-art methods. Further, we showed that on most of these tasks, even our 8-layer model outperforms standard 12-layer pretrained models. These experiments show the potential of the proposed DMI objective towards building dialog understanding models. We will make the code and pretrained model checkpoints available on request, instructions can be found here https://bsantraigi.github.io/DMI. Although we experimented only with dialog modeling in this paper, we believe that the proposed DMI objective is generic enough to be applied to any type of discourse in any domain. In the future, we would like to explore how to harness DMI representations for generative conversation modeling.

8 Acknowledgments

This work was partially supported by Microsoft Academic Partnership Grant (MAPG) 2021. The first author was also supported by Prime Minister's Research Fellowship (PMRF), India.

9 Ethical considerations

Like many other pretrained language representation models, the proposed model may also have learned patterns associated with exposure bias. Interpretability associated with the output is rather limited, hence users should use the outputs carefully. The proposed model ranks possible response candidates, and does not filter out any "problematic" candidates. Thus, for applications, where candidate responses could be problematic, (e.g., offensive, hateful, abusive, etc.), users should carefully filter them out before providing them as input to our model.

All the datasets used in this work are publicly available. We did not collect any new dataset as part of this work.

Banking77 Casanueva et al., 2020 has been obtained from https://github.com/PolyAI-LDN/task-specific-datasets. It is available under a Creative Commons Attribution 4.0 International license with details here[12].

SWDA Stolcke et al., 1998: The dataset has been obtained from http://compprag.christopherpotts.net/swda.html. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

E-Intent Welivita and Pu, 2020: The dataset was downloaded from https://github.com/anuradha1992/EmpatheticIntents. The original dataset is available at https://github.com/facebookresearch/EmpatheticDialogues which is under the Creative Commons Attribution 4.0 International license.

MuTual and MuTual-plus Cui et al., 2020: The datasets have been downloaded from https://github.com/Nealcly/MuTual. Licensing is unclear; the authors do not mention any license information or terms of use.

DailyDialog++ Sai et al., 2020: The dataset was downloaded from https://github.com/iitmnlp/DailyDialog-plusplus. The data is available under the MIT License.

[12] https://github.com/PolyAI-LDN/task-specific-datasets/blob/master/LICENSE

rMax or Reddit-727M conversational-data Henderson et al., 2019: the dataset has been obtained from https://github.com/PolyAI-LDN/conversational-datasets/tree/master/reddit. The dataset is available under the Apache License Version 2.0.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. 2018. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *EMNLP*, pages 2161–2174, Online. Association for Computational Linguistics.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Jiaming Song and Stefano Ermon. 2020. Understanding the limitations of variational mutual information estimators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, Carol Van Ess-Dykema, et al. 1998. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *EMNLP*, pages 1151–1168, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *EMNLP*, pages 917–929, Online. Association for Computational Linguistics.

Yi-Ting Yeh and Yun-Nung Chen. 2019. QAInfomax: Learning robust question answering system by mutual information maximization. In *EMNLP-IJCNLP*, pages 3370–3375, Hong Kong, China. Association for Computational Linguistics.

Kang Min Yoo, Hanbit Lee, Franck Dernoncourt, Trung Bui, Walter Chang, and Sang-goo Lee. 2020. Variational hierarchical dialog autoencoder for dialog state tracking data augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3406–3425.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

A Mutual Information Estimators

In this paper, we experiment with various different MI estimators, and found InfoNCE-S to be the best (both in terms of accuracy as well as training speed). The mathematical formulation of these estimators is provided below.

1. InfoNCE was proposed by Oord et al. (2018). It connects to the mutual information value $I(X;Y)$ as,

$$I(X;Y) \geq \log(N) - L_N$$
$$L_N = -\mathbb{E}_{P_{XY}}\left[\log \frac{e^{f(x,y)}}{\sum_{y' \in Y} e^{f(x,y')}}\right]$$

2. MINE (Belghazi et al., 2018)

$$I(X;Y) \geq sup_{\theta\in\Theta}[I_\theta^{(MINE)}(X;Y) = \mathbb{E}_{P_{XY}}[T(x,y)] - \log \mathbb{E}_{P_X \times P_Y}[e^{T(x,y)}]]$$

3. JSD (Hjelm et al., 2019)

$$I_\theta^{(JSD)}(X;Y) = \mathbb{E}_{P_{XY}}[-sp(-T(x,y))] - \mathbb{E}_{P_X \times P_Y}[sp(T(x,y))]$$

4. SMILE (Song and Ermon, 2020)

$$I_\theta^{(smile)}(X;Y) = \mathbb{E}_{P_{XY}}[T(x,y)] - \log \mathbb{E}_{P_X \times P_Y}[clip(e^{T(x,y)}, e^{-\tau}, e^{\tau})]$$

Use of the InfoMax objective for self-supervised learning has been more prevalent in the computer vision domain than in NLP. Although as Kong et al. (2020) have previously shown, many existing loss functions used for training NLP models can be derived directly from the InfoMax framework. Kong et al. (2020) had only focused on various language model objectives that focus on words given the surrounding context. The authors showed that this objective translates to maximizing mutual information between the context and the missing word within the context.

In dialog domain also, InfoMax-equivalent loss functions have been used. First, Henderson et al. (2020) used contrastive formulation of the response selection task as a pretraining objective for dialog representation. Other prior works on response selection models often used a binary-cross entropy loss for training. Both these loss functions are actually equivalent to various lower bound estimators for mutual information. In the QAInfoMax model (Yeh and Chen, 2019), the authors used the DeepInfoMax loss function (Hjelm et al., 2019) as a regularizer and showed that representations learned with or in-presence of an InfoMax regularizer are more resilient to adversarial attacks while maintaining the same level of task performance. We also observe the same effect in our DD++/cross experiments. This is because of the self-supervised yet task-specific nature of the loss function.

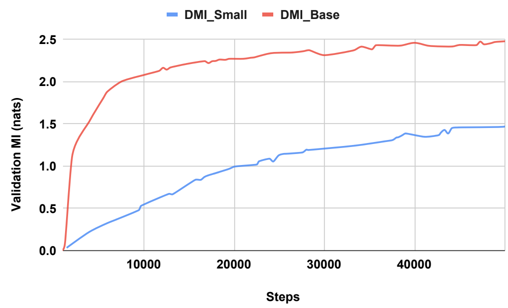

Figure 3: Validation-MI profile during pretraining

B Response Retrieval Experiment

We wanted to investigate if the proposed model can rank good responses higher compared to more generic/bland ones. Hence to test against an extreme setting we simulate a response selection task for a very large pool using the test set of Daily Dialog (Li et al., 2017) dataset. We took all the ∼7000 responses from test set of the daily dialog dataset as the response pool. Next, for a few randomly selected context examples, we illustrate the top two ranked as well as ground truth responses for two full conversations in Tables 5 and 6. Of course, the ground truth response was removed from the pool for each context. The ranking of responses were done using the $f(c, r)$ function from the trained DMI_Base model. From the examples of response selection, we can observe that the model is able to both avoid blend responses and select responses that are relevant to the current context even from such a large pool. This shows the usefulness of dialog specific pretrained representation trained using the DMI objective.

C Prediction Samples and Error Analysis

In Table 7, we show sample predictions from the DailyDialog++/cross task (Sai et al., 2020). As DialoGPT has the best performance in the probe setup, on this task among the baselines, we choose it for error analysis. We randomly sampled 11 instances where the DialoGPT model made a mistake and observed the behavior of our DMI_Base model on these samples. We see that our model correctly predicts for all 6 out of 6 negative samples and out of the 5 positive samples DMI_Base predicts the label of 2 samples correctly (overall 8/11 correct predictions by our model). This shows that our model has a better understanding of the context and response inputs, which makes it robust against the adversarial negative samples. As can be seen in samples 2, 3, 5 and 6, the incorrect predictions by the DialoGPT model might have been caused by presence of common or similar meaning tokens (*cook, food; million; long; employee*) between context and response. This means that DialoGPT often relies on weak token-based cues for prediction.

For error analysis on the Empathetic-Intent (E-Intent) task (Welivita and Pu, 2020), we chose the ConveRT model as the baseline to compare against predictions from our DMI_Base model. First, we randomly select 10 samples from the test set of the E-Intent task where the baseline ConveRT model makes a mistake. Then the predictions from the DMI_Base model are observed on these 10 samples. The input utterances, true labels and the predictions made by the model are shown in Table 8. Out of these 10 samples, DMI_Base is able to predict the labels for 6 instances correctly. We notice that though sample inputs often contain more than one emotion, the one denoted by the gold label is generally the primary one. Our model is able to capture this emotion correctly more often than the baseline, with such mixed-emotion samples.

Fig. 4 shows the confusion matrix for our DMI_Base model for the Empathetic-Intent task. The accuracy is highest for afraid, acknowledging and questioning classes (each above 95%). Some of the most confusing pairs of classes are (annoyed, wishing), (anxious, apprehensive), (caring, confident), (content, grateful), (content, lonely).

Dialog	Rank-1 Response	Rank-2 Response
S1: Hey man , you wanna buy some weed ?	Yes , smoking or non-smoking ?	Sure , what ' s up ?
S2: Some what ?	My cigarettes .	Come on man ! I even got dope and acid ! Try some !
S1: Weed ! You know ? Pot , Ganja , Mary Jane some chronic !	I'm sorry , Mrs . Lin . We don't allow smoking in the office .	Ha ! That's right . You used to smoke . I forgot !
S2: Oh , umm , no thanks .	Come on man ! I even got dope and acid ! Try some !	Come on , you can at least try a little , besides your cigarette .
S1: I also have blow if you prefer to do a few lines .	No , thanks . Oh , yes , could I have some juice for the mini bar ?	No , thanks . I ' Ve had enough . I'll have my bill , please .
S2: No , I am ok , really .	Great ! Would you prefer smoking or non-smoking ?	Come on , you can at least try a little , besides your cigarette .
S1: Come on man ! I even got dope and acid ! Try some !	I will , but let's go to the disco after dinner .	No , thanks . I've had more than enough.In fact , I must be running along .
S2: Do you really have all of these drugs ? Where do you get them from ?	I found them at JC Pennies .	They are my personal effects .
S1: I got my connections ! Just tell me what you want and I ' ll even give you one ounce for free .	Hmm ... I'll tell you what . I'll lend you four thousand dollars , but you have to pay me back next week .	No , thanks . Oh , yes , could I have some juice for the mini bar ?
S2: Sounds good ! Let ' s see , I want .	Here you go.On any other day , it would cost me a fortune , but it ' s on special offer today .	All right , Maria . I ' ll give you until tomorrow at 4:00 to produce a satisfactory piece of work , but otherwise , you ' ll have to re-do it .
S1: Yeah ?	But of course ! Well , it ' s been great talking to you , but I have to get going .	Good . I was thinking that I ' d like to invite you to watch a movie . I can meet you at the cinema gate .

Table 5: Retrieval Example 1: Top two ranked responses from a large pool, as well as ground truth response for a conversation. Note that every line corresponds to one utterance in the conversation either from speaker S1 or speaker S2. **How to read the table:** For any context (all entries in first column upto any row), the ground truth response can be automatically obtained from the immediate next entry in first column. The response selected by the model in any cell $(t, 1 \text{ or } 2)$ is for the true context from row 1 to row t in the first column.

Dialog	Rank-1 Response	Rank-2 Response
S1: Could I have my bill , please ?	Sure . Here is your receipt .	Sure . Your cash back is $ 13 . And we'll bring out your fries in two minutes .
S2: Certainly , sir .	Thanks . Now can I make the full deposit ?	Thank you ! Would you like an aisle seat or a window seat ?
S1: I'm afraid there's been a mistake .	I am really sorry too , maybe I can give you a call sometime .	Oh , I'm sorry.However , if you could help me out , I'll double the pay for the hours worked .
S2: I'm sorry , sir . What seems to be the trouble ?	Not much . I had to pay an unexpected bill , so I needed the money back .	Oh , nothing special . I'm just a bit tired .
S1: I believe you have charged me twice for the same thing . Look , the figure of 6.5 dollar appears here , then again here .	One moment , please , sir . ... Here's your bill . Would you like to check and see if the amount is correct ?	Sir , I deleted the $ 10 , but I had to add a $ 2 service charge to your bill.

Table 6: Retrieval Example 2: Top two ranked responses from a large pool, as well as ground truth response for a conversation. Note that every line corresponds to one utterance in the conversation either from speaker S1 or speaker S2. **How to read the table:** For any context (all entries in first column upto any row), the ground truth response can be automatically obtained from the immediate next entry in first column. The response selected by the model in any cell $(t, 1 \text{ or } 2)$ is for the true context from row 1 to row t in the first column.

ID	Context	Candidate Response	Gold	DialoGPT	DMI_Base
		DMI_Base predicts correctly			
1	All right. I'll take it. __eou__ Do you like to use chopsticks __eou__ Yes, I like using chopsticks.	When you get closer, you see that each horizontal section is made up of two pieces that converge in a right angle.	0	1	0
2	And you'll have to sell your motorcycle. And your cameras. Right? __eou__ Maybe I'll cook once or twice a week. How is that?	I go to the temple twice a week so I prefer vegetarian food.	0	1	0
3	But I heard the box office rose up to 15 million in the first week. __eou__ Box office can't explain everything. I do not think it is cheerful or well-made. The plot is old and the female character is not pretty. __eou__ My sister has given me two tickets for tonight. It is called' The life of Rose', a French movie.	I got 1 million views on my youtube channel in one week.	0	1	0
4	Glad you like it. By the way, is this your first time to China, Mr. White? __eou__ Yes, as a representative of IBM. I hope to conclude some business with you. __eou__ We also hope to expand our business with you.	May I know what and all process you have?	1	0	1
5	Good. I have to go right now. I really hope this meeting doesn ' t last too long. __eou__ They usually go on for ages. __eou__ I ' ll stop by if I have time later. Make sure everyone knows that we must stick to the deadlines.	I don't cut my hair because I really like to keep it long.	0	1	0
6	Of course. The main thing is that all our work must be completed on schedule. We even allow our employee to go home early if they finish their work early. __eou__ How often do you have meetings? __eou__ You should attend a department meeting every Monday morning. There are other meetings for people working together on certain projects. Department heads also attend an interdepartmental meeting each week.	In the newsletter, I gave employees column references this week.	0	1	0
7	Sounds interesting! That must be very convenient. __eou__ Yes, you're right. I can blog wherever and whenever I'm on the move. It's especially good when I'm on a business trip and my laptop happens to be away from me. __eou__ How can you do that?	I sank parents money into my business it is not convenient.	0	1	0
8	There is a wait right now to use the computers. __eou__ That ' s fine. __eou__ Would you please write your name on this list?	Sure, please give me a pen.	1	0	1
		DMI_Base predicts wrongly			
9	How much cash would you like? __eou__ I want $150. __eou__ Here ' s your $150.	Well! I never forget your help.	1	0	0
10	I see, sir. This one is very good. __eou__ Is it? __eou__ You may rest assured. It sells well. __eou__ May I have a look at the introduction?	It has been recommended by top nutritionists.	1	0	0
11	Sir, tell us about your experience with Super Bulk-up. __eou__ Well, it's completely changed my life. __eou__ Tell us how.	The change is right in front of you, isn't it?	1	0	0

Table 7: Sample Predictions from the DD++/Cross task. In DD++/Cross, the models are trained using randomly sampled negatives and tested on curated adversarial negative samples. In each sample, the input context comprises the utterances, previous to the response, spoken by the two participants. Such utterances within a context are delimited by a special token "__eou__".

ID	Input Utterance	Gold Label	ConveRT	DMI_Base
	DMI_Base predicts correctly			
1	i feel very thankful for everything that i have, i live a really good life in my liking	grateful	content	grateful
2	I'm training a new girl at work. She is doing so good for her first week!	proud	confident	proud
3	It broke my heart today when I went to the grocery store and found out that they were out of Dean's French Onion Dip.	disappointed	devastated	disappointed
4	My wife's birthday is coming up. I got her a gift and the party planned out way ahead of time this year.	prepared	surprised	prepared
5	My friend helped me to pack	grateful	trusting	grateful
6	For two years now I've been walking with help of a walker, following a botched hip operation. Recently, at a physical therapy session, I was able to walk with a cane the length of the treatment room. I felt quite good about myself!	proud	caring	proud
	DMI_Base predicts wrongly			
7	I was trying to plan my wedding by getting a caterer, and they kept blowing us off over and over again.	furious	disappointed	disappointed
8	Being a successful single mothr.	proud	content	content
9	We were over at our friend's house for a dinner and I was in the kitchen helping her cook. I had melted butter in a baking dish to make dessert, and I poured cold milk into it like the recipe said to do. It ended up cracking the dish. I felt bad. I offered to buy her a new one.	guilty	caring	ashamed
10	One time I had done really well in a class. I fully expected to get an A in it	anticipating	confident	disappointed

Table 8: Example Predictions on the Empathetic-Intent (E-Intent) task by ConveRT and our DMI_Base model.

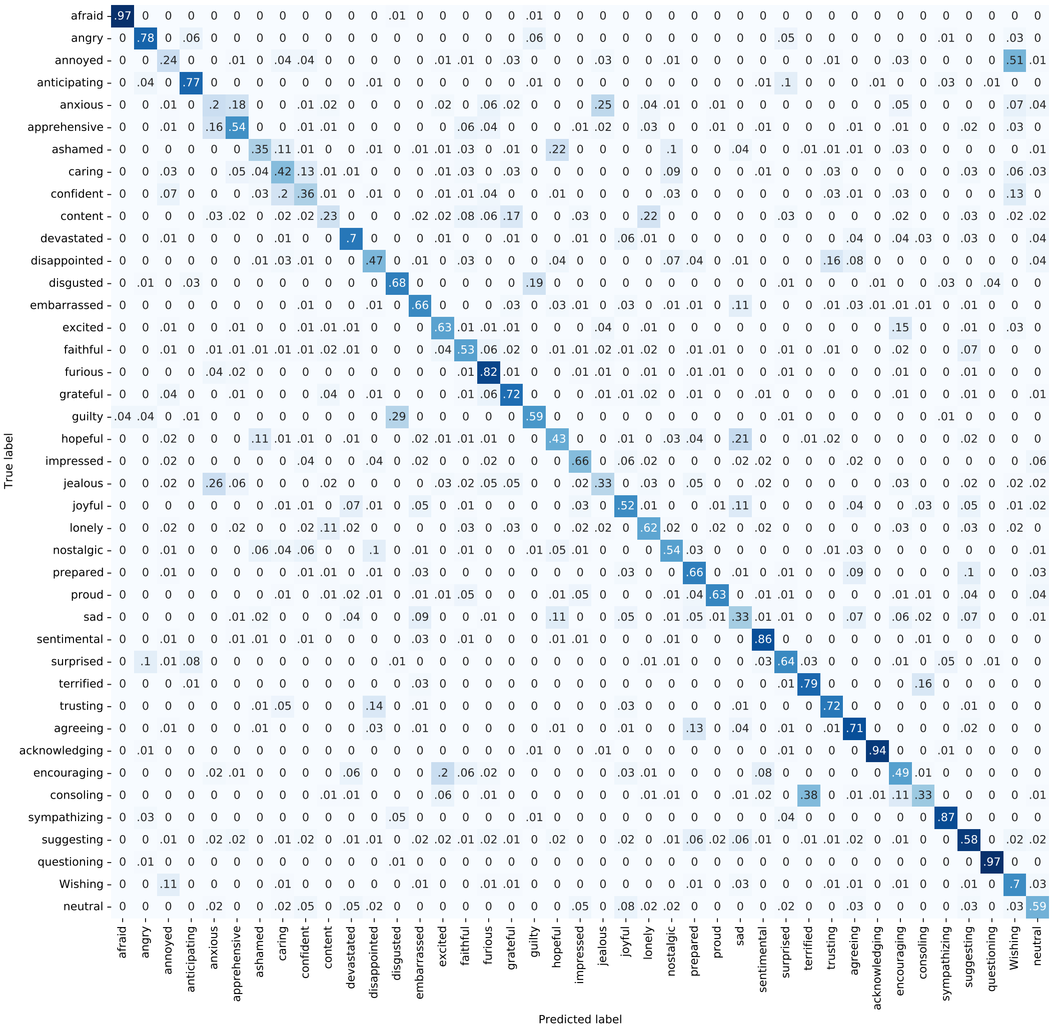

Figure 4: Confusion Matrix for our DMI_Base model for the Empathetic-Intent task.