Methodologies for Aligning Pretrained Dialog Representations to Aspects of Human Judgments

Project-I (CS47007) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Abhinandan De

(19CS10069)

Under the supervision of

Prof. Pawan Goyal

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Autumn Semester, 2022-23

7th November, 2022

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: 9th November, 2022	(Abhinandan De)
Place: Kharagpur	(19CS10069)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Methodologies for Aligning Pretrained Dialog Representations to Aspects of Human Judgments**" submitted by **Abhinandan De** (Roll No. 19CS10069) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2022-23.

Date: 9th November, 2022

Place: Kharagpur

Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Abstract

Name of the student: **Abhinandan De** Roll No: **19CS10069**

Degree for which submitted: **Bachelor of Technology**

Department: **Department of Computer Science and Engineering**

Thesis title: **Methodologies for Aligning Pretrained Dialog Representations to Aspects of Human Judgments**

Thesis supervisor: **Prof. Pawan Goyal**

Date of thesis submission: **7th November, 2022**

Dialog research is a booming field owing to the growing number of dialog systems in today's world. We are in the world of open-domain where dialog agents are expected to converse fluently on a variety of topics while keeping their content engaging and grounded. To facilitate the development of such models, automatic evaluation metrics are the need of the hour. In this project, we come across a new family of metrics and make an attempt to find out whether our model (DMI) would be a good fit for the evaluation pipeline. We extensively evaluate our model and experiment with alternate techniques of evaluation in an attempt to correlate well with human judgments. Our results are promising and we even outperform SOTA correlation values under some aspects of human judgment. An ablation for retrieval based analysis also shows the scope for an alternative metric design.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Pawan Goyal for his prompt assistance every time I needed his support. He helped me focus on the important topics and ensured that I carried out my work in a goal-oriented fashion. I also thank my mentor Bishal Santra for helping me understand the problem statement, patiently clearing all my doubts, and assisting me whenever I ran into issues.

Furthermore, I would like to thank the Department of Computer Science and Engineering, IIT Kharagpur for all the facilities and support, that contributed to this project. Finally, I also thank my family and friends for their constant encouragement.

Contents

Chapter 1

Introduction

With the advent of AI-based conversational agents, the need for well-defined evaluation metrics for dialog systems becomes even more necessary. [4]

Dialog systems can be categorized into 2 types: (1) Task-oriented systems, which focus on completing a specific task (2) Non-task-oriented systems (also known as chat-bots) which aim to converse with human beings in an open-domain setting.

Evaluating the former is an easier task as one can easily check if the task being pursued was completed. However, for open-domain generation which comes with an inherent one-to-many flavor to it, evaluation becomes increasingly difficult.[4]

Here, we study the performances of several evaluation metrics on the Topical Chat [5] and the Persona Chat[18] datasets. We also incorporate some changes in the SOTA evaluation models with the recently proposed DMI model [15] and report our results.

1.1 Motivation

In order to facilitate the development of dialog models, it is necessary to have well-designed metrics in the first place. Finding well-correlated metrics is an active research domain in the field of dialog systems. It was evident that we needed better metrics for evaluation than ROGUE [8] and BLEU [12] since these failed to correlate well with human judgments. This brought into consideration the new transformer-based models [16, 19] where one could extract the embeddings and calculate the similarity scores for the given context and response. The work of CTC[2] fascinated us since it drew parallels with information alignment for scoring, something which was a pretraining objective in the DMI model [15] involving concepts from information theory. As an ablation, we also drew inspiration from the recent retrieval-based models [13, 14] and explored how these systems correlate with human evaluations. Hence our main goal is **to align different pretrained representations through a secondary formulation viz. CTC and a retrieval based ranking model**.

Chapter 2

Background

2.1 Related work

Human evaluation is regarded as the oracle for dialog systems, but it is extremely expensive and time-consuming to obtain. The need for well-designed evaluation metrics is evident from the recent development in the world of virtual assistants.

The lack of meaningful automatic evaluation metrics has hindered the development of open-domain dialog research [10]. The roots of automatic evaluation metrics lie in the introduction of BLEU [12], and ROUGE [8] which measures the n-gram overlap between the generated text against human written ground truths. METEOR [1] was a development over BLEU owing to the fact that it considered synonyms while finding the overlap between two sentences.

However, the very nature of n-gram overlap doesn't consider the one-to-many nature of dialog models. Hence, it is necessary to go beyond n-gram matching to achieve a closer correlation with human evaluation. This led to the development of learnable evaluation metrics involving large-scale pretrained models [3, 9] based on transformers [17].

However most of these techniques performed well on the datasets they were evaluated upon, but they failed to generalize for unseen datasets.[10] This necessitated the use of a metric that was universal for natural language generation. We then came across a family of evaluation metrics proposed in [2] where the authors argue that NLG can be broadly classified into 3 tasks: 1. Compression (Summarization), 2. Transduction (Style transfer, eg: translation), and 3. Creation (Dialogs). All these methods have their roots in maximizing the information alignment between a given context and response. They incorporate various pretrained large-scale transformer models such as BERT [3] and RoBERTa [9] to approximate their formulation of information alignment.

This helped us to draw parallels from the concepts of information theory where dialog systems are trained to maximize the mutual information between a context and response [11, 15]. The recent success of another model which aims to maximize mutual information as a pretraining objective was the main motivation behind the project. We desired to check the performance of the DMI model as an evaluator hoping that two similar objectives would boost correlations with human evaluations. Later on, inspired by the recent retrieval-based systems used in popular NLG pipelines [6, 7, 13] we also experiment with a retrieval-based evaluation as a comparison with the CTC metrics.

2.2 Objectives

We may break down the objectives of this project into three parts:

- To survey the different available evaluation metrics.
- To see how well the DMI model correlates to human judgments when substituted in an automatic evaluation pipeline.
- To test an additional retrieval-based architecture for evaluation. (ablation)

2.3 Our approach

Since we were mostly concerned with the evaluation of dialog systems, we experimented with the *creation* aspect of the CTC evaluation. The authors argue that a single metric would naturally not suffice for analyzing all aspects of the annotated test sets. Hence, they improvised a family of metrics that had their roots in information alignment. 2.3.1. We first formalize a few basic definitions to present the final mathematical formulae devised for the metrics.

2.3.1 Information Alignment

The information alignment is measured at the token level. If a is a vector of tokens of length N and b is arbitrary data:

$$align(a \rightarrow b) =< \alpha_1, \alpha_2, ..., \alpha_N > \tag{2.1}$$

Here $\alpha_n \in [0, 1]$ is the probability that the n-th token in a is grounded by b.

There are two important aspects that must be noted here:

1. Alignment is unidirectional i.e. from a to b and not vice versa.
2. It is a vector of N values corresponding to each token and not a scalar quantity.

This basic formulation is then extended to design a family of intuitive metrics which are calculated using large-scale pretrained models. We now dive into the definitions of metrics for dialog evaluation by building upon alignment estimation. For each of the following definitions, we assume x is the input, y is the output generated and c is any other relevant context. [1]

[1] In general x and c may be of multimodal forms, however, we restrict ourselves to the textual analysis here.

2.3.2 Engagingness

This metric checks if the generated response is not bland/dull and provides interesting facts to the partner. Ideally, an engaging response y that satisfies the following:

1. Provides high volume of information
2. Acknowledges the input x and puts in relevant facts from the context c

Hence, the intuitive way to formulate engagingness is as follows

$$Engagingness(y, x, c) = sum(align(y \rightarrow [x, c])) \tag{2.2}$$

Here, the history x and dialog context c are concatenated and we measure the extent of overlap between them. The use of sum hints at the importance of the volume of information and successfully avoids the generation of bland responses such as *I don't know.*

2.3.3 Groundedness

This metric is essentially a subset of the engagingness measure and just measures the information overlap between the response and context

$$Groundedness(y, c) = sum(align(y \rightarrow c)) \tag{2.3}$$

One must note that we still stress upon the importance of the length of the response in an attempt to prevent generic answers. There are some models which estimate alignment and we take a deep dive into each of them in our formal description of the methodology.

Chapter 3

Exploring our dataset

3.1 The training set

3.1.1 Persona Chat

This dataset was constructed to train dialog agents with the goal of producing specific, consistent, and engaging responses. The authors sought to provide some profile information to the interlocutors a priori (aka persona) in an attempt to improve the groundedness of dialogs. The whole dataset consists of 162,064 utterances where dialogs are generated between randomly paired crowd workers. They were given a specific persona and were expected to converse with each other in such a way that it is consistent with their personas.[18]

3.1.2 Topical Chat

This is a knowledge-grounded human-human conversational dataset consisting of $\sim 11K$ conversations where the underlying knowledge spans 8 broad topics and conversation partners don't have explicitly defined roles. The interlocutors just

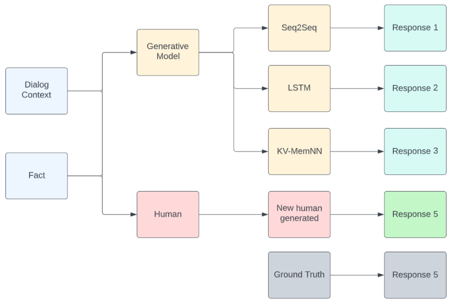

FIGURE 3.1: Persona chat pipeline

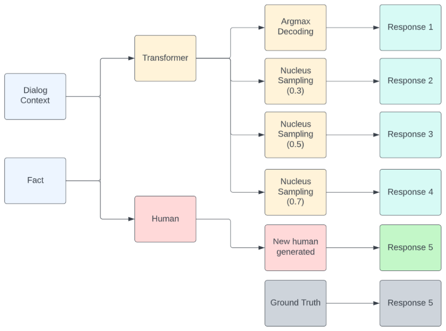

FIGURE 3.2: Topical chat pipeline

have a topical reading set and are expected to engage in coherent and grounded conversations as per their reading sets. [5]

3.2 The test set

3.2.1 Generation

For topical chat, the vanilla transformer[17] is trained to produce the response r given the context c. Different outputs are obtained for a given context owing to different decoding strategies: argmax sampling and nucleus sampling at three different rates: $p = 0.3, 0.5, 0.7$. This is augmented with a ground truth and a new human-generated response, thus amounting to 6 responses per input.(Figure: 3.2) Since there were 60 input utterances, 360 context-response pairs were generated.(Table: 3.2)

On the other hand for persona chat, three language models: a) Seq2Seq, b) LSTM and c) Key-Value Profile Memory Network were trained. These three responses were similarly augmented with the ground truth and a new human-generated response, thus leading to 5 responses for a given context (Figure: 3.1) For this dataset, 60 utterances led to 300 context response pairs in the test set. (Table: 3.2)

3.2.2 Evaluation

The dataset for evaluation was created in [10]. The authors avoid the use of public annotation platforms and instead borrow help from six experienced dialog researchers to obtain high-quality annotations. Each response was rated based on 6 (5 individual and 1 cumulative) aspects (Table: 3.1)

Metric	Range	Description
Understandable	(0-1)	Indicates if the response is valid given the context.
Natural	(1-3)	Tells if the response is something a human would naturally say.
Maintains Context	(1-3)	Indicates if the response is engaging.
Interesting	(1-3)	Indicates if a response is not dull.
Uses Knowledge	(0-1)	Reveals if a response is grounded with context.
Overall	(1-5)	Overall judgment score.

Table 3.1: Metrics used in [10] with descriptions

System	**Und (0-1)**	**Nat (1-3)**	**MCtx (1-3)**	**Int (1-3)**	**UK (0-1)**	**OO (1-5)**
	Topical-Chat					
Original Ground-Truth	0.95	2.72	1.92	2.64	0.72	4.25
Argmax Decoding	0.60	2.72	1.93	1.94	0.47	2.76
Nucleus Sampling (0.3)	0.51	2.08	2.01	1.82	0.42	2.40
Nucleus Sampling (0.5)	0.48	2.13	1.87	1.72	0.34	2.29
Nucleus Sampling (0.7)	0.52	2.02	2.92	1.80	0.37	2.39
New Human Generated	0.99	1.90	2.93	2.90	0.96	4.80
	Persona-Chat					
Original Ground-Truth	0.99	2.89	2.49	2.67	0.56	4.36
Language Model	0.97	2.82	2.70	2.24	0.08	2.98
LSTM Seq2Seq	0.92	2.63	2.18	2.29	0.47	3.47
KV-MemNN	0.93	2.02	2.97	2.56	0.17	3.25
New Human Generated	1.00	2.64	2.88	2.87	0.96	4.80

Table 3.2: Mean scores as per annotations by the dialog researchers.[10]

3.3 Dependency Analysis

The reason why one might defend a family of metrics instead of choosing a “one size fits all” [10] approach is because of the multi-faceted nature of dialog quality. On top of that, a smaller dataset allowed the authors to analyze stuff at a finer granularity since the views of the annotators were moderately correlated. A normalized regression map shows the importance of each weight in deciding the final overall score metric. (Figure: 3.3)

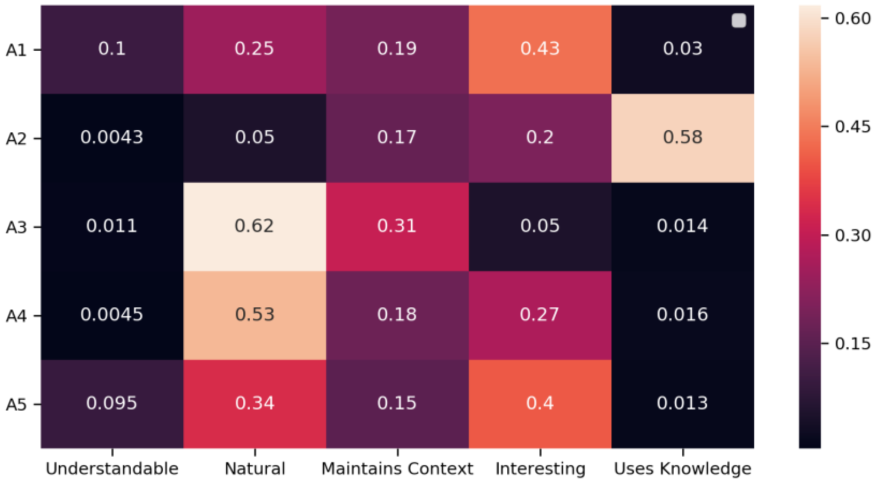

FIGURE 3.3: Dependency of metrics on overall score obtained by regression [10] . A lighter color indicated higher weightage. One can check the varying importance assigned to each metric for deciding the overall score.

Chapter 4

Model

4.1 Alignment Estimators

Broadly, we leverage 3 models as indicated in [2]. We now describe these models and show how $align(a \rightarrow b)$ is estimated. We keep in mind that alignment is a vector of N values, one for each token in a.

4.1.1 Embedding matching(E)

This is the simplest greedy estimation method and doesn't require fine-tuning. We begin by extracting embeddings for each token from pretrained models such as BERT[3] and RoBERTa[9]. This is followed by normalization. After that, we compute the alignment α_i of each token in a as the maximum dot product taken with its counterparts in b. [19]

4.1.2 Discriminative Model (D)

This necessitates the use of fine-tuning. The authors introduce weak supervision via the use of masking models [1] and paraphrasing techniques [2]. After that, the model is trained to predict a probability score between $[0, 1]$ for each token in a as per its alignment in b.

4.1.3 Aggregated Regression (R)

This model reuses the regression-based architecture proposed in [16]. The authors reuse the weights proposed in the same paper and train on the same weakly supervised data. The only difference here is that we consider the sum of scores for each token since the volume of information is important for dialog. This helps to prevent the generation of dull and generic responses such as *"I don't know"*

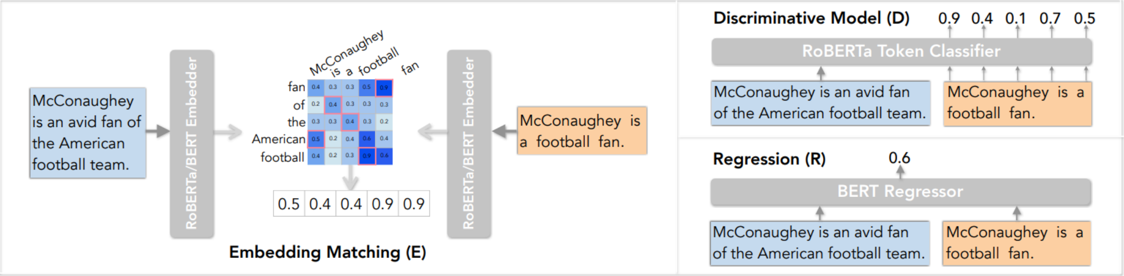
McConaughey is an avid fan of the American football team.
RoBERTa/BERT Embedder
McConaughey
is
a
football
fan
fan 0.4 0.3 0.3 0.5 0.9
of 0.2 0.4 0.3 0.3 0.3
the 0.3 0.3 0.4 0.3 0.2
American 0.5 0.2 0.4 0.6 0.4
football 0.4 0.2 0.3 0.9 0.6
0.5 0.4 0.4 0.9 0.9
RoBERTa/BERT Embedder
McConaughey is a football fan.
Embedding Matching (E)
Discriminative Model (D)
0.9 0.4 0.1 0.7 0.5
RoBERTa Token Classifier
McConaughey is an avid fan of the American football team.
McConaughey is a football fan.
Regression (R)
0.6
BERT Regressor
McConaughey is an avid fan of the American football team.
McConaughey is a football fan.

FIGURE 4.1: Illustration of the different approximators of Information alignment[15]

4.2 DMI model

This is a model based on the vanilla transformer [17]. The reason for the success of this model is the introduction of a structure-aware loss function [15] that takes into account the context while formulating the loss. It also overcomes the shortcomings

[1] https://github.com/nikitakit/self-attentive-parser
[2] https://huggingface.co/Vamsi/T5_Paraphrase_Paws

of the cross-entropy loss function as it assumes the existence of a single ground truth for a given context. Such a loss function isn't expected to do justice to a dialog system that has an inherent one-to-many flavor attached to it. Hence the loss function here is formulated in terms of the mutual information between the context and the response, something which is synonymous with the aspect of groundedness as proposed in [2].

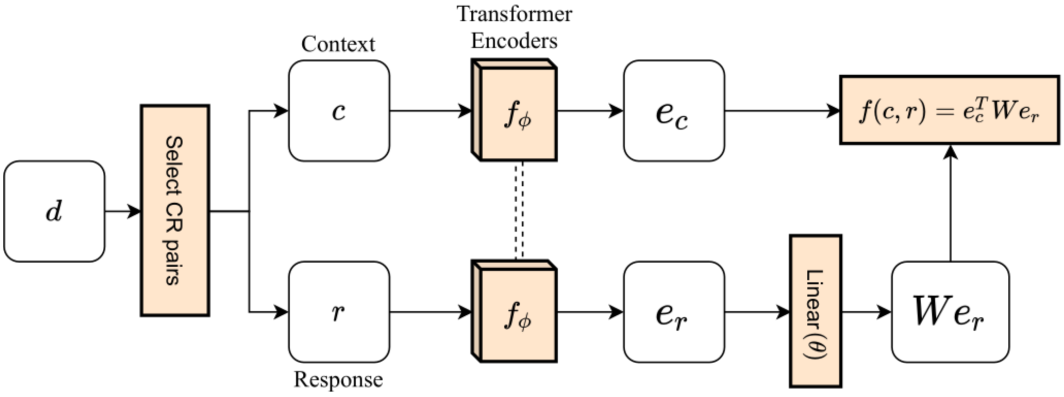

Figure 4.2: The transformer-based DMI model as proposed in [15]. f_ϕ denotes the transformer and d denotes samples from the training dataset

Chapter 5

Experiments

We recall that our main objective was to check if the similar goals of DMI and information alignment reinforce each other to outperform SOTA correlations. Hence the general procedure was to replace the pretrained models with the proposed DMI model.

5.1 Reproducing results using DMI model

We reuse all open-sourced pipelines for each implementation. We fine-tune our models as per the specifications and carry out correlation tests with each of the 6 metrics proposed in [10].

5.1.1 Implementation details

For the E-based models, we directly substitute our model into their pipeline and carry out the evaluation. But, for the D and R-based architecture, we perform fine-tuning after substituting our DMI model. The dataset for finetuning is essentially some weakly supervised data that is constructed from the persona-chat and topical-chat datasets and modifications are done as per the objective functions. We evaluate and compare our correlations with the values published in the paper.

5.2 Retrieval based metric

Additionally, we also utilize a retrieval-based metric for evaluation where we introduce a simple rank-based similarity metric which derives its inspiration from the pretraining objectives of several language models [6, 14].

5.2.1 Implementation details

The dataset is constructed from the ParlAI framework. Here we sample a pool of utterances that act as negative samples and develop a rank-based system. Along with the generated response, every pair of context and response is assigned a score and the metric is decided as its rank when sorted as per their scores in ascending order. Hence a higher rank implies a higher score. We find out correlations for 5 samples to get an expected value and suppress the effects of randomization.

Scoring is done with the help of three models: 1) Blenderbot 2) DialoGPT 3) DMI. For DMI, we reuse the score $f(c, r)$ which was proposed in the paper. While for the first two models, we utilize the perplexity (Equation 5.1) as the scoring function for ranking.

$$PP(W) = \sqrt[n]{\frac{1}{P(w_1, w_2, ..., w_n)}} \tag{5.1}$$

Here W denotes the sequence of response tokens generated and we are required to compute the joint probability of co-occurrence of each token.

5.3 Choosing the output layer

An interesting fact about the DMI model is that the weights in the last layer never got trained via backpropagation. This happened because the only output that we cared about was the proposed score $f(c, r)$ which was computed using the output corresponding to the $[CLS]$ token. Hence, as an ablation, we also check the scores by using the outputs from some of the intermediate layers.

5.3.1 Implementation details

Implementing this is straightforward. We simply change the extracted embedding for further calculations and check correlations for values extracted from each of the last 10 layers.

Chapter 6

Results

We note that the internal transformer-based language model is different from the models used for approximating information alignment. The DMI model[15] that achieved the best AUC used a `roberta-base` implementation. However, the other models used `roberta-large` or `bert-base` implementations. Hence, we would analyze each model separately by making two broad comparisons: 1) With those estimations that use RoBERTa internally and 2) With all other models. We now report our results for each configuration.

6.1 CTC evaluation after DMI substitution

6.1.1 Correlations for engagingness and groundedness

Embedding based

For the embedding matching models, we find that our E-DMI model performs better than the proposed E-Roberta-based models for both Topical and Persona Chat. One may say that it isn't a fair comparison since DMI undergoes pretraining where it learns to approximate the mutual information between a context and response. However, the implementation they provide uses a much larger model with almost three times more parameters. This might lead to the ability to store more information in the pretraining phase: something which might work against our model

Interestingly the BERT[3] model outperforms both the RoBERTa[9] based models by a significant margin. This is contrary to what we expected since RoBERTa was an improvement upon BERT owing to dynamic masking and larger pretraining data. A key difference was in the removal of Next Sentence Prediction as a pretraining objective for RoBERTa. This was mostly used to improve performance on downstream Natural Language Inference tasks such as finding out the relationship between two sentences, which is exactly what we are doing here.

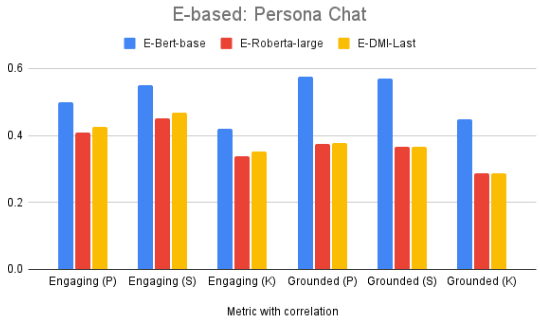

Figure 6.1: Correlations for E-based models considering engagingness and groundedness aspects of persona chat.

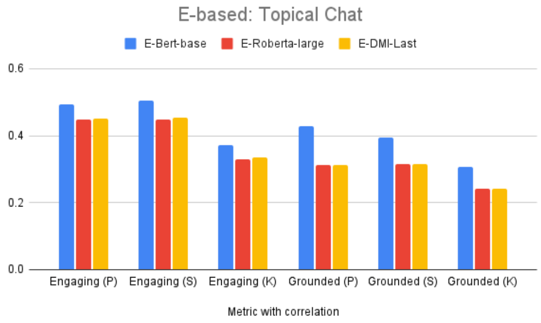

Figure 6.2: Correlations for E-based models considering engagingness and groundedness aspects of persona chat.

Discriminative based

In this scenario, our DMI model outperforms SOTA when we consider the engagingness aspect. However, for groundedness, the proposed RoBERTa model outperforms our DMI-Rob model. This is the case for both topical and persona chat. These values may be found in tables 6.1 and 6.2. One should note that the discriminative model is the only place where we are able to beat SOTA. This is because the pretraining objective of DMI aligns with the proposed method of calculating information alignment.

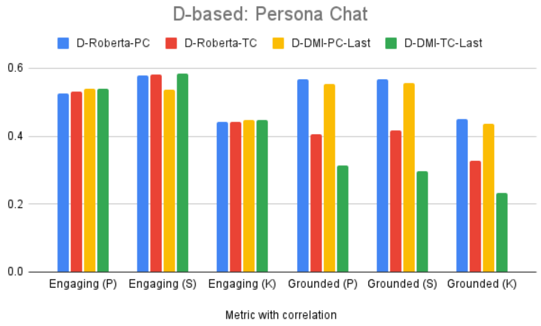

Figure 6.3: Correlations for D-based models considering engagingness and groundedness aspects of persona chat.

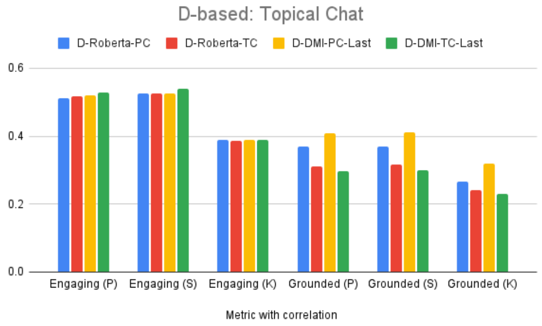

Figure 6.4: Correlations for D-based models considering engagingness and groundedness aspects of persona chat.

Regression based

For this, the implementation given in [16] uses a `bert-base-midtrained` model as was proposed in [16]. Originally it was found to outperform other metrics considering the groundedness aspect. In our substitution, for most of the cases, the BERT-based model outperforms our model. We may attribute this to the difference in the pretraining phases of both models and provide a similar reason as we did for the performance of our Embedding-based models. However, for some cases in persona chat, our DMI model does a better job.

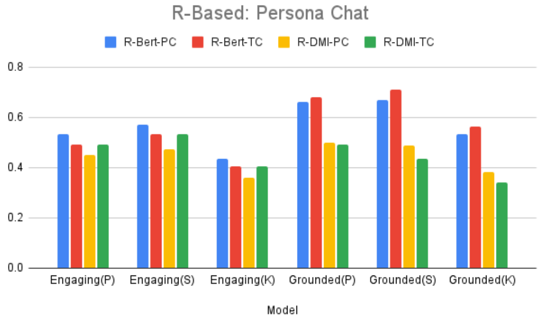

Figure 6.5: Correlations for R-based models considering engagingness and groundedness aspects of persona chat.

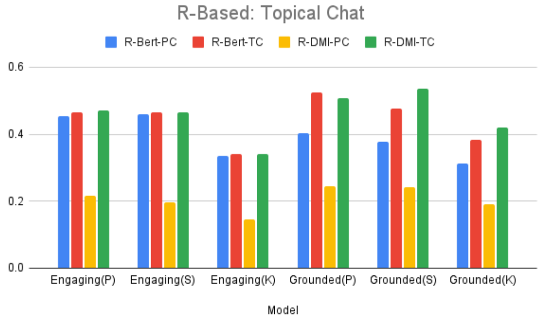

Figure 6.6: Correlations for R-based models considering engagingness and groundedness aspects of persona chat.

	Engaging			Grounded		
Model	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
E-Bert-base	0.5003	0.549	0.4193	0.5761	0.5683	0.4492
E-Roberta-large	0.4081	0.4502	0.3375	0.3758	0.3652	0.2862
E-DMI-Last	0.4241	0.4691	0.3517	0.3777	0.367	0.2863
D-Roberta-PC	0.5265	0.5793	0.4412	0.5683	0.5674	0.4505
D-Roberta-TC	0.5317	0.5818	0.4409	0.4056	0.4172	0.327
D-DMI-PC-Last	0.5384	0.538	0.4463	0.5541	0.5557	0.4367
D-DMI-TC-Last	**0.5392**	**0.585**	**0.448**	0.3137	0.2961	0.232
R-Bert-PC	0.532	0.5692	0.4346	0.6597	0.6689	0.5338
R-Bert-TC	0.4933	0.5333	0.4043	**0.6819**	**0.7113**	**0.5636**
R-DMI-PC	0.4519	0.4743	0.3596	0.5002	0.4871	0.3841
R-DMI-TC	0.4927	0.5336	0.4033	0.4903	0.4367	0.3423

Table 6.1: Correlations on the engagingness and groundedness aspects of persona chat

	Engaging			Grounded		
Model	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
E-Bert-base	0.4937	0.5047	0.371	0.4293	0.3949	0.3075
E-Roberta-large	0.4471	0.4479	0.3288	0.3122	0.3141	0.2421
E-DMI-Last	0.452	0.4543	0.3345	0.3131	0.3149	0.2422
D-Roberta-PC	0.5124	0.5245	0.3878	0.3697	0.3691	0.2656
D-Roberta-TC	0.5163	0.5253	0.3873	0.3099	0.3159	0.2421
D-DMI-PC-Last	0.5201	0.5256	**0.3898**	0.4075	0.4102	0.3192
D-DMI-TC-Last	**0.5282**	**0.5385**	**0.3896**	0.2967	0.2988	0.2289
R-Bert-PC	0.4542	0.4588	0.3357	0.4026	0.3788	0.3137
R-Bert-TC	0.4653	0.4643	0.3395	**0.5235**	**0.4768**	**0.3838**
R-DMI-PC	0.2151	0.1975	0.1449	0.2448	0.2423	0.1903
R-DMI-TC	0.4699	0.4654	0.3411	0.5067	**0.5349**	**0.4201**

Table 6.2: Correlations on the engagingness and groundedness aspects of topical chat

6.1.2 Combining all remaining aspects

For this, we reuse the models that perform best in the engagingness and groundedness evaluations. We find that the discriminative-based models outperform all others. Secondly, our model outperforms the proposed DMI model in 2 out of 8 scenarios. We also explore some additional retrieval-based methods where we achieve higher correlations with human judgments. (See table 6.3 and table 6.4)

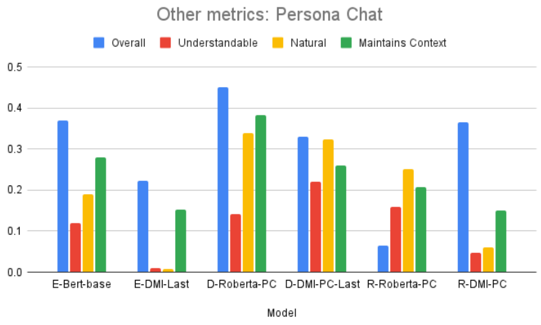

FIGURE 6.7: Correlation with the remaining metrics for persona chat

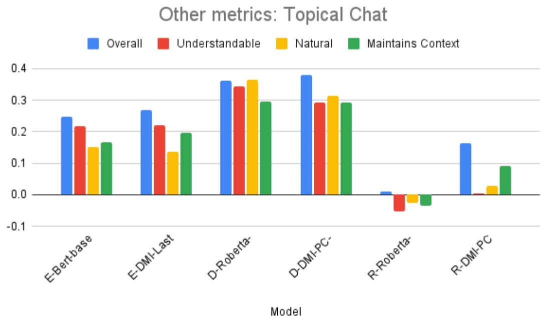

FIGURE 6.8: Correlation with the remaining metrics for topical chat

	Persona Chat			
Model	Overall	Understandable	Natural	Maintains Context
E-Bert-base	0.369	0.1185	0.1891	0.2786
E-DMI-Last	0.2235	0.0102	0.0076	0.1532
D-Roberta-PC	**0.45**	0.1421	**0.3384**	**0.3837**
D-DMI-PC-Last	0.3295	**0.2193**	0.3232	0.259
R-Bert-PC	0.0639	0.1595	0.2518	0.2076
R-DMI-PC	0.3642	0.0482	0.0598	0.1509

TABLE 6.3: Checking rest of the metrics for persona chat

	Topical Chat			
Model	Overall	Understandable	Natural	Maintains Context
E-Bert-base	0.2483	0.2168	0.1528	0.1669
E-DMI-Last	0.268	0.2209	0.1356	0.1969
D-Roberta-PC	0.362	**0.3433**	**0.3653**	**0.2969**
D-DMI-PC-Last	**0.3791**	0.2918	0.3136	0.2923
R-Bert-PC	0.0105	-0.0524	-0.0248	-0.0338
R-DMI-PC	0.1626	0.0036	0.0283	0.0922

TABLE 6.4: Checking rest of the metrics for topical chat

6.2 Results for retrieval-based scoring model

Table 6.5 shows us that the DMI model performs better than both Blenderbot and DialoGPT for most cases in persona chat. The best performance is seen for *Maintains Context* aspect for both datasets. However, it fails to correlate well with the engaging and uses knowledge aspects in topical chat (Table: 6.6). DialoGPT outperforms the other two models and correlates better than its counterparts in these situations.

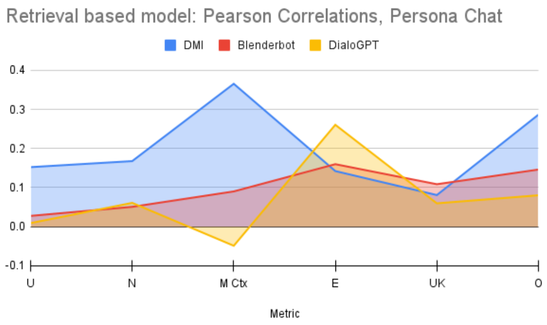

FIGURE 6.9: Pearson correlations for retrieval-based model on persona chat

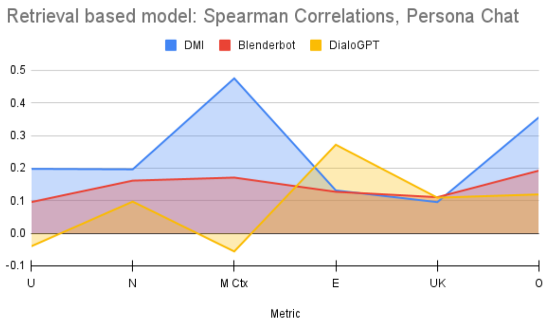

FIGURE 6.10: Spearman correlations for retrieval-based model on persona chat

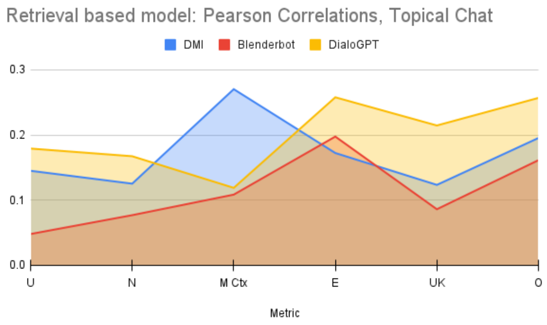

FIGURE 6.11: Pearson correlations for retrieval-based model on topical chat

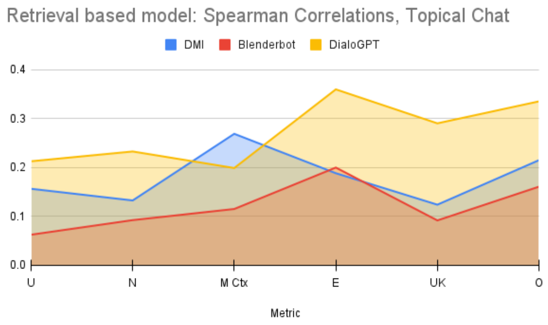

FIGURE 6.12: Spearman correlations for retrieval-based model on topical chat

	Pearson			Spearman		
Metric	DMI	Blenderbot	DialoGPT	DMI	Blenderbot	DialoGPT
Understandable	**0.1523354**	0.0277692	0.009365	**0.197575**	0.095623	-0.040086
Natural	**0.1677818**	0.0508486	0.0610028	**0.196062**	0.161438	0.09714
Maintains Context	**0.3654034**	0.0902726	-0.0486168	**0.475169**	0.170732	-0.055845
Engaging	0.1422836	0.15989	**0.2605532**	0.131414	0.127119	**0.271572**
Uses Knowledge	0.0809784	**0.1086642**	0.0597636	0.09562	**0.110561**	0.109144
Overall	**0.286453**	0.146188	0.0804444	**0.355731**	0.192084	0.119061

Table 6.5: Correlation values for the retrieval-based models on persona chat dataset

	Pearson			Spearman		
Metric	DMI	Blenderbot	DialoGPT	DMI	Blenderbot	DialoGPT
Understandable	0.1450364	0.048106	**0.1791504**	0.156578	0.062709	**0.212877**
Natural	0.1252882	0.076989	**0.1673346**	0.132721	0.092496	**0.23289**
Maintains Context	**0.2702156**	0.1084988	0.1188672	**0.268986**	0.115255	0.19903
Engaging	0.1723016	0.1975004	**0.2577412**	0.188429	0.200002	**0.359883**
Uses Knowledge	0.1234016	0.0860758	**0.2145094**	0.124054	0.09188	**0.290415**
Overall	0.1953214	0.1613102	**0.2567764**	0.215071	0.160862	**0.335334**

Table 6.6: Correlation values for the retrieval-based models on topical chat dataset

6.2.1 Layer wise check

A glance into the results (Table 6.7) for the layer-wise analysis confirms our suspect and shows us that values from the last layer aren't the best representations. We only consider embedding-based models here. The best correlations are achieved using representations from the fourth and fifth layers. Although the differences are lesser, this motivates us to use representations from other layers during fine-tuning for the D and R-based models.

6.3 Discussion

To summarize our interesting observations, we have found our model DMI to outperform SOTA correlation values for the discriminative models under the engagingness

Layer	PC-Engaging	PC-Grounded	TC-Engaging	TC-Grounded
E-DMI-Last	0.4241	0.3777	0.452	0.3131
E-DMI-2L	0.424	0.3801	0.4521	0.3132
E-DMI-3L	0.4256	0.3853	0.4533	0.3142
E-DMI-4L	0.4266	**0.391**	0.4522	**0.3169**
E-DMI-5L	**0.4267**	0.3856	**0.4525**	0.3133
E-DMI-6L	0.4259	0.3851	0.4508	0.31
E-DMI-7L	0.4254	0.3869	0.4513	0.3118
E-DMI-8L	0.4228	0.3907	0.4494	0.3146
E-DMI-9L	0.4217	0.392	0.4491	0.3143
E-DMI-10L	0.4179	0.3904	0.4448	0.3101

Table 6.7: Pearson correlations embedding based models after varying layers. The difference is minute but the 4th and 5th layers result in best correlations

aspect. Also, witnessing BERT perform better than our RoBERTa-based DMI certainly motivates us to experiment using BERT as the base language model for DMI. Removing the Next Sentence Prediction task from the pretraining phase possibly reduces the ability to perform Natural Language Inference tasks. Furthermore, our retrieval-based system shows much promise and achieves comparable results with the zero-shot approaches. Finally, our layer-wise analysis also confirms that harnessing representations from the last layer may not always be the best idea for the RoBERTa-based DMI model.

6.4 Conclusion and Future Work

As part of this project, the DMI model was extensively tested as an evaluator and was shown to even outperform SOTA in some scenarios. This establishes the fact that DMI does well as an evaluator. However, to fully leverage the strengths of the DMI pretraining, we are motivated to use the BERT model in the DMI pipeline owing to the incorporation of NSP which allows us to reason about relationships between sentences. Our retrieval based models also show an Our layer-wise analysis also further lays the foundation for fine-tuning the DMI-based model on the representations taken from layers other than the last one to achieve better results.

Bibliography

[1] Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

[2] Deng, M., Tan, B., Liu, Z., Xing, E., and Hu, Z. (2021). Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[3] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

[4] Finch, S. E. and Choi, J. D. (2020). Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

[5] Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

[6] Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020). Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

[7] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

[8] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[9] 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.

[10] Mehri, S. and Eskenazi, M. (2020). USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

[11] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding.

[12] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[14] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., and Weston, J. (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.

[15] Santra, B., Roychowdhury, S., Mandal, A., Gurram, V., Naik, A., Gupta, M., and Goyal, P. (2022). Representation learning for conversational data using discourse mutual information maximization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1718–1734, Seattle, United States. Association for Computational Linguistics.

[16] Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[18] 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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

[19] Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.