Methodologies for Aligning Pretrained Dialog Representations to Aspects of Human Judgments

Abhinandan De (19CS10069)

Supervisor: Prof. Pawan Goyal

Contents

What is this about?

Dialog systems: Massive growth over the last few decades

Need good evaluation systems to develop

Problem: Human evaluation is too costly

Solution: Automatic evaluation

Apprehension: Do they correlate well with human judgements?

Resolve: Let's find out with some pretrained dialog representations

What has already been done?

BLEU[6], **ROUGE** [4], **METEOR**[12]: mostly syntactic

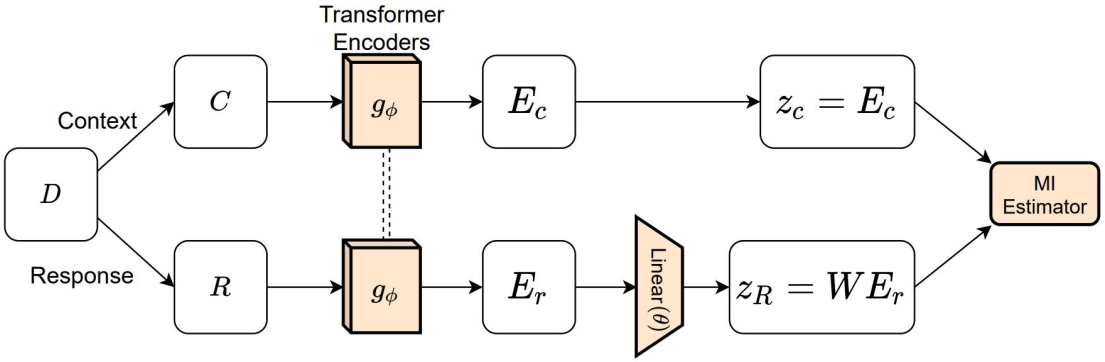
BLEURT[8], **BERTScore**[11]: Transformer based models[2, 5, 9] to calculate similarity

CTC[1] - Compression Transduction and Creation: family of unifying metrics

DMI[7] - Discourse Mutual Information: approximate MI between context and response

But why DMI + CTC?

Pre-training objective: approximate Mutual Information



CTC[1]: unifying perspective with its roots in information alignment

Question: Do they come “in phase” and enhance correlation?

Technicalities

Information Alignment: How well a aligns to b

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

Engagingness: Alignment of output with both input and context.

$$Engagingness(y, x, c) = sum(align(y \rightarrow [x, c]))$$

Groundedness: Alignment of output with context.

$$Groundedness(y, x, c) = sum(align(y \rightarrow, c))$$

Technicalities (contd)

E-based → D-based → R-based

Uses BERT-base or RoBERTa-large

No fine-tuning

$$\alpha_i = \max_{j \in [1,m]} sim(a_i, b_j)$$

Uses RoBERTa-large

Finetunes on weakly supervised data

$$\alpha_i = p(a_i|b)$$

Uses BERT-base

Finetunes on weakly supervised data

$$\alpha = \sum_i p(a_i|b)$$

Masking → Paraphrasing → Predicting

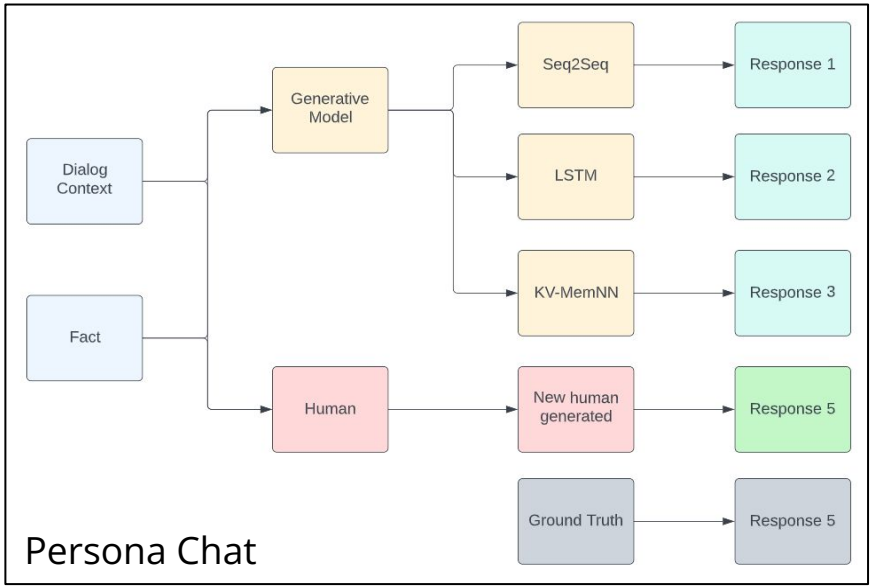
Weak supervision

Dataset

Persona chat[10, 13]: Chat in a way that's consistent with their personas.

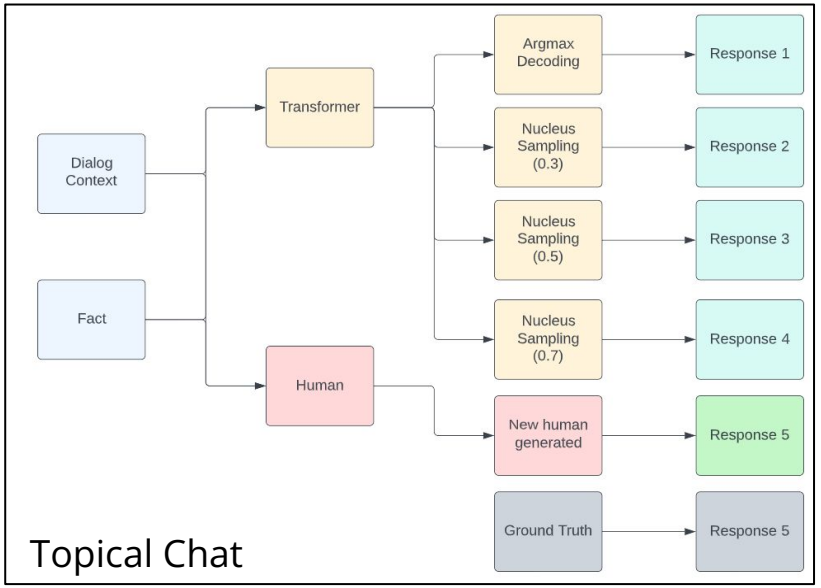
Topical Chat[3, 13]: Chat in a way that's grounded with the topic given.

$$60 * 5 = 300(c, r)pairs$$



Persona Chat

$$60 * 6 = 360(c, r)pairs$$



Topical Chat

Experiments

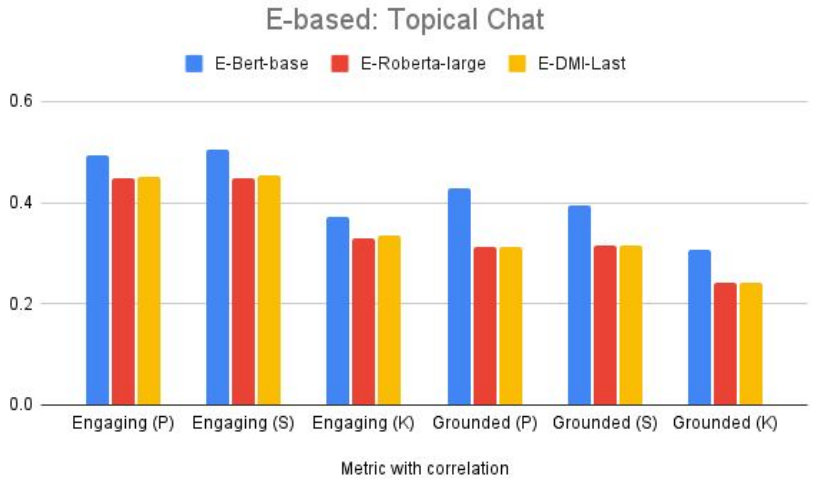
Expt 1: Reuse the pipelines, substitute with DMI

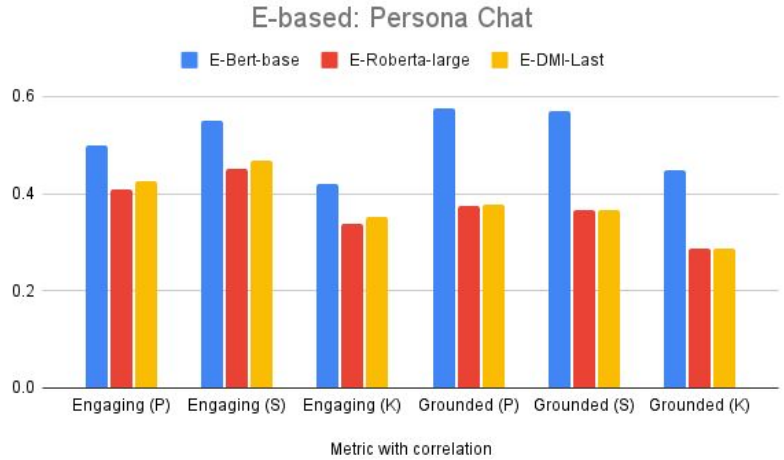
Expt 2: Explore a retrieval-based model

Expt 3: Explore representations from different layers (ablation)

Results: Embedding(E) based models

General trend: BERT > RoBERTa-base (DMI) > RoBERTa-large

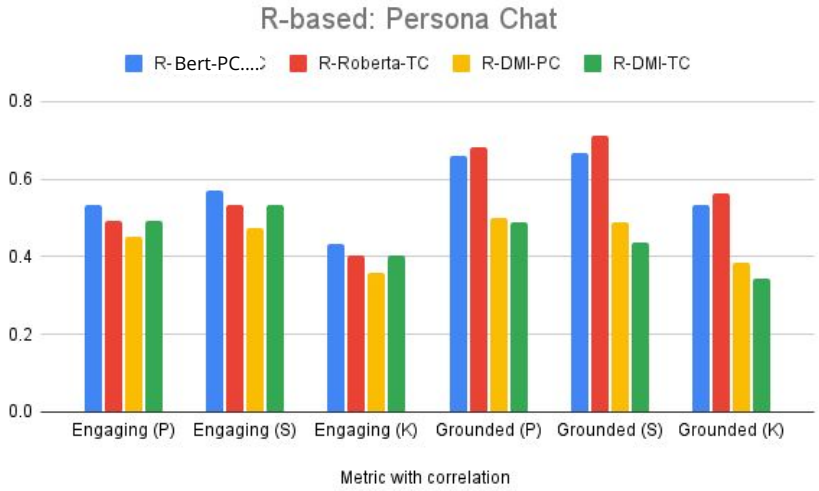


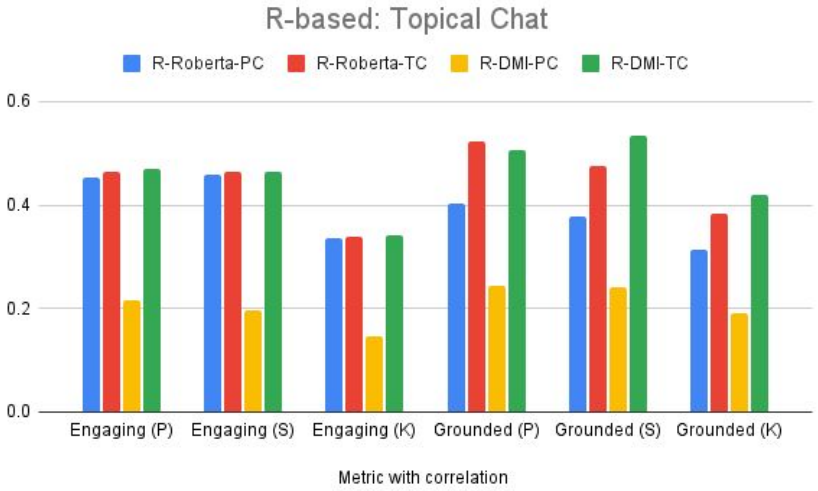


Possible reason for BERT > RoBERTa : Removal of **NSP** which assists NLI tasks

Results: Regression(R) based models

General trend: BERT > RoBERTa-base (DMI)



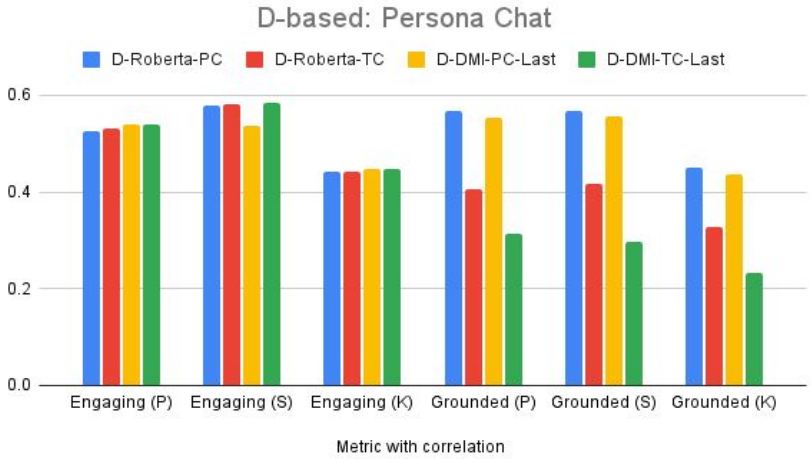


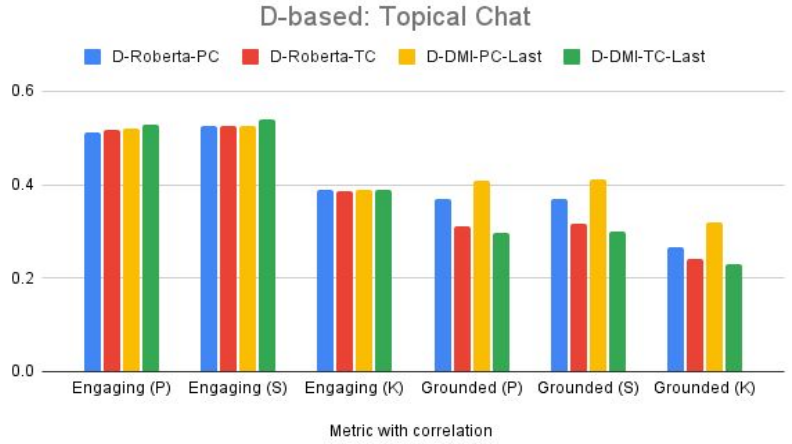
Possible reason for trend: Same as R-based analysis

Results: Discriminative(D) alignment models

Engagingness: RoBERTa-base (DMI) > RoBERTa-large

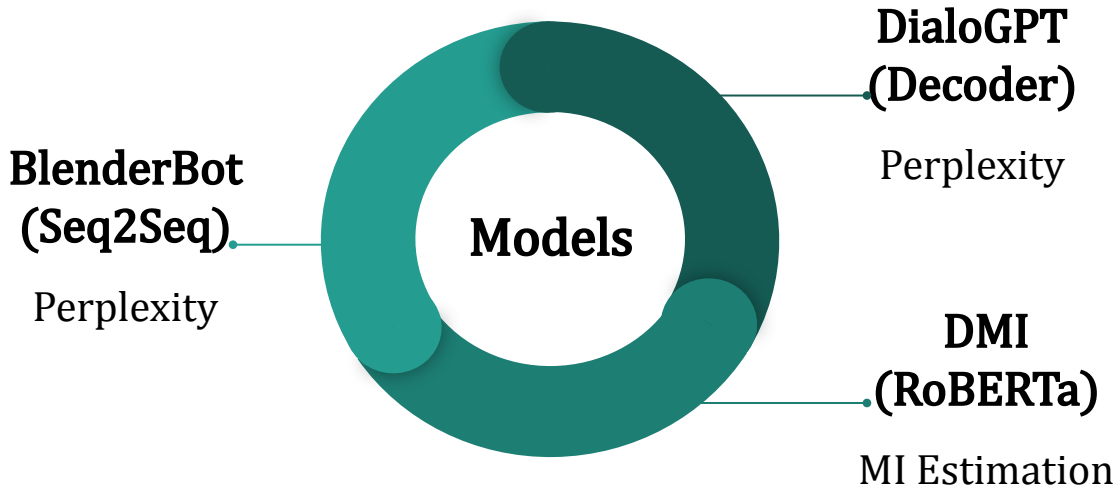
Groundedness: RoBERTa-large > RoBERTa-base (DMI)



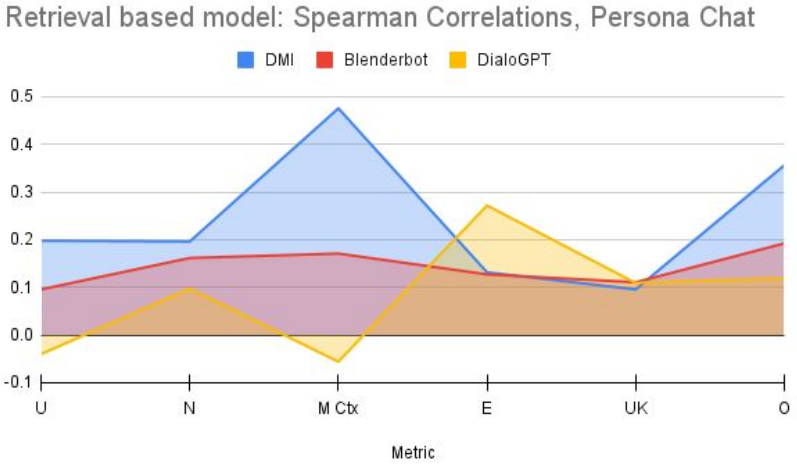


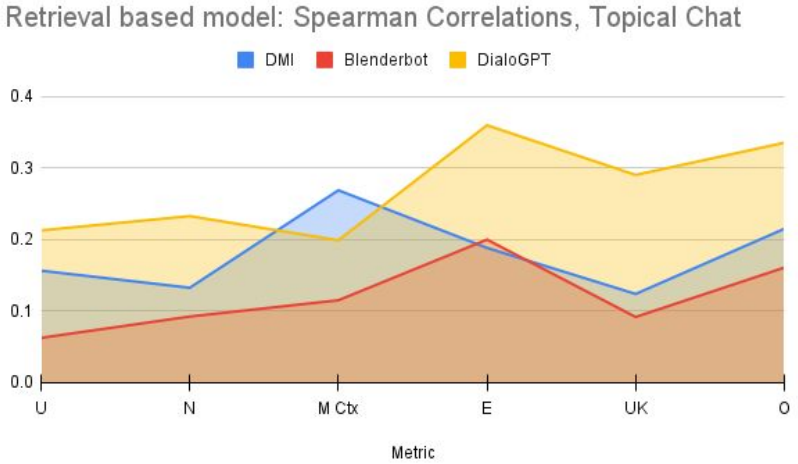
Possible explanation: The D-model formulation aligns with DMI pretraining

Experiment II: Retrieval based comparison



Results





Persona: DMI does better on most cases apart from uses knowledge
Topical: DialoGPT beats other models in all cases apart from maintains context
Overall: DMI always scores well in maintains context aspect

Layer-wise analysis

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

Note: Here L means Last; E-DMI-3L refers to representations harnessed from the 3rd last layer

Inference: 4th and 5th layers lead to best correlation!

Conclusion and Future Work

We beat their performance for the D-based model (engagingness).

We find promising results from the retrieval based model.

We get motivated to explore the application of BERT in DMI

We then use values from intermediate layers for fine-tuning.

References

[1] Deng, M., Tan, B., Liu, Z., Xing, E. P., and Hu, Z. (2021). Compression, transduction, and creation: A unified framework for evaluating natural language generation

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding

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

[4] 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

[5] 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

[6] 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.

References

[7] Santra, B., Roychowdhury, S., Mandal, A., Gurram, V., Naik, A., Gupta, M., and Goyal, P. (2021). Representation learning for conversational data using discourse mutual information maximization.

[8] Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need

[10] 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?

[11] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert.

[12] 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.

[13] Mehri, S. and Eskenazi, M. (2020). USR: An unsupervised and reference free evaluation metric for dialog generation.

THANK YOU!

Questions?