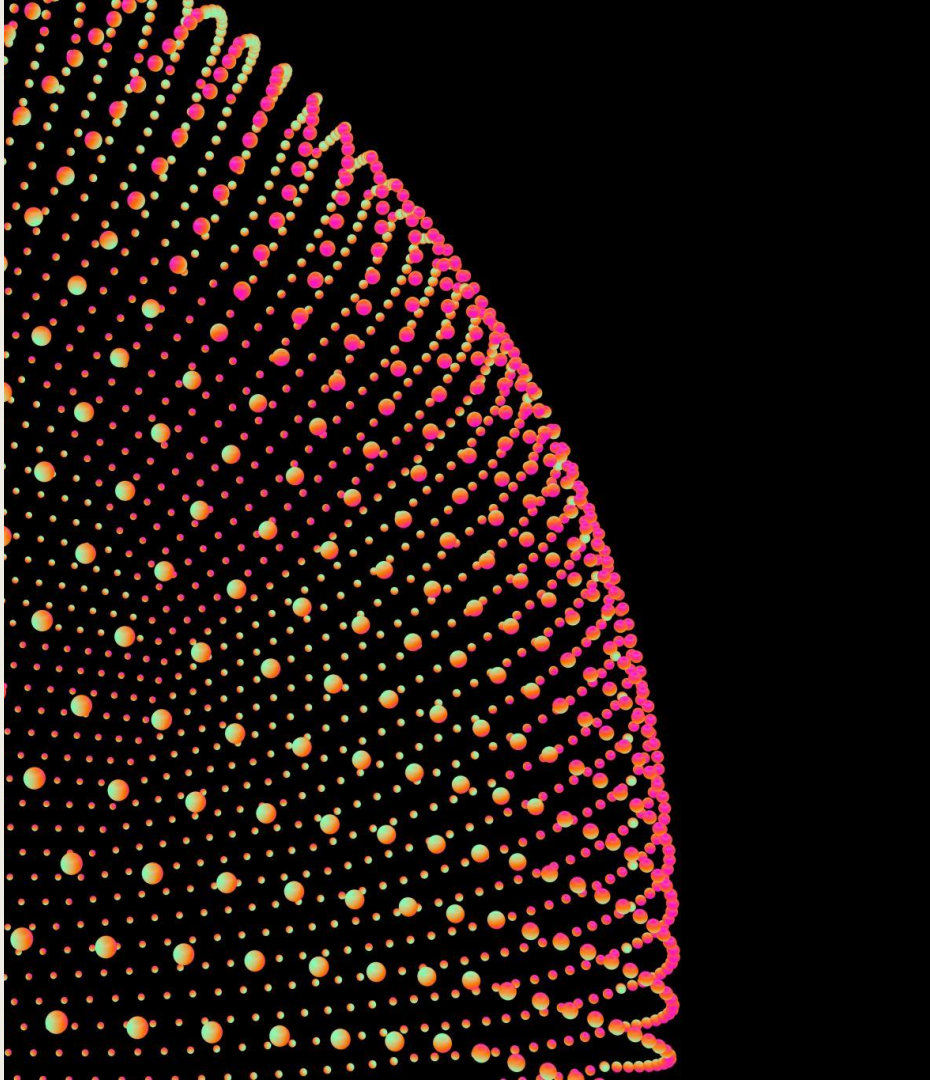
ASSESSING PERFORMANCE OF CONVERSATIONAL AGENTS

~ Group 5



Data Set and Task: Conversational Agent with MultiWOZ 2.2

- Problem Description
  - Given datasets and pretrained models, our project aims to explore the differences in performance between different models, different decoding techniques, different evaluation metrics, and by fine-tuning the pre-trained models.
  - In addition to the MultiWOZ2.2 dataset provided, we supplemented our fine-tuning dialogue with The Schema Guided Dialogue DataSet
- To carry out this problem, we used two different datasets:
  - MultiWOZ Data
  - The Schema Guided Dialogue DataSet

Data Sets

- Multi-Domain Wizard-of-Oz dataset (MultiWOZ)
  - a fully-labeled collection of human-human written conversations
  - size of 10k dialogues
  - at least one order of magnitude larger than all previous annotated task-oriented corpora
- Schema Guided Dialogue DataSet (SGD)
  - Multi-domain, task-oriented conversations between a human and a virtual assistant
  - size of 20k dialogues
  - conversations involve interactions with services and APIs spanning 20 domains

Implementation

- All training and evaluation performed on a single 8GB Nvidia GTX 1070

Changes made to codebase:

  - *Fine-tuning script modification*
  - *File directory structure*
  - *Automated evaluation loop*
  - *New decoding method (Top-K) and new metrics (BERTScore and ROUGE)*

Initial Results

Performance By Model, Decoding Technique, and Evaluation Metric

	Greedy Decoding			Top-k Decoding			Top-p Decoding			Beam Search		
Model	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**
GPT-1	0	0.722	0.083	0	0.769	0.085	0	0.796	0.068	0.039	0.716	0.093
GPT-2	0.098	0.758	0.105	0.113	0.778	0.093	0	0.799	0.08	0.1	0.775	0.112
GPT-1 Fine Tuned	0.075	0.782	0.167	0.087	0.782	0.156	0.063	0.782	0.157	0.07	0.777	0.154
GPT-2 Fine Tuned	0.264	0.85	0.263	0.189	0.848	0.212	0.169	0.848	0.206	0.28	0.831	0.253

Additional Results

- As mentioned, in addition to the MultiWOZ dataset, we supplemented the fine-tuning data, re-tuned the models, and once again evaluated. See the results below

	Greedy Decoding			Top-k Decoding			Top-p Decoding			Beam Search		
Model	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**
GPT-1 Fine Tuned-Supp	0.077	0.778	0.161	0.069	0.781	0.152	0.069	0.78	0.149	0.078	0.778	0.16
GPT-2 Fine Tuned-Supp	0.239	0.855	0.274	0.152	0.848	0.223	0.248	0.847	0.214	0.262	0.83	0.253

Our Approaches: Different Models
Background and Hypothesis

Our first approach was to compare pre-trained GPT vs. GPT2 models' zero-shot performance. Before analyzing the results, we want to highlight the key differences between the two models.

- GPT1
  - Semi-supervised learning.
  - Dataset: Bookscorpus
  - 12 layer decoder only + masked self attention (110M parameters)
- GPT2 (smallest version)
  - semi-supervised learning + task conditioning.
  - GPT2 has larger dataset -> WebText (40 M data)
  - 12 layer decoder only + masked self attention (117M parameters)

Since GPT-2 is pre-trained on a larger dataset and has slightly more tunable parameters, we hypothesize that the performance of the GPT2 model will outperform the performance of the GPT1 model both in zero-shot and fine-tuned evaluation.

Our Approaches: Different Models
Results and Interpretation: Zero-Shot

	Greedy Decoding			Top-k Decoding			Top-p Decoding			Beam Search		
Model	BLEU	BERTScore	ROUGE	BLEU	BERTScore	ROUGE	BLEU	BERTScore	ROUGE	BLEU	BERTScore	ROUGE
GPT-1	0	0.722	0.083	0	0.769	0.085	0	0.796	0.068	0.039	0.716	0.093
GPT-2	0.098	0.758	0.105	0.113	0.778	0.093	0	0.799	0.08	0.1	0.775	0.112

- GPT1
  - As we can observe above, the zero-shot performance of the GPT model was quite poor, with multiple BLEU scores rounding down to a value of 0. Given the difficult task of generating human-level conversation, especially without fine-tuning, these results do not come as a surprise
- GPT2
  - For GPT2, we see similar poor performance across the board, but with slightly better results than the GPT1 model.
- Compare and discuss results
  - As we predicted, the performance of GPT1 is worse than that of GPT2 for every combination of decoding method and evaluation metric used. As we explained previously, the larger amount (and potential quality improvement) of training data as well as the increased number of trainable parameters allows GPT2 to generate better human-like conversations.

Our Approaches: Different Models
Results and Interpretation: Fine-Tuned

	Greedy Decoding			Top-k Decoding			Top-p Decoding			Beam Search		
Model	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**	**BLEU**	**BERTScore**	**ROUGE**
GPT-1 Fine Tuned	0.075	0.782	0.167	0.087	0.782	0.156	0.063	0.782	0.157	0.07	0.777	0.154
GPT-2 Fine Tuned	0.264	0.85	0.263	0.189	0.848	0.212	0.169	0.848	0.206	0.28	0.831	0.253
GPT-1 Fine Tuned-Supp	0.077	0.778	0.161	0.069	0.781	0.152	0.069	0.78	0.149	0.078	0.778	0.16
GPT-2 Fine Tuned-Supp	0.239	0.855	0.274	0.152	0.848	0.223	0.248	0.847	0.214	0.262	0.83	0.253

- GPT1
  - Fine tuning the GPT1 model lead to modest performance increase (in terms of % improvement from the zero-shot performance) but the model still had fairly poor performance.
- GPT2
  - For GPT2, however, we see a significant increase in the performance across all metrics, with improvements ranging from about 10% (BERTScore) to over 250% (BLEU) depending on the decoding technique and metric used
- Compare and discuss results
  - Once again, as we predicted, the performance of GPT1 is worse than that of GPT2 for every combination of decoding method and evaluation metric used. The interesting thing to note here, is that while fine tuning improved both models, it improved the performance of GPT2 much more than it did GPT1. We believe that this is due to the larger number of tunable parameters in the GPT2 model, allowing it to better fine-tune the model in order to learn the most important features of conversations (at least based on the training data provided).

Our Approaches: Varying Decoding Techniques
Background and Hypothesis

- Greedy Decoding
  - Selection of word with highest probability
- Top-K
  - Redistribution of probability mass over K most likely next words.
  - For this experiment : k = 5.
- Top-P
  - Shortlisting smallest number of top tokens with sum of probabilities exceeding p.
  - For this experiment : p = 0.90.
- Beam Search
  - Expanding the highest probability beam by keeping the number of beams constant in each timestep.
  - For this experiment : # of beams = 5.
- We hypothesize that beam search will yield best results and greedy decoding will have worst results.

Our Approaches: Varying Decoding Techniques
Results and Interpretation

Ranking of decoding technique performance for each model and evaluation metric

	Greedy Decoding				Top-k Decoding				Top-p Decoding				Beam Search			
Model	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG
GPT-1	2	3	3	2.66	2	2	2	2	2	1	4	2.33	1	4	1	2
GPT-2	3	4	2	3	1	2	3	2	4	1	4	3	2	3	1	2
GPT-1 Fine Tuned	2	1	1	1.33	1	1	3	1.66	4	1	2	2.33	3	4	4	3.33
GPT-2 Fine Tuned	2	1	1	1.33	3	2	3	2.66	4	2	4	3.33	1	4	2	2.33
GPT-1 Fine Tuned-Supp	2	3	1	2	3	1	3	2.33	3	2	4	3	1	3	2	2
GPT-2 Fine Tuned-Supp	3	1	1	1.66	4	2	3	3	2	3	4	3	1	4	2	2.33
				2				2.28				2.8				2.33

- The table above details the rank of each model's performance by decoding method used
- Originally we predicted that beam search would provide the best results and greedy the worst, but we found that greedy decoding actually performed best, on average, with top-k next, followed by beam search, and ultimately top-p.
- One potential explanation for this, is that given a well trained (and fine-tuned model), a very good model should assign the highest probability to the "best" token consistently. This notion supports the fact that greedy decoding is the best, since it samples the fewest (0) tokens that are not the most probable.
- Following suit, top-k limits the number of tokens sampled to a prespecified value (in our case 5) and we predict that lowering that value would further improve performance while increasing it would cause performance to suffer.

Our Approaches: Varying Evaluation Metric
Background and Hypothesis

- BLEU Score
  - Evaluation of quality of machine translated text with natural language text.
- BERTScore
  - Leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference texts by cosine similarity.
- ROUGE Score
  - Evaluation of automatic summarization and machine translation software
- Hypothesis about performance using each
  - We hypothesize that evaluation metrics should not matter the best performance.

Our Approaches: Varying Evaluation Metric
Results and Interpretation

	Greedy Decoding				Top-k Decoding				Top-p Decoding				Beam Search				
Model	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	AVG
GPT-1	6	6	6	6	6	6	6	6	6	5	6	5.66	6	6	6	6	5.9
GPT-2	5	5	5	5	3	5	5	4.33	6	3	5	4.66	3	5	5	4.33	4.6
GPT-1 Fine Tuned	4	3	4	3.66	4	3	3	3.66	4	6	3	4.33	5	4	4	4.66	4.1
GPT-2 Fine Tuned	1	2	2	1.66	1	1	2	1.33	2	1	2	1.66	1	1	1	1	1.4
GPT-1 Fine Tuned-Supp	3	4	4	3.66	5	4	4	4.66	3	4	4	3.66	4	3	3	3.66	3.9
GPT-2 Fine Tuned-Supp	2	1	1	1.33	2	1	1	1.66	1	2	1	1.33	2	2	1	1.66	1.5

- The table above details the average rank of each model and decoding combination based on which evaluation metric was used
- The results were largely consistent across each metric, with some minor variability across the fine-tuned and supplemental fine-tuned models.
- This fits our hypothesis that using different evaluation metrics should not change which models and decoding techniques yielded the best results
- In every instance, GPT2 fine-tuned on either the original WOZ or supplemented dataset performed best, regardless of metric used
- In all but once instance, zero-shot GPT1 performed worst, regardless of metric used
- Top-p decoding was the only area in which we saw relative model rankings vary by more than 1 position based on which evaluation criteria was used.

Our Approaches: Supplementing Fine-Tuning Data
Background and Hypothesis

- Original Goal:
  - *Fine-tune on large general purpose conversational dataset*
  - *Fine-tune again on the MultiWoz dataset*
- Completed Goal:
  - *Combined SGD training dataset with the MultiWoz training dataset*
  - *Fine-tuned on combined (supplemented) dataset*
- As discussed earlier, we added data from the Schema Guided Dialogue DataSet (SGD)
- We believe that further supplementing the fine-tuning data will further improve the models' performance across all decoding and evaluation metric combinations
- Models were evaluated on the same test set (MultiWoz.test_b) throughout all experiments.

Our Approaches: Supplementing Fine-Tuning Data
Results and Interpretation

	Greedy Decoding				Top-k Decoding				Top-p Decoding				Beam Search				
Model	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	BLEU	BERTScore	ROUGE	AVG	AVG
GPT-1	6	6	6	6	6	6	6	6	6	5	6	5.66	6	6	6	6	5.9
GPT-2	5	5	5	5	3	5	5	4.33	6	3	5	4.66	3	5	5	4.33	4.6
GPT-1 Fine Tuned	4	3	4	3.66	4	3	3	3.66	4	6	3	4.33	5	4	4	4.66	4.1
GPT-2 Fine Tuned	1	2	2	1.66	1	1	2	1.33	2	1	2	1.66	1	1	1	1	1.4
GPT-1 Fine Tuned-Supp	3	4	4	3.66	5	4	4	4.66	3	4	4	3.66	4	3	3	3.66	3.9
GPT-2 Fine Tuned-Supp	2	1	1	1.33	2	1	1	1.66	1	2	1	1.33	2	2	1	1.66	1.5

- The table above is the same as the table presented when discussing the performance of each model based on evaluation criteria
- We see that the supplementally fine-tuned GPT1 model performed (marginally) better, on average, than the originally fine-tuned GPT1
- However, we see that the originally fine-tuned GPT2 model slightly outperforms the supplementally fine-tuned GPT2 model, on average

Other Approaches

- While we were able to realize performance increases between GPT1 and GPT2, using additional fine-tuning data or decoding techniques (other than greedy decoding) did not significantly improve the performance
- We believe that the following approaches would help to further improve the performance on the conversational agent task;
  - Using domain specific fine-tuning and testing data
  - Ensembling different models prior to decoding
    - Average logit values before feeding into a decoding technique
    - Use different models for different domains of conversations (hierarchical-esq)
  - Fine tuning on an even larger data-set

Summary and Conclusion

- GPT2 >> GPT1
  - This comes as no surprise, the model trained on more, higher quality data, with more tunable parameters performs best
- Greedy decoding performed better than we thought
  - We expected beam search to yield the best results, but on average, we saw the greedy decoding resulting in the highest evaluation metrics
- Supplemental fine-tuning had little to no impact
  - Adding additional data did not alter the models' performance significantly
  - May be better suited for additional fine-tuning to be domain specific, also paired with a domain specific test set
- Other ideas
  - Focus on a certain type of conversation (be more domain specific)
  - Use some sort of ensembling methodology to predict the next token
    - Can combine different models' logits differently depending on the decoding technique used