XCS224U: LITERATURE REVIEW - CHATBOTS

Akshay Agarwal, Shashank Maiya, and Sonu Aggarwal

General Problem Area/Task Definition

While task-oriented chatbots are a well-understood space, the area of open-domain chatbots has emerged as an exciting area of NLP research in recent years. On the one hand, open-domain SOTA chatbot models can generate very impressive examples of human-model conversations, capable of being very engaging, conveying rich personality, holding their own in conversations, and even demonstrating some reasoning and some ability to generate humor. On the other hand, open-domain chatbots are still nowhere close to being human-like in some dimensions; they tend to be prone to repetition, lean towards simple responses, and lack any deep understanding of the subject matter. The following illustrate some key research efforts/findings in the field in recent years:

- Systematic frameworks that collate datasets, models, and evaluation techniques (automated and crowdworker), a notable one being parl.ai.
- Models for generating next responses in multi-turn dialogs, some key models being Retrieval, Generative, and Retrieve-and-Refine. SOTA models generally leverage Transformers.
- Techniques for automatic evaluation of dialog models, a key metric for automatic evaluation being perplexity.
- Techniques for human evaluation of dialog models, leveraging crowdworkers scoring model performance in defined dimensions.
- Relating model size with effectiveness, SOTA models ending up being around the 2B-parameter range.
- Techniques for equipping dialog models with certain major capabilities, such as imbuing models with personas, knowledge, empathy, and conversational consistency, and supporting datasets and evaluation techniques.
- Techniques for injecting and controlling “softer” human factors into dialog models, such as listening, avoiding repetition, asking questions, etc.,, and efficient techniques for evaluating such models.

We propose to build off the last area of research above and further investigate “soft” factors in dialog models, with specific examples of research hypotheses in the “Further Work” section below.

Concise Summaries of Articles

This section summarizes key papers starting in 2017 likely to directly inform our investigation, following a “dataset-first”, “bottoms-up” sequence, building up to the latest 2020 SOTA models leveraging these techniques.

The datasets used in these papers are generally collated within the parl.ai framework (Miller et al 2017) maintained by Facebook AI research. We propose to leverage this framework and these datasets - specifically, we will likely leverage the PersonaChat dataset described in Zhang et al 2018.

Miller et al 2017, Parlai: A dialog research software platform.

This paper defines an open-source software platform by Facebook for dialog research implemented in Python. It provides a unified framework for training and testing dialog models, evaluation, and a repository of machine learning models for comparing with other models. The repository includes popular datasets such as Wikipedia, SQuAD, Personality dataset, OpenSubtitles, DialogueDatabase, etc. Different models like memory networks,seq2seq, transformers, LSTMS, etc. are also integrated with the ParlAI.

The paper describes the main concepts (classes) in the ParlAI framework and the code structure. The following pre-trained models can be fine-tuned on dialogue databases:

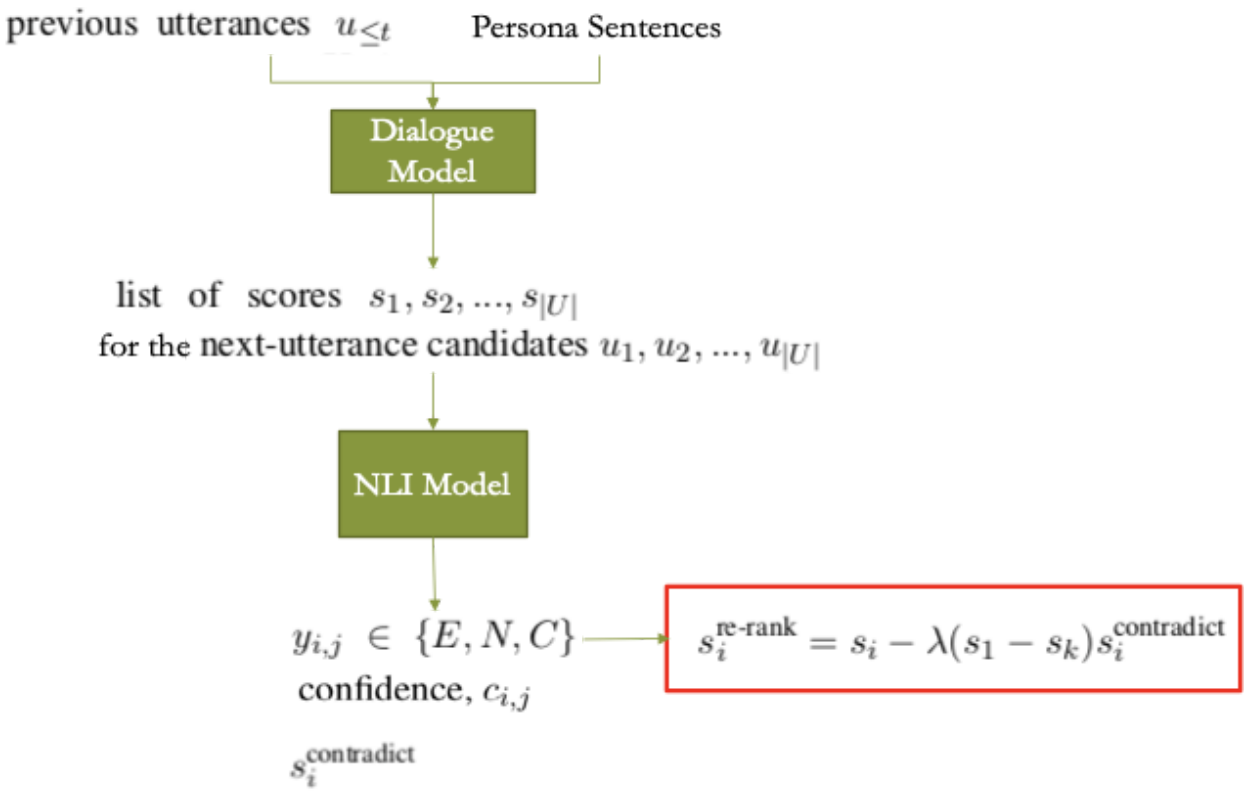
- Bert
- Fast_Text_Vectors
- Glove vectors
- Polyencoders
- KeyValue Memory Network
- Transformer Architecture
- Blender Bot

Welleck et al 2018, Dialogue natural language inference.

This paper provides a dialogue model to enhance the consistency of dialog generation, i.e. reduce contradictions and improve entailments of dialogs. It provides a new Dialogue NLI dataset, and develops efficient techniques for generating this dataset.

The model relies on a 2-step approach as described in the below figure:

1. In the first step, set of next-candidate along with their scores are obtained using a dialog model by using a set of previous utterances and a set of persona sentences,
2. In the second step, a NLI model is used to check for contradiction between these candidate utterances and any of the persona sentences. This is used to re-rank the candidate scores accordingly.



The dataset used to train the above NLI model (Dialogue NLI dataset) too, is obtained via a 2-step approach:

1. Based off the PersonaChat dataset in Zhang et al 2018, crowd-sourcing triple generation (entity 1, relation, entity2). This forms the Entailment dataset.
2. Applying automated techniques to construct entailment dataset (e.g., utterances and persona that share the same triple), contradiction dataset (e.g. via relation/entity swaps on the above, introducing numerics) and a neutral dataset (e.g. persona pairing, certain relation swaps based on the above).

Zhang et al 2018, Personalizing Dialogue Agents: I have a dog, do you have pets too?

Chit-chat models lack a consistent personality and are often not very captivating. To make chit-chat models more engaging, this paper defines a Personality dataset of 10k conversations created using MTurk. Ranking models, as well as Generative models, are used to train and evaluate on the dataset. Key-value (KV) Profile Memory Network (*Key-Value Memory Networks for Directly Reading Documents)* gives the best results on retrieval based systems while Generative Profile Memory Network (an extension of seq2seq) gives the best results on Generative models.

Rashkin et al 2018, Towards empathetic open-domain conversation models: A new benchmark and dataset.

This paper defines an Empathetic Dialogues dataset of 25k conversations using MTurk. Retrieval based as well as Generative methods are used to evaluate the model fine tuned on Empathy Dataset. Transformer-based Retrieval Architecture (*Learning Semantic Textual Similarity from Conversations)* and BERT model are used for Retrieval while Full transformer Architecture (with beam search) is used for Generative models. The model is pre-trained on Reddit Conversation and fine tuned on the empathy dataset. The candidates for Retrieval are chosen from Reddit Conversations, Daily Dialogue, and Empathy Dataset. Pre Built models (*FastText*) are used to add emotions as well as topics before training and inference. The generative model fine tuned on Empathy Dataset showed the best results (BLEU score), and Retrieval w/BERT showed the best results on Human Evaluation.

Dinan et al 2019, Wizard of Wikipedia: Knowledge-powered conversational agents.

The paper investigates unstructured Knowledge from Wikipedia across a broad set of topics to make more knowledgeable chatbots. It defines a large dataset with conversations directly grounded with knowledge retrieval from Wikipedia using Mturk. Knowledge retrieval is done via a pre-trained Information Retrieval system, which collects data from Wikipedia. Retrieval as well as Generative models were used to fine-tune on the Wikipedia dataset and evaluate the results. For generative models, two types of architectures were discussed

1. Two-stage models which first trains the task of selecting the most appropriate response from the knowledge base(Wikipedia) and then computing the utterance prediction
2. End-to-End model trained to produce the next utterance given all the sentences from the knowledge base(Wikipedia)

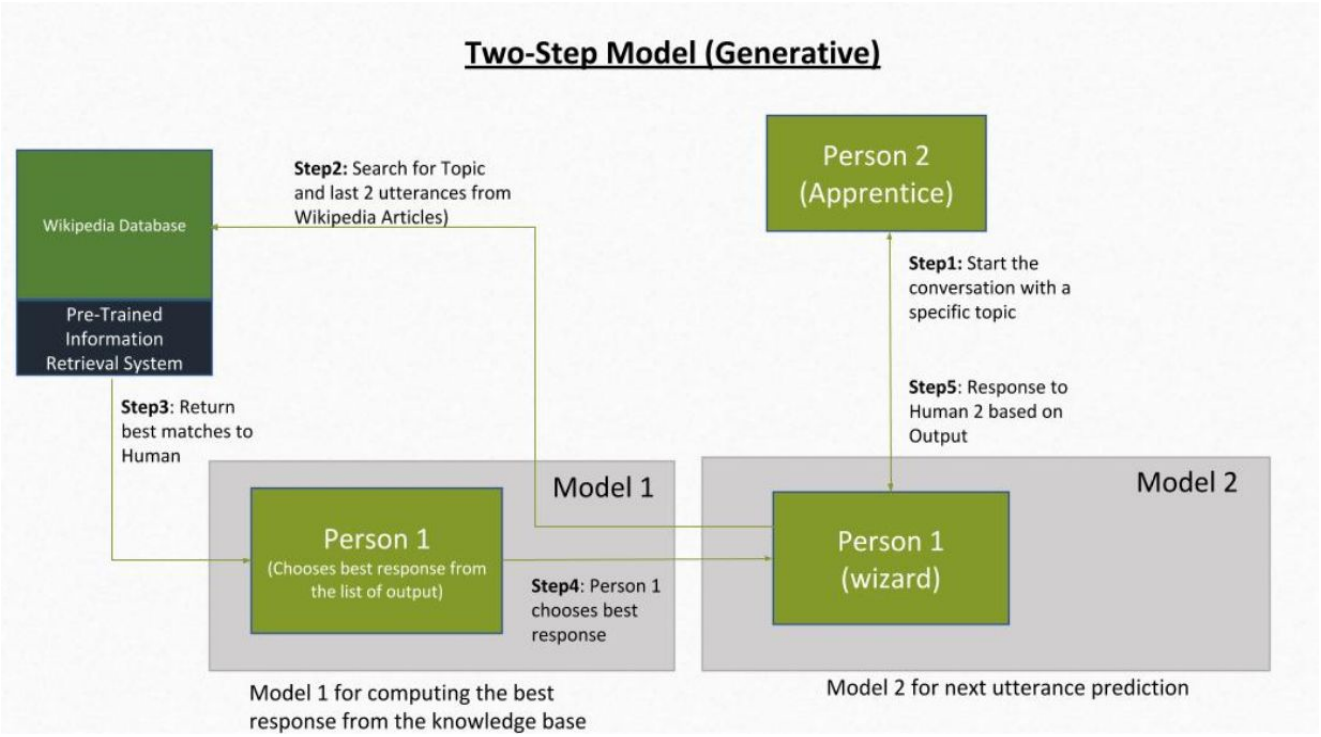


Figure: Two-stage Generative Model for next utterance prediction described in the paper

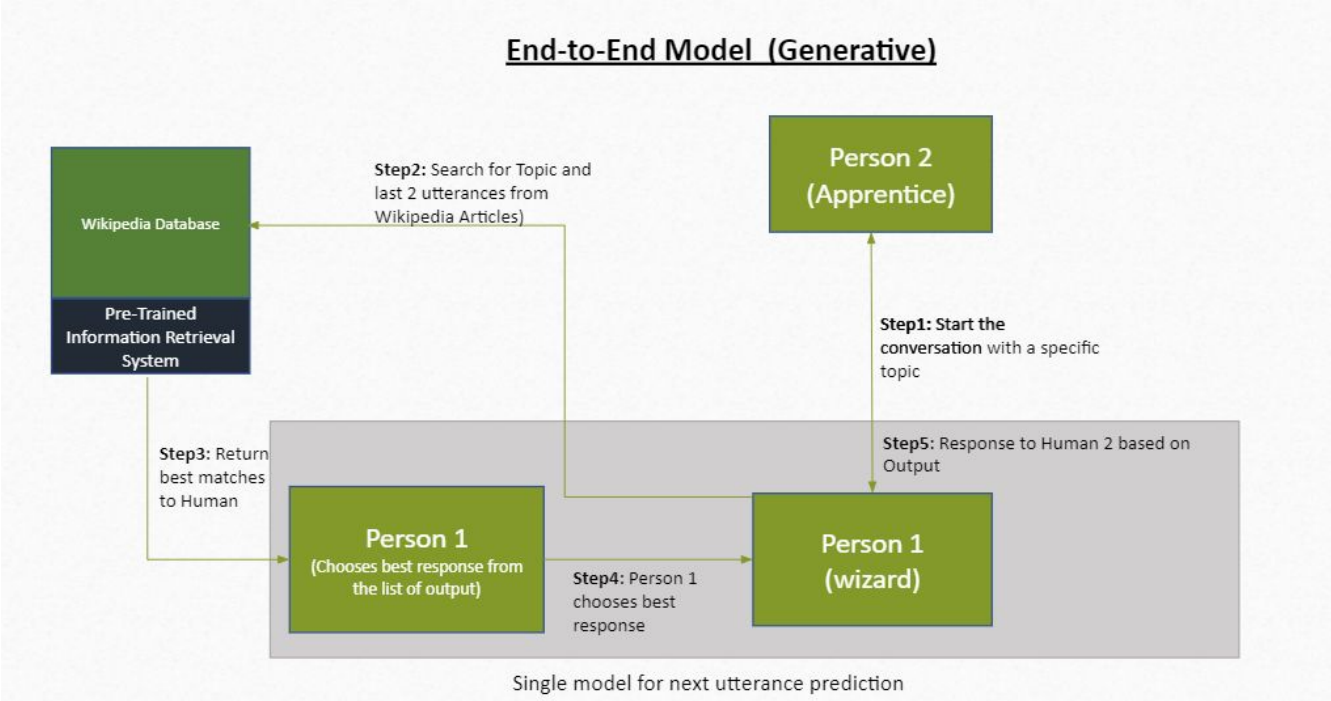


Figure: End-to-End model Generative Model for next utterance prediction in the paper

Transformer architecture(Vaswani et al 2017) was used to encode knowledge sentences and dialogue context. Attention mechanism is used to concatenate both the outputs which is then fed into the transformer decoder to generate the next utterance. The results for retrieval as well a generative model improves with the addition of a knowledge base.

See et al 2019, What makes a good conversation? How controllable attributes affect human judgments.

This paper describes how to make a good conversation in general. This is achieved by

- Reducing repetition (self repetition across utterances, self repetition within an utterance, repeating the partner utterance)
- Being more specific. For e.g., i/p : Yes, I'm studying law at the moment. Baseline o/p : That sounds a lot of fun! More specific o/p : I majored in practising my spiritual full time philosophy test
- Responding within the latest context of conversation. For e.g. the last i/p: My grandfather died last month. Baseline o/p: Do you have any pets? More related o/p : Im so sorry. Were you close to your grandfather?
- Asking suitable number of questions to keep the user more engaged

Li et al 2019, ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons.

This work provides a novel procedure to compare model effectiveness, involving humans comparing 2 full dialogues, where a human judge is asked to pay attention to only one speaker within each, and make a pairwise judgment between the 2 dialogues. The judgment question is a carefully worded question with 2 choices: speaker A or B, where the question measures a desired quality such as which speaker is more engaging, interesting, or knowledgeable.

This work:

- Defines a new (human) evaluation method ACUTE-EVAL with a clear mechanism that provides fast, cheap iteration. This evaluation method allows efficient reuse of data from prior papers, allowing new models to be evaluated independently of baselines, and significantly reduces the cost of annotation.
- Provides an explicit benchmark comparison between then-SOTA retrieval and generative models on PersonaChat and Wizard of Wikipedia, ranking various models on these tasks, using ACUTE-EVAL.
- Shows that ACUTE-EVAL human evaluations can be applied to model-model self-chats to compare models against one another, providing very efficient human evaluation (i.e. it might only take a human 2-3 hours to determine the better of two models with very high statistical significance.) Shows that the ranking against these self-chats is identical to ranking when humans are one end of the conversation.

- Provides optimized question choices based on experimentation, to maximize inter-annotator agreement.

Smith et al 2020, Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills.

This paper primarily talks about designing a model that can produce utterance from one of multiple skills (3 considered in this paper - persona, a particular topic or empathy). Firstly, a blended data set is constructed by crowd sourcing wherein the two speakers are given one skill from each of the pools and asked to converse (more details below). Later, several models are trained and performance is evaluated on these datasets.

Adiwardana et al 2020, Towards Human-like…

This paper:

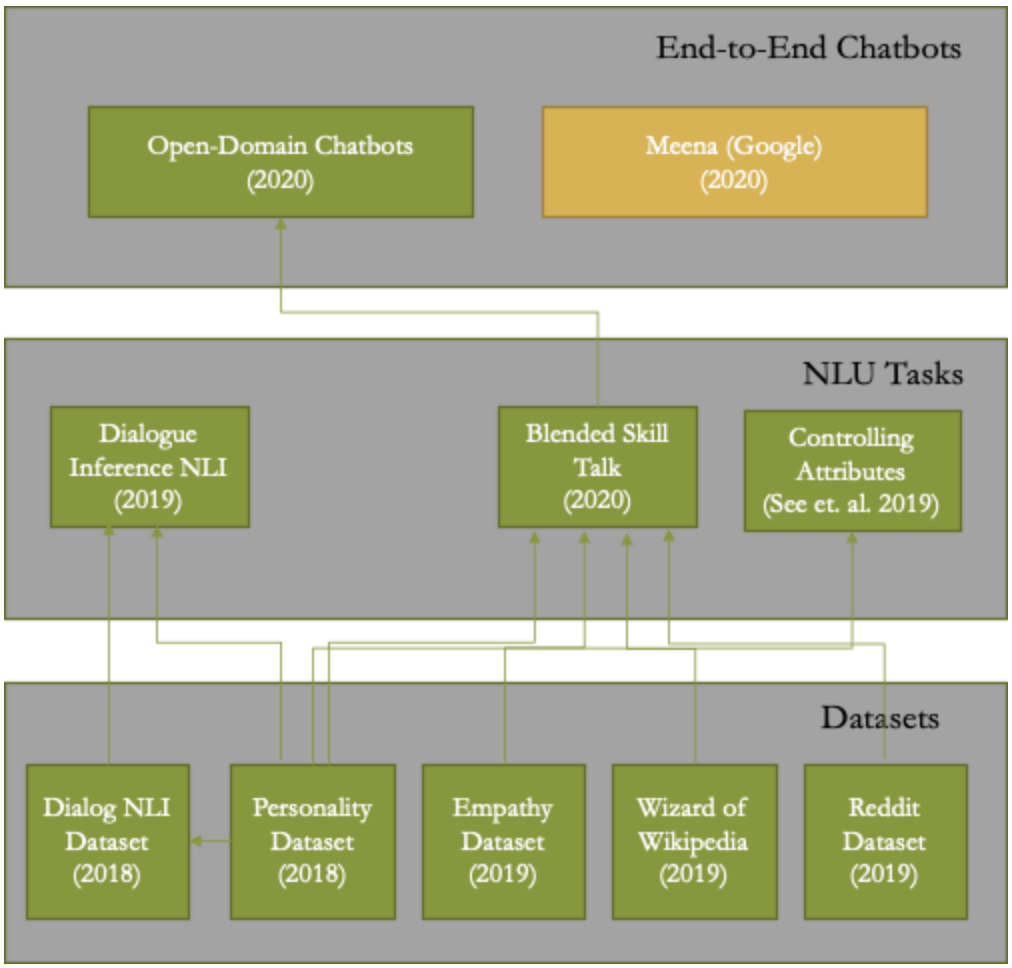
- Defines a human evaluation metric called Sensibleness and Specificity Average (SSA), where humans rate a multi-turn conversation based on sensibleness and on specificity.
- Concludes - via crowdsourced experimentation - that SSA is a good metric to capture the “humanness” of a bot interaction, as judged by humans. Determines the max possible SSA to be 86%, the SSA of humans themselves.
- Finds that SSAs of various chatbots tend to highly correlate with their perplexity, a simple automatic measure of any generative chatbot model. (The lower the perplexity, the higher the SSA.)
- Presents Meena, an open-domain chatbot built on an end-to-end NN (seq-to-seq, Evolved Transformer) with 2.6 B parameters.
- Evaluates Meena against other SOTA chatbots - Mitsuku, Cleverbot, DialogGPT, and XiaoIce. Meena tops at 79% SSA - pretty close to human 86% SSA.
- Presents fascinating examples of open-domain Meena chats.

Roller et al 2020, Recipes for building an open-domain chatbot

This paper shows how a chatbot that blends together a combination of conversational skills - engagingness, knowledge, and empathy - is superior to chatbots that do not. 3 versions of chatbots are evaluated - with 90M, 2.7B and 9.4B parameters respectively - and these generate some very impressive examples of open-domain conversations with humans. (These large-scale models, weights, agent code, evaluation code are all publicly available - at parl.ai/projects/recipes.) 3 decoding strategies are evaluated: “Retriever” (picking the best “canned response line” from a large corpus of training dialogs), “Generator” (using a Seq2Seq transformer to generate responses via a language model), and “Retrieve & Refine” (blending aspects of Retriever and Generator, with an interesting Knowledge retrieval variant of retriever - using TF-IDF lookups over a Wikipedia dump). Ultimately, the generative models appear to show the best performance.

Compare and Contrast

Miller et al 2017 provides a common Parl.ai platform that is leveraged by much of the later work for datasets, models, and evaluation.



The above diagram summarizes the dataset, NLU tasks and the end-to-end chatbots that were compared in our report. A number of papers build fundamental capabilities essential for developing functional open-domain chatbots in certain domain areas:

- Welleck et al 2018 develops the Dialogue Natural Language Inference (DialogueNLI) technique to improve the consistency of dialogue generation,
- Zhang et al 2018 imbues models with a sense of persona to make model responses more interesting and consistent (PersonaChat),
- Rashkin et al 2018 develops a technique to embed empathy into model responses, adapting a model response to the human context before it (Empathetic Dialogues, or ED),
- Dinan et al 2019 develops a technique to embed knowledge into model responses (Wizard of Wikipedia, or WoW), the knowledge in this case being retrieved from Wikipedia lookups.

See et al 2019 then develops techniques for adding additional “softer” human-like attributes to model behavior, such as reducing repetition, being more specific, listening, and asking questions. It finds that by carefully balancing model behavior in each of these dimensions, models can accomplish SOTA performance with much more limited training data.

Smith et al 2020 develops a technique to blend together multiple skills above in one model (Blended Skill Talk, or BST). This leverages the PersonaChat, WoW, and ED capabilities from the above.

Li et al 2019 develops ACUTE-EVAL, a more efficient and accurate (though still human) dialog evaluation framework for evaluating model performance, esp. on “soft” human measures (such as engagingness and humanness).

Finally, Adiwardana et al 2020 and Roller et al 2020 put together many of the above techniques to create and evaluate large SOTA models. Adiwardana et al (authored at Google) show that automated evaluation - via Perplexity - is a close approximation to human evaluations, and shows that its bot Meena beats previous models on this metric, by significant margins. Finally, Roller et al (authored at Facebook) claim that Perplexity - and automated techniques - are inadequate to evaluate chatbots on human-like dimensions above, and use the ACUTE-EVAL human evaluation framework to evaluate their own SOTA models, which leverage blended skills using the Blended Skill Talk (BST) framework of Smith et al 2020.

Future Work

The following are potential research directions to kick off in the context of our XCS224u project (our aim is to narrow down to one of these, or a specific aspect of one of these, for our project)

1. Can we calibrate humans (using their existing dialogs in ParlAi datasets) on “soft skills”, using the techniques of PersonaChat, WoW, ED, DNLI, and See et al 2019? Can we cluster humans based on “soft skill” metrics (defined in these papers, or perhaps introducing additional ones) to start identifying real “personalities” ?
2. Using techniques proposed in See et al - Reduce repetitions, improve specificity, response relatedness and engagingness by asking questions - on other datasets.
3. Does the See model translate well from a generic dataset to a domain specific dataset? Does it help become a better domain expert? [These might be more practical additive contributions]
4. Can we devise an additional “soft” attribute and measure model performance curve by controlling for it, as in See et al 2019? Examples of additional soft attributes we could devise:
    a. Topic-reuse from earlier in the conversation
    b. Acknowledging the speaker. (I see, you’re right, Hmmm, …)
    c. Exploring the agreement-disagreement spectrum. (“You’re right” vs “I don’t think so”)
5. Model’s tendency to clarify or ignore speaker? Keeping track of the topic or topics?Are there dimensions to the correlation between perplexity (or other automatically-evaluated metrics)and human-evaluated results, or other ? (Perplexity is one number; human evaluation can pivot on a number of different “soft” attributes…) [Powell fdbk: If theoretical - ask Chris Potts for permission if only computing metrics or only doing statistical work. This course focuses on coming up with an NLU/NLP task involving actual ML.]
6. Use an alternative ranking function for boosting entailment scores in Welleck et al 2018?
7. UDoes adapting a model’s responses to the “soft” attributes of a human speaker improve model performance? (E.g. somehow “matching” the speaker’s personality - e.g. roughly matching model response length to human “request length”, as one example?)
8. Investigate evaluator bias (or evaluator personality) in scoring model performance? This is a variation of #1 above.
9. Applying psychoanalytic diagnosis techniques to constructing rich bot personalities, modeling the multiple dimensions of actual human personalities seen in practice (e.g. “Obsessive-Compulsive”, “Narcissist”, “Manic-Depressive”, etc.), and seeing if that improves model performance, just like it might for real human-human conversations.
10. Improve information retrieval system to get more accurate knowledge results (e.g. in WoW-like tasks)?
11. Blending open-domain with task-oriented chatbot behavior? (We haven’t studied task-oriented behavior at all in our lit-review…)
    a. Blending emotions (using Empathy dataset) into a task-oriented chatbots
12. Use FastText API (for emotions and topics) in the Dialogue Inference NLI model? Try out models like GPT2?

References

1. Adiwardana, Daniel, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang et al. "Towards a human-like open-domain chatbot." *arXiv preprint arXiv:2001.09977* (2020).

2. Bo, Lin, Wenjuan Luo, Zang Li, Xiaoqing Yang, Han Zhang, and Daxin Zheng. "A Knowledge Graph Based Health Assistant." AISG/NeurIPS (2019)

3. Dinan, Emily, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. "Wizard of wikipedia: Knowledge-powered conversational agents." *arXiv preprint arXiv:1811.01241* (2019).

4. Ghandeharioun, Asma, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. "Approximating interactive human evaluation with self-play for open-domain dialog systems." In *Advances in Neural Information Processing Systems*, pp. 13658-13669. 2019.

5. Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. "The curious case of neural text degeneration." *arXiv preprint arXiv:1904.09751* (2019).

6. Humeau, Samuel, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. "Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring." *arXiv preprint arXiv:1905.01969* (2019).

7. Li, Margaret, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. "Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training." *arXiv preprint arXiv:1911.03860* (2019).

8. Li, Margaret, Jason Weston, and Stephen Roller. "ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons." *arXiv preprint arXiv:1909.03087* (2019).

9. Miller, Alexander H., Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. "Parlai: A dialog research software platform." *arXiv preprint arXiv:1705.06476* (2017).

10. Rashkin, Hannah, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. "Towards empathetic open-domain conversation models: A new benchmark and dataset." *arXiv preprint arXiv:1811.00207* (2018).

11. Real, Esteban, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc Le, and Alex Kurakin. "Large-scale evolution of image classifiers." *arXiv preprint arXiv:1703.01041* (2017).

12. Roller, Stephen, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu et al. "Recipes for building an open-domain chatbot." *arXiv preprint arXiv:2004.13637* (2020).

13. See, Abigail, Stephen Roller, Douwe Kiela, and Jason Weston. "What makes a good conversation? how controllable attributes affect human judgments." *arXiv preprint arXiv:1902.08654* (2019).

14. Smith, Eric Michael, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. "Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills." *arXiv preprint arXiv:2004.08449* (2020).

15. So, David R., Chen Liang, and Quoc V. Le. "The evolved transformer." *arXiv preprint arXiv:1901.11117* (2019).

16. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

17. Welleck, Sean, Jason Weston, Arthur Szlam, and Kyunghyun Cho. "Dialogue natural language inference." *arXiv preprint arXiv:1811.00671* (2018).

18. Zhang, Saizheng, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. "Personalizing dialogue agents: I have a dog, do you have pets too?." *arXiv preprint arXiv:1801.07243* (2018).