From Hybrid Dialogers to Neural Responders

João Luís Lins, Nikhil Reddy, Abdul Rafae Khan, Md Kowsher, Abhijeet Gusain, Yeshwanth Reddy, Xuting Tang, Preet Jhanglani, Nusrat Zahan, Mengjiao Zhang, Yu Yu, Darshil Shah, and Jia Xu

Stevens Institute of Technology

Abstract

Conventional open-domain dialogue modeling combines hand-designed dialogues and machine learning models. The fundamental architecture contains predefined conversational flows and responder's selection strategies as essential components to guide the chat. This design usually directs dialogues with hard-coded rules that easily fail from rare, new scenarios or out-of-domain conversations. For example, a topic classification model used to select an appropriate responder may be inaccurate. Or even, unseen words may cause mistakes in identifying the corresponding responder. This error propagation from a hybrid system largely hinders a chatbot's generalization ability and decreases the conversation quality. In this work, we introduce a pure neural responder aggregated architecture equipped with the knowledge base. Our main architecture does not contain rules or predefined heuristics, but only neural responders and rankers. We show that this simple and elegant architecture potentially outperforms the traditional logic flow and selection strategy-driven dialogue management on open-domain dialogue modeling. It is easy to code, simple to optimize, and robust to generate responses in new domains.

1 Introduction

This paper describes our models and systems developed at Alexa Prize SocialBot Grand Challenge 5 (SGC5) (Johnston et al., 2023). Performing open-domain dialogues with chatbots is a widely known challenge. Handling the size and breadth of the associated conversation context and selecting the most suitable approach to respond to each utterance are aspects that contribute to the complexity of conversations, even for humans (Csaky, 2019). To accomplish such a difficult task, conventional open-domain dialogue systems combine hand-designed dialogues and machine learning models to steer the chats. The existing architecture contains conversational graphs and selection strategies as essential components, following a hybrid chatbot design to control dialogue flow grounded on Natural Language Processing (NLP) components such as topic classification and sentiment analysis.

Hybrid approaches in conversation control often involve utilizing if/else logic gates and pre-scripted responses. These techniques can influence the direction of dialogue but may also introduce some level of bias towards more commonly encountered conversational paths. This can impact the desired balance of predictability in interactions. Additionally, the use of fixed heuristics in selecting responses from multiple systems can sometimes lead to coherence issues. For example, employing a topic classification model for response selection may encounter challenges in accurately identifying the most appropriate responder due to potential accuracy gaps or unfamiliar context. Such instances of error propagation can potentially limit the chatbot's ability to generalize effectively and could impact overall quality.

Envisioning a method that does not rely upon a static decision process, we propose a pure neural responder aggregated architecture equipped with knowledge-based responders. Our architecture does not depend on rules but is mainly based on neural responders. We show that this simple and elegant

5th Proceedings of Alexa Prize SocialBot Grand Challenge (Alexa Prize 2023).

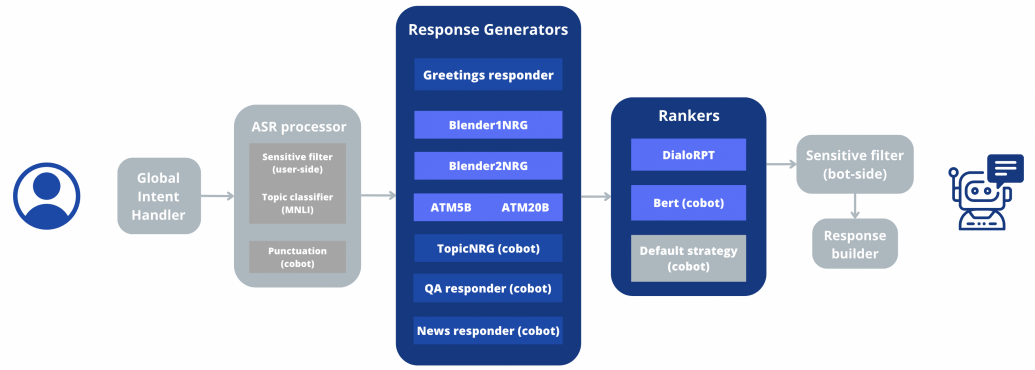

Figure 1: Architecture of our bot with the components from the Cobot toolkit (Khatri et al., 2018) along with pure neural responders.

design can potentially outperform the traditional graph-based and selection strategy-driven dialogue management on open-domain dialogue modeling. It is easy to code, simple to optimize, and robust to generate responses on dynamic utterances.

Our architecture illustrated in Figure 1 is comprised of three key parts: neural responders, rankers, and the baseline components. First, as our key components, we include several state-of-the-art large language models and introduce novel chat models as our neural and knowledge-based responders. Second, we use ranker to select the best response. Third, we streamline both the pre-processing and post-processing tasks by utilizing standard Automatic Speech Recognition (ASR), filters, and response builder modules. Our method mainly contains the following innovations:

- Neural Responders

  Relying on logic gates to activate a set of responders has shown to be effective for well-known and in-domain conversation flows and to adapt the system towards these same types of conversation. The error propagates into the rest of the pipeline whenever these models miss the target, compromising the quality of the conversation. In contrast, we turn off the selection strategy and build a pure neural-based chatbot, allowing all available responders to be triggered. We use several large language models, including Blenderbot 1 and 2, the Alexa Teacher model, to replace rules.
- Non-Neural Responders

  While LLMs learn human intelligence through massive data, some inputs can be extremely noisy and even false. On a few occasions, we call non-neural responders. When users ask a scientific question, we expect a professional answer with more confidence. Previously, such a knowledge base is commonly compiled in responders with a mixture of implementation of rules and information retrieval tools. To load updated results, we connect API interfaces to reach large knowledge databases, including Question/Answering (QA), Wikipedia, and News.
- Ranker Strategy

  By removing the logic flow and selection strategy, we rely on our ranker to meaningfully aggregate results from different responders. We use DialogRPT (Gao et al., 2020) and Bert ranker (provided by the SGC5 (SGC)).

Our experiments show promising results on reaching the state-of-the-art chatbot system performance without using rules in producing responses, demonstrating great potential for a new type of chatbot purely based on neural responders.

2 Related Work

Previous work on open-domain chatbots focuses on one of two directions: methods to achieve better neural generation or techniques to fuse knowledge bases into generative models. While systems that

prioritize pure generative models without attaching knowledge bases (Roller et al., 2020) increase the heterogeneity of the chat, they also prone the system to hallucinate. On the other hand, focusing too much on information retrieval (Ghazvininejad et al., 2018) can apply consistency and factualness to responses, but potentially lack the diversity of pure neural responses. To escape this paradox, instead of having just one LLM as the responder, one possible alternative is to have a good mix of them, each aiming to accomplish a specific task within a conversation. So, another architectural relevant aspect to contrast between different systems is the number of LLM responders and how the candidates are chosen as the final response. The usage of multiple LLMs in conjunction with ranking strategies, then, is a common approach used in numerous chatbots (University of California, 2021; in Prague, 2021; University, 2021), however, the general approach used to control the decision process of their activations is to embrace pre-scripted conversational flows (University, 2019), feature extractor models (Sun et al., 2019; Yoshikoshi et al., 2022) and selecting strategies in order to trigger a set of responders and finally rank the candidates with the help of a single classifier (Harrison et al., 2023). In our work, instead, we present an ensembled ranker to power a hybrid setup that can learn how to navigate the flow of the dialogue and finally select the best candidate response without hard-coded conversational graphs.

3 Model

In the following sections, we will describe all components in detail, including neural responders, rankers, and knowledge-based responders, focusing on their novelty compared to the existing work.

3.1 Neural Responders

Large language models (LLMs) is one of the most prominent recent innovation this year. Thus, updating our chatbot with LLMs facilities is inevitable. After comparing their capability, efficiency, and size, we directly host several state-of-the-art LLMs. Specifically, we use Blenderbots (Roller et al., 2020; Komeili et al., 2021; Xu et al., 2021), ATM5B (Soltan et al., 2022), ATM20B (Soltan et al., 2022), and Topic NRG. (Khatri et al., 2018).

3.1.1 BlenderBot 1 and 2

BlenderBot 1 (Roller et al., 2020) plays a major success in the conversational AI field mainly because it blends communication styles through the method of multi-task fine-tuning, in which multiple datasets, each with its own unique characteristics, are used to apply the required aspects to the output responses. Advancing further in the concept, we perform multiple fine-tuning experiments. In order to effectively apply a more engaging and extensive fashion to the response output, we use Parlai (Miller et al., 2017) default checkpoint with many others (BST (Smith et al., 2020), WoW (Dinan et al., 2019b), ConvAI2 (Dinan et al., 2019a), DailyDialog (Li et al., 2017), PersonaChat (Zhang et al., 2018)).

Taking advantage of the rapid and uncomplicated deployment framework of the Sagemaker/Huggingface (SM/HF) Deep Learning Containers (DLC) endpoints and envisioning testing purposes, we started using the default Hugging Face checkpoint of the model as one of the responders in our system. Subjective comparisons made by the test team show a significant increase in the quality of the responses, moreover, we also notice an increased user engagement seen through conversation logs, consequently pointing the direction in which we needed to invest our time: mine our own data to accomplish our unique desired style of communication.

In spite of BlenderBot 1's (Roller et al., 2020) excellent conversational abilities, two of its major drawbacks still need to be addressed: frequent hallucinations on grounds of a knowledge base; and the short-term memory in consequence of the limited context window size (model's positional embeddings). The second version of BlenderBot (Xu et al., 2021; Komeili et al., 2021) conveniently tackles those issues through the use of an attached knowledge base, and Long Term Memory, respectively.

Although the latency of BlenderBot2 is considerably higher, averaging 2.5s, its responses have proven to be more accurate and complete than the predecessor.

Entering the subject of deployment, in the first experiments phase, due to the more time-consuming implementation of the Parlai (Miller et al., 2017) framework, we maintained the HF/SM DLC

endpoint by copying the updated parameter weights into the HF Blenderbot 1 model's Architecture. But from the second experiment onwards, both with BlenderBot 1 and 2, we started using the Parlai Agent itself to provide inference with the help of Cobot's remote modules. This update is a major progress in our system by virtue of:

- Substantially reducing the latency of the inference from an average of 1.7s to 0.7s.
- Applying modularity to the system and, in consequence, more flexibility to be auto-scaled accordingly to user traffic.
- Opening the possibility of using diverse Parlai decoding hyper-parameters such as: beam, delayed beam, nucleus, topk inference and beam size.

3.1.2 Alexa Teacher Models

While BlenderBots generally exhibit impressively engaging performance, we occasionally observe instances where they fall short in providing satisfactory responses due to limited training data for certain scenarios. To ensure a more reliable and consistent performance, we have integrated Alexa Teacher Models (ATM5B and ATM20B) (Soltan et al., 2022) into our system alongside BlenderBots. The integration of these models aims to complement the capabilities of BlenderBots, particularly in scenarios where it may struggle due to limited training data.

By incorporating ATM models, we leverage their unique strengths and enhance the overall performance of our chatbot. The ATM models bring their own expertise and linguistic knowledge, enabling them to generate more contextually appropriate and informative responses in situations where Blenderbots may fall short. This combined approach of leveraging multiple LLMs enables us to offer a comprehensive and robust conversational experience, ensuring that our chatbot is well-equipped to effectively handle a wide range of user interactions.

3.1.3 Topic NRG

In our exploration of state-of-the-art LLMs, which include Blenderbots, ATM5B, ATM20B, and Topic NRG, we observe that each model has its own strengths and weaknesses. While Blenderbots generally perform well when users engage in longer and more detailed conversations, we notice that some users tend to provide short utterances consisting of only one or two words, such as "yeah", "no", "okay" or "cool". In these cases, we find that Topic NRG produces more engaging and appropriate responses. Based on this observation, we incorporate Topic NRG into our system to better cater to users who communicate using shorter utterances. This has significantly improved user ratings, especially for conversations where users are less engaging and provide minimal input.

3.2 Non-Neural Responders

In addition to the LLMs mentioned earlier, we also incorporate non-neural responders into our system to address specific scenarios where factual information is required. These responders, provided by the Alexa team, serve as valuable resources in generating accurate responses. Specifically, we employ two non-neural responders: the QA responder and the News responder.

3.2.1 QA Responder

The QA responder is a valuable component in our chatbot system, especially when users seek information on specific topics. While the neural responders excel at engaging in general conversations, their knowledge may be limited to what they have been trained on. However, the world is vast and constantly evolving, and there may be areas where neural responders lack expertise. Its purpose is to extract relevant information from its knowledge base and generate concise and precise responses.

3.2.2 News Responder

Similarly, we employ the news responder to address users' inquiries about current events and news. The world is dynamic, and new information emerges rapidly. While the neural responders strive to provide up-to-date information, there may be instances where their knowledge is not current. In such cases, the news responder becomes instrumental in delivering the latest news updates to users. By leveraging the news responder, we ensure that users receive accurate and timely news information,

enhancing their overall chatbot experience. Whether users are interested in global events, sports, entertainment, or any other news category, the news responder is adept at fetching relevant and current news articles to satisfy their information needs.

3.3 Ranking Strategy

In this section, we detail our ranking strategy to enhance response candidate selection and improve user experience.

3.3.1 BERT Ranker

Initially, we employ the default BERT Ranker's API (SGC5-provided) for ranking multiple response candidates. However, this yields sub-optimal outcomes in around 30% cases, either due to failure in selecting the best response or inaccuracies in candidate scoring.

3.3.2 DialogRPT

To address this issue, we experiment with the Dialog Ranking Pretrained Transformers (DialogRPT)(Gao et al., 2020), a DialoGPT model(Zhang et al., 2020) trained on Reddit feedback data (Gao et al., 2020). This ranker assesses the relevance of candidates to the ongoing conversation, considering up to five previous turns of dialog as context and selecting the most appropriate response.

4 Evaluation and Experiments

In the following, we will describe our offline evaluation methods and experimental results.

4.1 Data Acquisition

In an open-domain conversational task, data plays an important role in building a chatbot that can handle good conversations. We need control over what features our model should learn and what it should avoid. Therefore, we carefully create our own dataset by focusing on diverse strategies/domains/personas in order to accomplish engaging, coherent, factual, and diverse dialogues. We have collected conversational speeches based on ChatGPT and GPT3.5/4 (Brown et al., 2020; OpenAI, 2023) API.

4.1.1 Source

After developing the mining script to collect data from BlenderBot 3 UI (Shuster et al., 2022) and GPT 3.5/4 (Brown et al., 2020; OpenAI, 2023) API, we notice that the latter has more potential as we can not only collect entirely new conversations but also customize style, domain, length and practically every aspect of them via appropriate prompts.

We continuously analyze real user conversations from the current SGC5 with the purpose of getting insights about the topics that users are mostly interested in or even about their communication style. Then, to mine conversations in which such aspects appear as traits, we adapt the mining algorithm, accomplishing, finally, synthetic conversations over various structures and topics (ranging from movies, music, day-to-day activities, etc.).

Data is collected with the purpose of covering all conversational scenarios/intents (e.g. giving an opinion, suggestion, agreement/disagreement, encouragement, questioning, expressing emotion). A variety of prompts are made and programmatically filled in order to create different dialogue scenarios. We put as many different prompts to GPT models so that our models can handle different scenarios in conversation without relying on any rule-based algorithm. This strategy provides our system a robust capability of talking with any user in a changing from one topic to another in a continuous manner, in contrast to a tree-based dialogue flow in which the transitions are discrete and blind to not defined situations.

Topics	%	Average Turns approx	Average words per conversation
Movies	33	10	377
Sports	14	10	404
History	12	13	334
Technology	11	13	352
Music	9	12	545
Food	5	12	467
Arts	4	9	481
Topic Change	4	12	469
Fun random questions	3	5	667
Day-to-Day Activities	2	12	514
Health	2	14	628
Hobbies	1	11	504

Table 1: Data Distribution

4.1.2 Domains

The data collected along with their distribution of topics is shown in Table 1. Although we collect data over the whole period of competition, we conveniently divide the experiment into three phases, namely, Phases 1, 2, and 3. Some types of prompts used for data acquisition are shown below:

- In Phase 1, we collect data about the most popular topics (movies, music, sports, and food). The main aim of these prompts is to have a baseline for increasing qualitative responses from the bots. We notice an improvement in these topics but also realize a few problems related to the bot not being able to handle some specific user personas, such as the ones who repeatedly reply with short utterances. In this case, particularly, our model tends to prompt the end of the conversation conversations, not being able to change the topic or suggest a new one to extend the dialogue.

  Examples of such prompts:

  - Generate a long conversation between a human and a bot on sports. The human starts the conversation and the bot ends the conversation. The bot is knowledgeable, friendly, and provides engaging responses.

- In Phase 2, we focus on a wide variety of topics, including arts, hobbies, day-to-day activities, etc. We also collect a specially designed range of conversations to specifically cover short utterance-styled users. Prompts are also focused on creating a more specific type of conversation by introducing personality to each speaker. This introduces the bot having a personality of his own based on the topic of the conversation also the length of the conversation increased significantly.

  Personality-based prompts:

  - Generate a long conversation between a human and a bot on football. The human starts the conversation and the bot ends the conversation. The bot is a football coach and the human is a football fan.
  - Generate a long conversation between a human and a bot on football. The human starts the conversation and the bot ends the conversation. The bot is a football news reporter and the human is a professional football player.

- Phase 3 prompts focus on topic suggestion and topic switching in the conversation. They are more challenging to create as there is a limit on the number of turns per conversation from OpenAI APIs. We needed to get a few turns of conversation on some initial topic and then change the topic to a sub-topic of the same domain or a completely different domain. We also handled short utterances which we usually receive in user interaction.

  Short utterance prompts: Generate a long conversation between a human and a bot in the domain of sports. The human starts the conversation and the bot ends the conversation. The bot is knowledgeable, friendly, and gives engaging responses. Humans respond with extremely short replies 40% of the time.

Topic change prompts: Generate a long conversation between a human and a bot. Start a conversation on basketball and bot change the topic after 6 turns.

4.2 Evaluations

A critical aspect is the creation of a robust external evaluation framework. To achieve this, we develop a custom evaluation framework. This framework's primary aim is to evaluate these models' performance before deployment, by comparing model outputs with reference responses across specified domains.

LLMs, including BlenderBot1, BlenderBot2, Falcon (Penedo et al., 2023), ATM 5B, and ATM 20B, underwent a thorough offline examination to gauge their effectiveness and efficiency before deployment.

The evaluation process involves using user utterances from our reference conversations as inputs to the language models and subsequently recording the models' responses. These responses are then compared against the reference set to measure accuracy and relevance.

To choose the most effective evaluation metric, we explore several methods, such as BLEU score (Papineni et al., 2002), ROUGE score (Lin, 2004), token accuracy, and cosine similarity between sentence embeddings. The choice of evaluation metric is critical as it determines how the machine's responses will be scored and compared to human-generated responses.

Despite being widely used in machine translation and summarization tasks, BLEU and ROUGE scores present some limitations in this context. BLEU, which primarily measures precision by quantifying how many words in the machine's output appear in the reference translations, often overlooks contextual and semantic nuances. Likewise, the ROUGE score, which predominantly focuses on recall, assessing how many words in the reference summaries appear in the machine-generated summary, may also miss subtle semantic differences.

However, cosine similarity proves to be a more suitable choice in this scenario. By measuring the cosine of the angle between two non-zero vectors, cosine similarity captures semantic and syntactic relationships in a multi-dimensional vector space. This approach ensures that not only the presence or absence of specific words but also their semantic roles and contextual relations in sentences are considered when scoring model responses.

In conclusion, adopting cosine similarity as an evaluation metric, given its strength in maintaining semantic integrity, is a significant breakthrough in our chatbot development process. This metric enables a more comprehensive, context-sensitive, and semantically faithful evaluation of chatbot responses, thereby enhancing the reliability of our evaluation framework.

4.3 Experiments

Baseline Components

1. Global Intent Handler: Situated at the initial stage of the Cobot pipeline, the Global Intent Handler plays a crucial role in generating the initial portion of the response. Typically, it responds to Launch Requests with a welcome prompt and to Stop and Cancel intents with a goodbye prompt. Its primary function is to set the tone and context for the subsequent stages of the pipeline.
2. ASR Processor: Positioned prior to the NLP feature extractor, the ASR Processor analyzes the ASR (Automatic Speech Recognition) text and has the capability to generate a response based on its analysis. If the user's utterance contains offensive language, the default ASR Processor redirects the topic by providing a prompt.
3. Response Builder: Serving as the final step in the Cobot pipeline, the Response Builder is responsible for packaging and delivering the response to the Alexa Skills Kit API. It performs two key tasks: converting the chosen response into a Text-to-Speech (TTS) response for ASK and generating the re-prompt speech.

To compare the results of each combination/hyper-parameter selection, we adopt offline evaluation metrics, namely, Token Accuracy and Perplexity with respect to test sets of different datasets.

Multi-tasking approaches using industrially known datasets such as BST (Smith et al., 2020), ConvAI2 (Dinan et al., 2019a), WoW (Dinan et al., 2019b), DailyDialog (Li et al., 2017) do not provide desirable results, overfitting the model to a shorter response style of communication already overused by the market. Our best model, then, was trained purely with our proprietary dataset.

As shown in Table 3, when analyzing some written outputs, remarkable differences are observed in terms of proficiency of the model in the trained domains, moreover, paying attention to the offline evaluation metrics (see Table 2), we can statistically assert that, indeed, the model is learning the response style of our dataset. We also use these metrics to choose between versions of models trained with different hyper-parameters. We perform this selection while taking care to not make the model unlearn traits from the original dataset that BlenderBot 1 has been ultimately fine-tuned on (BST (Smith et al., 2020)), as generally it is academically known as a good chit-chat source.

We upgraded BlenderBot 1 (Roller et al., 2020) two more times using our own ChatGPT collected data, each aiming for a specific strategic improvement while not only progressively increasing the volume of dialogue turns (400k / 550k) but also developing a more refined process to clean the collected data. In the second and third phases, to accelerate the training, we use model parallelization with 4 A10G GPUs (4x24GB).

Model	Own Test Set		BST Test Set	
Default	0.57	6.31	0.573	6.313
Version 1	0.707	3.23	0.393	21.83
Version 2	0.691	3.38	0.427	15.81
Version 3	0.642	4.17	0.42	16.35
	Token Accuracy	Perplexity	Token Accuracy	Perplexity

Table 2: Offline evaluation metrics (BlenderBot 1 experiments)

Model	Utterances
Input	Let's talk about sports
Default	Do you like sports? I'm a big fan of the NFL. I love to talk about it.
Version 1	Sports are a great way to stay active and stay healthy. What's your favorite sport?

Table 3: Comparison between default checkpoint and V1 outputs. The input utterance is an example generated by our team.

The fine-tuning process and implementation of Blenderbot 2 (Komeili et al., 2021; Xu et al., 2021) follows the same methodology as the former version, the major faced challenge is how to fit such a big model in the remote instance without running out of space in the default disk partition. As a solution, we use the new cobot's EFS (Elastic File System) integration to store the model and parallelize the inference with 4 A10G GPUs.

4.4 Testing

4.4.1 Offensive Filter

After receiving a text from ASR, to handle sensitive topics and terms, we initiate an offensive filter. The process is composed of two phases, in the first, making use of a topic classifier model (BART-large (Lewis et al., 2019) trained in the MNLI dataset (Williams et al., 2018), we determine whether each user utterance is in the domain of politics, finance, medical consulting, or any other controversial topic that we would like to avoid. Given the utterance, the model predicts scores for each specified label, we then define a threshold to compare the scores with. In the second phase, we split the utterance into 1-7 N-grams to check if any word is present in an extensive prohibited list. Finally, if any offensive utterance is found, the response is built by randomly selecting a topic redirection prompt from a set that has been continuously tuned along the competition. In this sense, we can rapidly deliver an appropriate response without having to run the rest of the system pipeline. Such topic redirection is also used for handling offensive words to continue the conversation without conflicting with the user's negative emotions. We experiment with designing our redirection prompts

in diverse ways, from the usage of jokes or witty phrasings to specifically suggesting another topic to talk about. The best one so far has proven to be explaining to the user that we are not proficient on the topic and asking for one to steer the conversation in another direction. Lastly, it's worth mentioning that dealing with intentionally offensive users is still a challenge due to the sensible balance between being more rigid and classifying normal utterances as offensive or being lenient and incurring in undesirable conversation domains.

4.4.2 Capabilities and Limitations

Capabilities:

- Broad Topic Coverage: The SocialBot exhibits the ability to engage in conversations on a wide range of topics, displaying versatility in its conversational repertoire. These encompass but are not limited to sports, music, food, movies, art, hobbies, health and fitness, fun random questions, technology, history, and day-to-day activities.
- Expertise in Specific Topics: The SocialBot demonstrates the capacity to delve deeply into specific subjects, providing expert-level insights and information to users seeking in-depth discussions. For instance, it can provide detailed information about various sports, including regional styles of play, prominent players, and top teams. Furthermore, it can discuss team strategies, win records, and other relevant details.
- Topic Suggestions: When the conversation reaches a point where the user appears to have exhausted their input, the SocialBot proactively suggests new topics, facilitating the smooth transition to fresh conversational avenues.
- Initiative in Starting Conversations: The SocialBot possesses the capability to initiate conversations with users, introducing new topics and initiating engaging interactions.
- Long-term Engagement: The chatbot excels in sustaining lengthy conversations by attentively discussing the user's interests, ensuring a high level of user engagement and continued interaction.

Limitations:

- Mathematical Calculations: The SocialBot cannot perform mathematical calculations, rendering it unable to provide precise answers to math-related queries.
- Reasoning and Logical Questions: The SocialBot is not equipped to handle reasoning tasks or answer logical questions, limiting its capacity to engage in complex cognitive processes.
- Retention of Previous Conversation: Due to the limited input-token capacity, the SocialBot faces challenges in retaining the entire context of previous conversations, potentially leading to a loss of continuity and coherence.
- Smoothness of Topic Changes: Occasionally, the chatbot experiences difficulties in effecting seamless transitions between topics, resulting in less fluid conversational shifts and potentially impacting user satisfaction.

5 Conclusion

We move from the rule-based, dialogue-manger-driven, and knowledge-based controlled conventional dialogue modeling to a novel, simple, and elegant learning structure mainly based on an aggregation of neural responders. Our architecture is easy to implement, flexible to adapt to new components, and robust to predict accurate and interesting conversation responses. This work shows a new possibility of a pure machine learning constructed chatbot that challenges needed rules.

6 Acknowledgement

We appreciate Amazon Alexa Prize SocialBot Grand Challenge 5 to fund and support our project. We are thankful to OpenAI providing us access to data acquisition. We acknowledge the individuals who have contributed to the project.

References

Alexa Prize SocialBot Grand Challenge 5. https://www.amazon.science/alexa-prize/socialbot-grand-challenge.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Richard Csaky. 2019. Deep learning based chatbot models.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019a. The second conversational intelligence challenge (convai2).

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Vrindavan Harrison, Rishi Rajasekaran, and Marilyn Walker. 2023. A transformer-based response evaluator for open-domain spoken conversation.

Czech Technical University in Prague. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization. In *Alexa Prize SocialBot Grand Challenge 4 Proceedings*.

Michael Johnston, Cris Flagg, Anna Gottardi, Sattvik Sahai, Yao Lu, Samyuth Sagi, Luke Dai, Prasoon Goyal, Behnam Hedayatnia, Lucy Hu, Di Jin, Patrick Lange, Shaohua Liu, Sijia Liu, Daniel Pressel, Hangjie Shi, Zhejia Yang, Chao Zhang, Desheng Zhang, Leslie Ball, Kate Bland, Shui Hu, Osman Ipek, James Jeun, Heather Rocker, Lavina Vaz, Akshaya Iyengar, Yang Liu, Arindam Mandal, Dilek Hakkani-Tür, and Reza Ghanadan. 2023. Advancing open domain dialog: The fifth alexa prize socialbot grand challenge. In *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, Chandana Satya Prakash, Mukund Sridhar, Fabian Triefenbach, , and Apurv Verma. 2022. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China. Association for Computational Linguistics.

Stanford University. 2019. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations. In *Alexa Prize SocialBot Grand Challenge 3 Proceedings*.

Stanford University. 2021. Neural, neural everywhere: Controlled generation meets scaffolded, structured dialogue. In *Alexa Prize SocialBot Grand Challenge 4 Proceedings*.

Santa Cruz University of California. 2021. Athena 2.0: Discourse and user modeling in open domain dialogue. In *Alexa Prize SocialBot Grand Challenge 4 Proceedings*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation.

Takumi Yoshikoshi, Hayato Atarashi, Takashi Kodama, and Sadao Kurohashi. 2022. Explicit use of topicality in dialogue response generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 222–228, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?

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.