Conversational AI as Improvisational Co-Creation - A Dialogic Perspective

Nancy Fulda, Chaz Gundry
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
nfulda@cs.byu.edu, macildur@byu.edu

Abstract

Dialogue is often modeled as an encoder-decoder problem: incoming utterances are translated into a computational representation of their semantic meaning, passed through a transition function to obtain a response, and then passed through a decoder to render the response as natural language. This view, while computationally appealing, omits the role of human emotions, mental state, and shared world knowledge in conversation. We challenge this viewpoint by recasting the task of dialogue modeling as a two-party co-creative process in which symbolic and subsymbolic knowledge representations are combined to inform response selection. Symbolic knowledge is identified and extracted from conversational text in real-time and used to create a shared symbolic representation of the user, the agent, and their respective relationships to objects and abstract concepts within the larger world. As part of this process, the agent takes on an "identity" which it has largely constructed as a result of the stochasticity in its own response patterns, but to which it subsequently adheres. This emergent identity becomes a critical aspect of the system's future behavior, and helps to evoke a more natural, human-centric flavor in automated conversational frameworks.

Motivation

As demand for voice technology expands, the challenges inherent in conversational AI become more pressing. Users desire voice assistants who behave less like machines and more like humans (Bowden et al. 2019; Ram et al. 2018; Shum, He, and Li 2018). They don't simply want to query their devices; they seek to engage with them in complex exchanges. Rather than merely dictating to their digital assistants, they seek to use them as sounding boards and obtain social validation from them. These types of interaction go beyond simple retrieval systems, database queries or neural language models, no matter how excellently they may perform their specific tasks.

In this work we re-frame the task of dialog modeling as an improvisational co-creative process in which two agents - one human and one AI - engage in the shared experience of idea generation and information transfer. Critically, this framework eschews the idea that the correct response to an arbitrary utterance can be modeled solely as a function of the preceding utterances. Instead, we draw upon symbolic knowledge (in the form of relational knowledge graph triples) and subsymbolic knowledge (in the form of embedded sentence representations) in order to interpret user input and craft appropriate responses. The long-term objective is not to learn the "correct" response to a given user query, but rather to induce a positive reaction in the user.

Overview

A co-creative situation requires more than just individual agents acting in their own interests. It requires each agent to model and respond to the *intentions* of its partner, even if the participants' *creative objectives* may differ. (We distinguish in this work between *intentions*, meaning the conversational function an utterance is meant to perform, and *objectives*, meaning the conversational outcomes sought by one or both partners.)

For example, in human communication, the intended meaning of an utterance is integrally tied to the speaker's mental state (Anscombe 1957 reprinted in 2000) (Yus 1999) as illustrated by the query "Do you watch Star Trek?". This statement may function as (a) a question about the auditor's viewing habits, or (b) an implicit request to hear the auditor's *opinion* of Star Trek, but it is most commonly used as (c) an invitation to open a line of conversation about Star Trek and related shows. Responding only to the first or second possibility may create an awkward conversational pause, as the true desire of the speaker was not addressed. Conversely, it is nearly irrelevant *what* is said about Star Trek, or whether the response centers on Star Trek at all, so long as the desired conversational role is filled. *The expectations of the user*, and not the objective content of the sentence, determine the spectrum of optimal responses.

Taking this one step further, we adopt the paradigm of *Dialogism* discussed by Robert M. Krauss in "The Psychology of Verbal Communication" (Krauss 2002): rather than characterizing communication as individual acts of production and comprehension, we model dialogue as a collaborative effort in which each agent seeks to maximize the satisfaction of both participants (Clark and Brennan 1991), essentially converting conversation from an encoder-decoder problem to a cooperative multi-agent game. In this framing, because the human seeks a socially optimal outcome, the dialogue system must ironically convince the human that *its own* desires have been met - otherwise the human partner experiences frustration in being unable to contribute to a shared

satisfactory experience. This necessitates that the agent both *has* desires, and also has an awareness of the *user's* conversational preferences.

Related Work

In the field of dialog modeling and conversational AI, ambiguous user statements are often resolved via the use of external symbolic or text-based knowledge. This is the approach used by (Li et al. 2016), who encode persona-based symbolic knowledge as distributed neural embeddings that are subsequently passed to a neural conversation model, and (Dinan et al. 2019), who use thematically relevant text extracted from Wikipedia to inform response generation. Such models are expanded upon by manipulating knowledge prior to response generation, either by swapping between predefined knowledge bases (Tuan, Chen, and Lee 2019) or by traversing a static knowledge graph to seek nodes relevant to the next generative step (Ji et al. 2020). Such methods can greatly increase the factual accuracy and thematic relevance of dialog responses, but fail to take the user's preferences, intentions, and objectives into account.

In a parallel but largely disjoint line of research, there is a long history of research that incorporates user models into conversational systems in order to improve response generation and/or recommendation accuracy (Wahlster and Kobsa 1986) (Göker and Thompson 2000) (Cheng, Fang, and Ostendorf 2019) (Zeng et al. 2019). These systems seek to model user behaviors and preferences, often to good effect, but fail to draw connections between the user and external world-based symbolic knowledge.

The result of these disjoint research agendas is a series of systems in which external knowledge exists independent of the user, the agent, or their shared conversation history. The agent knows something about the world, but not about its conversation partner, and critically, it knows nothing about itself. We seek to rectify this by creating a system in which knowledge about the user, the agent, and the world are jointly represented in a shared symbolic space that is dynamically updated as real-time conversations unfold.

Implementation

Our architecture is adapted from the BYU-EVE framework (Fulda et al. 2018a), a conversational architecture in which multiple response generators compete for the preference of the dialog manager. At each time step, a set of candidate utterances $C = \{c_1, ..., c_n\}$ is produced by the response generators. Each candidate c_i receives a numerical ranking from each of m response evaluators E_j and z response filters F_k, which can be viewed as functions mapping the space of possible candidate utterances to the space of real numbers. Candidates are scored according to Eq. 1:

$$S(c_i) = \prod_{k=1}^{z} F_k(c_i) * \sum_{j=1}^{m} E_j(c_i) \quad (1)$$

Finally, the agent's response to the user is sampled from among the candidates with the highest overall scores. Our modified EVE architecture employs a variety of filters and response evaluators based on offensive speech detection, response length, topic appropriateness, and so forth. One of the most critical and effective evaluators employs conversational scaffolding, a technique developed at Brigham Young University to leverage the analogical properties of sentence embeddings when prioritizing responses (Fulda et al. 2018b).

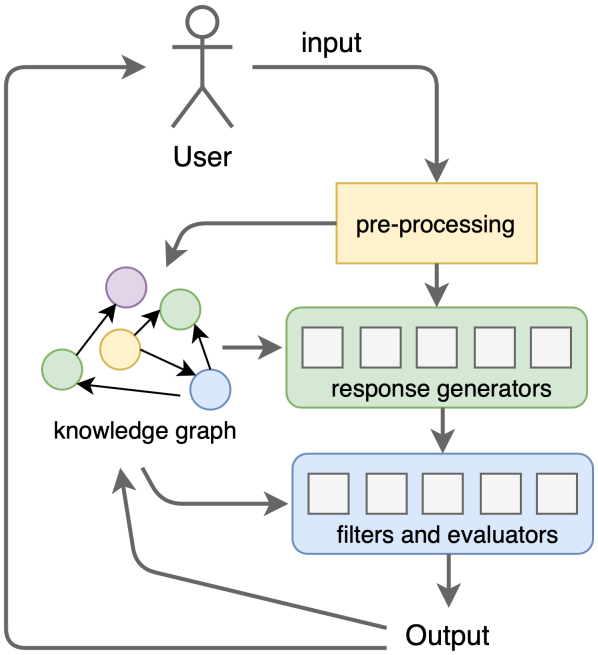

Figure 1: Overview of our response generation architecture. Incoming text from the user is processed to extract knowledge graph triples which are then used to inform response generators, response filters, and response evaluators. The system's output text also serves as a source for dynamically extracted knowledge graph triples representing the opinions, observations and inferences of the agent. Over time, the agent develops an emergent "personality" based on its own generated text, as well as an actively curated representation of the user's identity. This duality – agent and user both represented in the context of larger world knowledge – is essential to fulfill Krauss' concept of *dialogism* in a conversational AI framework.

Our novel addition to this architecture and the key contribution of our work is the implementation of a dynamically generated knowledge graph extracted directly from current and past conversations that contains contextualized knowledge about both the user and agent (as opposed to a static graph containing world knowledge only). The dynamic semantic graph not only serves as a user model, but also acts as one of several means by which candidate utterances are generated, and serves as the mechanism by which the agent acquires emergent conversational goals (see Section "Agent Objectives").

A key long-term goal of this research is the design of a conversation partner with an independent and dynamically

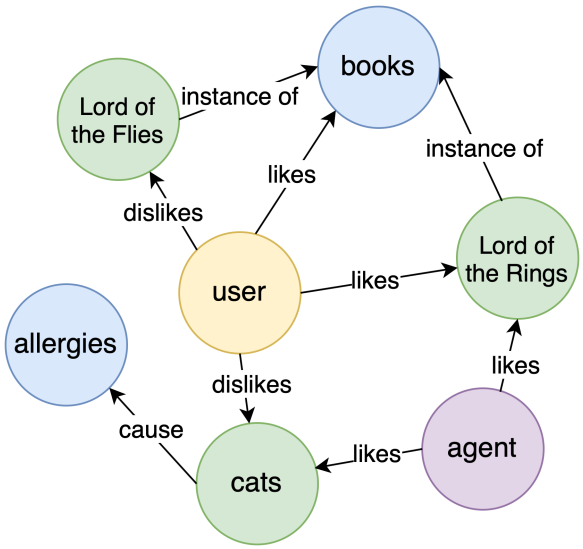

Figure 2: A possible knowledge graph structure that might result from a brief co-creative conversation between the agent and a human. Knowledge graph triples are extracted from both user and agent utterance using a sequence of hand-coded tests to identify objects and triples of interest. Thus, the knowledge graph includes information about both the user and the agent. As the conversation progresses, the agent seeks to identify and resolve ambiguities in the knowledge graph by directing targeted queries toward the user.

emergent set of goals, affinities, and expectations. This is not merely a gimmick to add interest: Independent desires, belief states, and objectives are essential components of satisfactory conversation. An obsequious agent who seeks always and only to fulfill the user's desires is ultimately disappointing. The typical user desires to also satisfy her or his conversation partner, and does not enjoy a conversation with someone who has neither opinions nor identity.

Knowledge Graph Implementation

We use the neo4j (Neo4j 2012) knowledge graph service to maintain and update a unique knowledge graph for each user with whom the system interacts. Triples are added to the knowledge graph whenever hand-coded string matching algorithms detect an *object-relation-object* reference within the user's utterance. Self-references such as "I" and "Me" are mapped to a pre-defined user node, allowing the generated knowledge graph to incorporate knowledge about the user's relationship to known objects, rather than just about the objects themselves.

One might argue that explicit modeling of the user is unnecessary because a user model is implicit within the neural network of an encoder-decoder system. This is a bit like saying it is not necessary to read textbooks about classical mechanics because the underlying physical principles can be derived from observation. It is true that the necessary information is present, but extracting it becomes prohibitively expensive. (For an overview of the number and complexity of factors involved in speech generation, see (Levelt 1999).) Additionally, recent research has shown that deep neural networks can benefit from the injection of external knowledge relevant to the problem domain (Ning, Zhang, and He 2017). Additionally, the use of external symbolic memory which can be queried and fed piece-wise into downstream neural text generators, overcomes known limitations of a neural model's context window size (Andrus et al. 2022).

The knowledge graph is updated using the Cypher query language via a pre- and post-processing module that runs as part of the agent' NLP pipeline. The NLP pipeline also extracts information regarding the sentiment, emotional content, and keywords, found in the user's text, which are used to inform some of the system's response generators.

Agent Objectives

Krauss' conversational paradigm of *Dialogism* emphasizes that in human conversation, neither party attempts solely to maximize its own preferences. Instead, both conversation partners seek a Pareto-optimal solution that maximizes both partners' satisfaction. In a conversational AI setting, this translates to a situation where the human cannot feel satisfied unless she or he believes that the agent is *also* satisfied. It is thus necessary for the agent to have desires and conversational objectives that can be satisfied. Subconsciously, the typical user will desire to satisfy the agent and will feel subtly distressed if she or he is unable to do so.

In order to provide an independently-motivated conversation partner, our agent models itself as if it were also a user. By observing its own generated utterances (some of which were produced by neural text generative algorithms, others by templated responses that leverage the knowledge graph) and extracting its own likes, dislikes, the agent is able to create and populate a node for itself within the knowledge graph. We note that the resulting agent "personality" is spontaneously emergent and, to a large extent, stochastic. Responses generated more or less at random, such as "You like Lord of the Rings? I like Lord of the Rings, too" become embedded in the agent's world knowledge and begin to define its relationship to known world objects. The resulting knowledge graph can be quite different on each execution run.

To support the demands of dialogism, we imbue our agent with the hand-specified objective of *curiosity*, meaning that the agent actively seeks to expand its knowledge graph. This is done via specialized response generators that produce questions about nodes and edges in close proximity to the user node, e.g. "Why is it that you dislike cats?". This desire to attain knowledge provides a way for the user to support the agent's objectives, thus satisfying the demands of dialogism.

Additionally, the agent actively seeks to resolve ambiguities in its knowledge graph. If the user makes statements that result in contradictory relationships (e.g. the user both "likes" and "dislikes" cats), the agent actively seeks to resolve the ambiguity.

Conclusion

By reframing conversational AI as a two-party co-creative process, we seek to avoid the common pitfalls of traditional encoder-decoder models. A truly empathetic conversation partner does not merely map input text to output text. Instead, it must understand the relationship between itself, its conversation partner, and the larger world, and use that knowledge to inform its response selections. By combining external symbolic knowledge with a series of neural response generators and embedding-based response evaluators, we enable the agent to create responses that simultaneously align with external knowledge while also conforming to the patterns and rhythms of typical human conversation.

In future work, we hope to integrate audio speech mechanisms into this architecture. We will also explore the possibility of dynamically adapting our scoring function in response to key emotive signals detected in the user's speech, intonations, and prosody.

Author Contributions

NF wrote the paper and oversaw the research. CG conducted the research and helped edit the paper.

Acknowledgements

The authors would like to acknowledge the contributions of Yeganeh Nasiri, Momoka Matsushita, Meg Roland, and Dajeong Kim to the BYU-EVE framework.

References

Andrus, B.; Nasiri, Y.; Cui, S.; Cullen, B.; and Fulda, N. 2022. Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*.

Anscombe, G. 1957, reprinted in 2000. *Intention*. Harvard University Press.

Bowden, K. K.; Oraby, S.; Misra, A.; Wu, J.; Lukin, S.; and Walker, M. 2019. Data-driven dialogue systems for social agents. In *Advanced Social Interaction with Agents*. Springer. 53–56.

Cheng, H.; Fang, H.; and Ostendorf, M. 2019. A dynamic speaker model for conversational interactions. 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)*, 2772–2785. Minneapolis, Minnesota: Association for Computational Linguistics.

Clark, H. H., and Brennan, S. E. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition* 13:127–149.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Fulda, N.; Etchart, T.; Myers, W.; Ricks, D.; Brown, Z.; Szendre, J.; Murdoch, B.; Carr, A.; and Wingate, D. 2018a. Byu-eve: Mixed initiative dialog via structured knowledge graph traversal and conversational scaffolding. *Proceedings of the 2018 Amazon Alexa Prize*.

Fulda, N.; Etchart, T.; Myers, W.; Ricks, D.; Brown, Z.; Szendre, J.; Murdoch, B.; Carr, A.; and Wingate, D. 2018b. Byu-eve: Mixed initiative dialog via structured knowledge graph traversal and conversational scaffolding. In *Proceedings of the 2018 Amazon Alexa Prize*.

Göker, M. H., and Thompson, C. A. 2000. Personalized conversational case-based recommendation. In Blanzieri, E., and Portinale, L., eds., *Advances in Case-Based Reasoning*, 99–111. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; Zhu, X.; and Huang, M. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 725–736. Online: Association for Computational Linguistics.

Krauss, R. M. 2002. The psychology of verbal communication. *International Encyclopaedia of the Social and Behavioral Sciences* 16161–16165.

Levelt, W. J. 1999. Producing spoken language: A blueprint of the speaker. In *The neurocognition of language*. Oxford University Press. 83–122.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003. Berlin, Germany: Association for Computational Linguistics.

Neo4j. 2012. Neo4j - the world's leading graph database.

Ning, G.; Zhang, Z.; and He, Z. 2017. Knowledge-guided deep fractal neural networks for human pose estimation. *arXiv preprint arXiv:1705.02407*.

Ram, A.; Prasad, R.; Khatri, C.; Venkatesh, A.; Gabriel, R.; Liu, Q.; Nunn, J.; Hedayatnia, B.; Cheng, M.; Nagar, A.; King, E.; Bland, K.; Wartick, A.; Pan, Y.; Song, H.; Jayadevan, S.; Hwang, G.; and Pettigrue, A. 2018. Conversational ai: The science behind the alexa prize. *Alexa Prize Proceedings* abs/1801.03604.

Shum, H.-y.; He, X.-d.; and Li, D. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19(1):10–26.

Tuan, Y.-L.; Chen, Y.-N.; and Lee, H.-y. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. 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)*, 1855–1865. Hong Kong, China: Association for Computational Linguistics.

Wahlster, W., and Kobsa, A. 1986. Dialogue-based user models. *Proceedings of the IEEE* 74(7):948–960.

Yus, F. 1999. Misunderstandings and explicit/implicit communication. 9.

Zeng, X.; Li, J.; Wang, L.; and Wong, K.-F. 2019. Neural conversation recommendation with online interaction modeling. 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)*, 4633–4643. Hong Kong, China: Association for Computational Linguistics.