Conversational AI Take-Home Exam

Daan Brugmans
Radboud University
daan.brugmans@ru.nl

Despite significant improvements in language modelling techniques in the past half decade, current state-of-the-art conversational interfaces are far removed from completely mimicking human conversation. There exist varying aspects of human-to-human conversations that modern conversational AIs are not making use of yet, not to their fullest extent, or even at all. In this essay, I will briefly touch on some of these foundations of human conversation, and relate them to the Alexa voice assistant.

Skantze (2021) provides a broad overview of the many components that underlie the turn-taking process in human interaction. Turn-taking stands at the core of human conversation, and in his review paper, Skantze shows that humans use many varying cues during conversation to indicate turn-taking, -yielding, and -holding. These include, but are not limited to, utterance syntax, semantics, pragmatics, and prosody, and the current speaker's gaze, gesturing, and breathing. Human-to-human conversation is thus multimodal, and uses both auditory and visual cues to determine turn-holding and -yielding cues. This multimodality of conversation is something that most state-of-the-art conversational AIs lack: for example, Alexa can only process auditory information, so visual turn-taking cues cannot be conferred onto her. Skantze's Furhat is an example of a conversational AI that can interpret auditory and visual cues simultaneously, utilizing the multimodality of conversation.

However, even Furhat will have trouble when attempting to truly mimic human-to-human conversation. Levinson (2019) shows that the multimodality of conversation extends beyond the turn-taking mechanism, and is built into human conversation itself. Human conversation, they state, consists of many design features, which include not only turn-taking, but also metacommunication, such as repair mechanisms and backchannels, and a motivation underlying the conversation - two aspects of conversation relevant for this essay - and more. Metacommunication, for example, has been shown to be vital in human-to-human conversation: according to work by Dingemanse and Enfield (2024), the repair mechanism, for example, seems to help listeners in decreasing cognitive load significantly by bouncing the load back to the speaker. By initiating repair, the speaker can assist the listener in decreasing the cognitive load, making the conversation more of an effort of both parties. This utilization of repair is lacking in conversational AI systems, as it is very hard to implement; the open-endedness of conversation not only causes the range of potential repair initiations to be very large, but as Dingemanse and Enfield show, the formulated repair initiation should preferably be as specific as possible in order to minimize cognitive load. Due to the cost of repair initiation in AI systems, it is often foregone as a feature, as is the case with Alexa, who will only make clear that she has not understood the speaker once they are done speaking entirely, and her "repair" will be very non-specific ("Sorry, I didn't get that").

Elaborating on Alexa's turn-taking skills, Alexa is not actually really capable of picking up natural turn-taking cues at all: she must rely on a manual "waking word" that tells her that she should listen to the speaker, and that the turn will be passed to her when the speaker is finished speaking. This design exists due to the purpose that Alexa has: being a voice assistant who will help a user fulfill requests. Enfield and Sidnell (2017) talk about the concept of "action in interaction", which means that (a purpose of) action is rooted into conversation. When we communicate, we have a reason for doing so, and we use conversation to relay the action that we wish to fulfill. When we speak to Alexa, she is expected to listen to what we have to say and is expected to fulfill the request embedded into our utterance; she was even designed this way. Her design supports her in making her role in the conversation clear, as otherwise, she would have

difficulties in participating in worthwhile conversation. This is another issue conversational systems face: in human conversation, the speaker and listener hold each other accountable, and they can be "punished" on the basis of this accountability. Conversational AIs do not have a sense of such accountability and do not seem to be able to "understand" it. For example, requesting Alexa to play a song that she cannot find may have her say "Sorry, I don't know that". Alexa's reaction is quite worthless, but her lack of sense of accountability prevents her from understanding that.

In fact, it may be argued that Alexa does not really "understand" the conversation at all. Turing (1950)

References

Mark Dingemanse and N.J. Enfield. 2024. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42.

NJ Enfield and Jack Sidnell. 2017. On the concept of action in the study of interaction. *Discourse Studies*, 19(5):515–535.

Stephen C. Levinson. 2019. Natural Forms of Purposeful Interaction among Humans: What Makes Interaction Effective? In *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*. The MIT Press.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

A.M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.