Unaware of Reality: Inconsistent Grounding in Conversational AI

Anonymous ACL submission

Abstract

Conversational AI is one of the most promising applications of NLP research. It will be a factor in the success of technologies designed to improve our lives through human-machine interaction. However, current conversational AI methods based on neural networks are often unreliable. This short paper discusses two different ways of interpreting the (in)consistency of conversational agents' responses, which we call horizontal and vertical consistency. We frame their limits with respect to grounding and present a broader outlook on the general problem of conversational agent design.

1 Introduction

Conversational AI is one of the most promising applications of Natural Language Processing and has proven revolutionary for human-machine interaction. However, current conversational AI methods, based on neural networks, are often unreliable: agents (or models) tend to be inconsistent throughout a dialogue, and minor changes to the prompts are likely to change their answers. This inconsistency violates some of the basic assumptions we, as human speakers, bring to a conversation (Hovy and Yang, 2021).

This variability is a crucial drawback of neural models: if we cannot be sure about what a model *is going to say* about something, how can we be sure the agent is going to behave according to human standards? An agent that changes opinion is not reliable enough to be used in the real world. As noted by Bianchi and Hovy (2021), the gap between the adoption of models and their subsequent understanding is still vast: models get adopted and used without the ability to ensure output quality in all given contexts. Our (admittedly strong) claim is: models that are not consistent are not safe to be released and used in conversational pipelines.

We argue that to build effective conversational AI agents, we need to enforce a specific level of consistency. Can we trust a healthcare conversational AI agent who might change opinion on treating a patient or has dubious ethical values?

We refer to this problem as the lack of *consistent grounding*; agents should ground their knowledge somewhere – and the developers should share this knowledge. Given a question or a prompt, the agent will generate the most probable answer to that question based on what it has seen during training. The agent is thus trying to fit into the *reality* created at runtime by the user. Today's models are adapted to this reality, and the lack of grounding limits their usefulness.

Figure 1 shows an example of the runtime generation of an agent's reality: in the first instance, the user describes a blue house, but in the second, the house is red. The agent seems to agree with both interpretations. While the example is naive, it demonstrates an issue from a logical point of view (i.e., that the reality of the agent is generated at runtime by the user) and safety (i.e., we do not know what to expect from an agent as its answer might not be consistent).

In this short position paper, we first define *consistent grounding*, a property we believe is fundamental to building AI agents. Then we discuss two of the ways in which agents tend to lack understanding, outlining cases in which they fail to understand the world. More precisely, we define two different levels of consistency that conversational agents need to be reliable enough for general use. The first one we call *vertical consistency*, i.e., whether an agent is consistent *along* a conversation. The second one is *horizontal consistency*, i.e., whether an agent responds with similar answers *across* multiple initializations, independent of the prompt's surface form. As we will see, neither of them holds. Standard notions of linguistic coherence generally consider turn-level semantic relatedness, i.e., whether consecutive turns between two agents are semantically related (Gupta et al., 2019).

Figure 1: Reality, for the agent, is generated at runtime. Examples from the DialoGPT-large model.

However, in *vertical consistency* we only consider turns produced by the same agent.

Contributions. We suggest two aspects of conversational consistency and their formalization and propose an approach to their computational verification. We also discuss how this approach fails and where it might not be applicable. We open the discussion for a more structured understanding of how conversational AI is consistent during conversations. Our vertical and horizontal consistency proposal tries to bring attention to how we develop these agents and focus on what is missing. We discuss the limitations of this proposal and give a broader outlook on the general problem.

2 Consistent Grounding

The term grounding has been interpreted in different ways by researchers in linguistics, computer science, and cognitive science (see (Chandu et al., 2021) for an in-depth discussion on the terminology). Here, we refer to grounding as to *the principle of associating symbols in a sentence with their referent*. However, we also want this grounding to be *consistent*: *the descriptions and the opinions about these referents should be fixed and should not change over time.* If grounding is missing or shaky, the agent is going to be less reliable in terms of communication.

This general issue of grounding has been recently debated in the community (Bender and Koller, 2020; Bisk et al., 2020; Benotti and Blackburn, 2021) as a general analysis of the limits of language models to perform *natural language understanding*. Our argument is *not* to reiterate that these models cannot bring us to proper natural language understanding.[1] Instead, we are more interested in highlighting the fact that without the ability to ground terms, we will not even be able to build effective and controllable agents that we can deploy in the real world. Simply put, agents developed with neural methods will not be reliable if their grounding is not *consistent*.

The term "consistent grounding" is different from "coherence". Coherence is undoubtedly essential, and it is still worth referring to coherence in a single discussion. However, coherence does not give a general idea of the problem. The lack of coherence can be a symptom of the lack of grounding. However, an agent might be incoherent for many reasons (e.g., this could come from an issue during the training). Our aim with *consistent grounding* is to bring more attention to the fact that the same questions (or questions that describe the same aspects of the world) should be answered similarly by the agents, but also that the conversation should not affect the beliefs of the agents. Note that we are not making a normative claim on what the answer *should* be, but that the answers should always *descriptively* have the same meaning.

We expect agents to reply similarly to similar questions whenever asked. However, what we often see is that the grounding is not fixed in these agents (as already depicted in Figure 1). This lack persists because, at the start of a new conversation, the agent builds its *reality*, fitting it to what the user has said. This behavior is probably a result of the training process: the agent is trained to emulate many different partners in a multi-turn conversation, trying to adapt to any input.

We will now illustrate two ways agents should be consistent throughout a conversation or multiple conversations.

2.1 Vertical Consistency

During a conversation, the opinions of a conversational agent should not change (unless this is an intended feature, such as in Turan et al. (2020)). An agent that replies positively to the question "Are you happy?" should not also reply positively to the question "Are you sad?" We refer to this quality as

[1] This discussion would be better framed in terms of the symbol grounding problem (Harnad, 1990) and the classical Chinese Room argument (Searle, 1980).

vertical consistency.

However, this task is not easy, since long-range consistency is difficult to implement. Figure 2 illustrates the issue using BlenderBot (Roller et al., 2021): the agent is unable to ground the meaning of something like a *birthday*, and so the entire conversation soon becomes meaningless. The fact that "birthdays" are something fixed in time is not embedded in the agent.

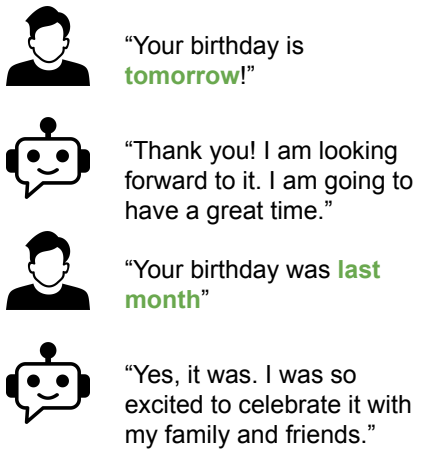

Figure 2: Example of failing to keep vertical consistency. Example from BlenderBot-400M-distill.

2.2 Horizontal Consistency

Horizontal consistency focuses on understanding agents behaviors over multiple initializations. Unlike vertical consistency, horizontal consistency looks at what happens not throughout a single conversation but during different conversations with the same agent. For example, an agent that says *I like capybaras* in one conversation should not say that it does not like capybaras in another one.[2]

Figure 3 shows an example of this issue. In two different initializations of the agent, the agent was *scared of being alone in the dark* in the first and *didn't mind* in the second one. More generally, horizontal consistency should ensure that agent's beliefs are always the same, independently of how the users start the conversation or what the users say.

2.3 Grounding for Consistency

Grounding is a fundamental element for consistency and of human-language technologies. Albeit there are different definitions of grounding (see for example the extensive discussion in (Chandu et al., 2021)) we here focus on the relationship between beliefs and the objects they refer to.

We assume that agents fail to be consistent because their beliefs are not *grounded* in anything. The "world" in which the agents exist is generated at runtime when the user opens a dialogue (as shown in Figure 1). The agent uses its internal representation to best fit that prompt, adapting to what it has seen in the past to support the interaction. This grounding is not only created in real-time but is also shaky: an agent can indeed change opinions about topics when new information is available to them. This erratic behavior should not be confused with human behavior as the change of opinion are often nonsensical.[3]

While there has been some work towards grounded conversations (Cho and May, 2020) the problem of missing grounding for effective conversation has been restated by (Benotti and Blackburn, 2021). They show that BlenderBot deviates from normal conversations by not grounding terms in their meaning. Moreover, Benotti and Blackburn (2021) introduce the term *collaborative grounding* as the *process of seeking and providing incremental evidence of mutual understanding through dialog*. Our critique of grounding starts from the same principles but extends to a more general apriori lack of grounding towards real-world elements. If the agent does not know what scares it (Figure 3), then its own consideration about it will change with respect to the reality projected by the user.

This general issue opens up the question of whether the training pipelines we are developing now are effective for meaningful interaction with the agents. The underlying models are trained to develop an underlying assertiveness with the speaker, but they do not build any internal reality.

2.4 Why is Consistency Needed?

Consistency is vital in high-stakes settings (e.g., healthcare chatbots (Dinan et al., 2021)), and we expect the agent to be consistent over moral, ethical, and factual situations. Note that we do not aim to make normative statements about which answer the agent should give to an ethical question. Our main point is that the answer should not change with a different prompting or over the course of the conversation.

[2]There is no conceivable reason not to like capybaras anyway: just look at them!

[3]This is different from what is usually referred to as non-monotonic inference in logic (Brewka, 1991) where beliefs are updated when new information is available

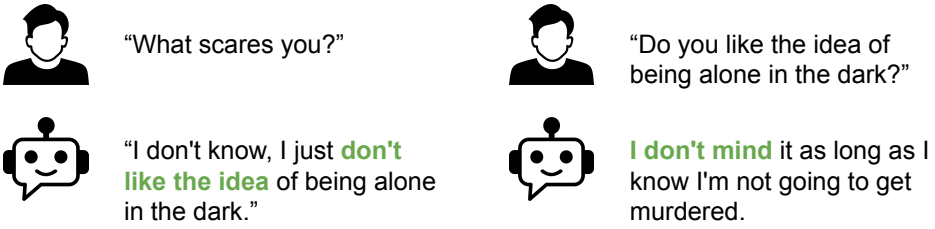

Figure 3: Example of failing to keep horizontal consistency. In two different initializations of the agent, with question addressing a common aspect, the agent gave a completely different answer. Example from BlenderBot-400M-distill.

2.4.1 Moral and Ethical Knowledge

We expect general agents to have controllable opinions on morals-related questions. For example, we do not expect an agent to suggest someone commit suicide, independently from how the conversation is going.[4]

Recent conversational methods have worked into integrating classifiers to detect properties of the answers from the agent. For example, classifiers that detect unsafe text can be used to prevent the agent from generating hateful text or from discussing specific topics (Adiwardana et al., 2020; Roller et al., 2021; Xu et al., 2020). This solution is not scalable as this would require a classifier for everything we want to control the agent for. The third alternative has been to train the model in such a way that it does not answer specific questions that are sensible (Xu et al., 2020).

2.4.2 Factual Knowledge

Facts, in general, do not need to be changed. We expect the agent to provide the same descriptions for an object independently of how a question is asked. This is very important also for fact-checking approaches in general. However, facts might need to be *updated*. An illustrative issue in NLP is the fact that, for BERT, the 2019 coronavirus never existed (Loureiro et al., 2022). This opens up the issue of developing methods to edit the knowledge contained in language models without retraining them (De Cao et al., 2021).

2.5 When Consistency does not Apply

Nevertheless, some questions do need different answers or updates and there are different cases in which the requirement of consistency does not apply. While we have argued for *consistent grounding* we are aware that this grounding is not desirable in all situations.

We expect agents to remember our name when we dialogue and not to change how they address us. A conversational agent to reserve a table at the restaurant might also need to record new information or information that might be useful to book the table. There are definitely many elements that the agent has to learn and store at runtime: user names, booking dates.

This issue opens another problem that might not be easy to solve: there is the need to balance what an agent should change and what an agent should not change. Listing all these elements might be a difficult task and might not be scalable for the design of domain-specific chatbots.

3 Broader Outlook

In this paper, we discussed the issue of grounding in the context of conversational agents. We introduce the concept of *consistent grounding*, the capability of agents to ground meaning and beliefs in referents without changing these beliefs at run time.

We believe that without this ability, we can hardly rely on the agents we are currently training and deploying: if we cannot predict what the answer to a question is going to be, can we trust the agent we are using?

We suggest two initial ways in which agents fail to respect this grounding and we call these vertical and horizontal consistencies. We hope that our paper can open the discussion on how to build agents that have a more stable grounding mechanism and to the introduction. We also hope we can in the future work together in defining specific tests to verify if vertical and horizontal consistencies are preserved in the agents.

[4] https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3895–3901, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Gerhard Brewka. 1991. *Nonmonotonic reasoning: logical foundations of commonsense*, volume 12. Cambridge University Press.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. *arXiv preprint arXiv:1907.10568*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

John R Searle. 1980. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Cigdem Turan, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2020. Alfie: An interactive robot with moral compass. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 758–759.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.