What do you mean?
Eliciting enthymemes in text-based dialogue

Ebba Axelsson Nord, Vladislav Maraev, Ellen Breitholtz and Christine Howes

Department of Philosophy, Linguistics and Theory of Science

University of Gothenburg

ebbaxelsson@hotmail.com; vladislav.maraev@gu.se;
ellen.breitholtz@ling.gu.se; christine.howes@gu.se

Abstract

We report a proof of concept text-based chat study which inserts spoof questions in a real time conversation in order to elicit participants reasoning. The results show that different questions are more or less effective at eliciting explicit enthymematic reasons, which is a technique that could be employed to augment conversational artificial intelligence systems to improve their reasoning abilities.

1 Introduction

Reasoning is crucial for humans and for dialogue agents, and much research has been dedicated to enabling computers to reason from premises to conclusions. Nevertheless, how people interactively reason in natural dialogue is still poorly understood, since much reasoning in human dialogue is enthymematic, i e it relies on non-logical common sense principles of reasoning (Breitholtz, 2020). No existing artificial intelligence system is able to make use of this type of reasoning, which people find so natural.

Besides having little knowledge about reasoning in natural dialogue for building better dialogue agents, we also lack adequate data resources, especially given that important reasoning often takes place "behind the scenes", and can't be extracted directly from dialogue transcripts. This work can be viewed as a step towards collecting data for dialogue agents which are capable of reasoning.

Specifically we present a proof of concept chat tool study to investigate whether people can be prompted to provide their reasoning in an unobtrusive way. As noted by (Schlöder et al., 2016), one way of probing enthymematic reasoning is through the use of questions like "why?" or "what do you mean?".

2 Method

2.1 DiET chat tool

The Dialogue Experimental Toolkit (DiET) chat tool (Healey et al., 2003) is a text-based chat interface into which interventions, such as adding fake turns, can be introduced into a dialogue in real time, thus causing a minimum of disruption to the 'flow' of the conversation. For this experiment we used the new mobile version of DiET, which runs through the messenger app Telegram.[1]

2.2 Task

The subjects discussed the balloon task – a moral dilemma known to elicit dialogues containing extended reasoning sequences. Participants are instructed to reach agreement on which of four passengers should be thrown out of a hot air balloon that will otherwise crash, killing all the passengers, if one is not sacrificed. The four passengers are:

Mr Tom Harris – the balloon pilot who is the only passenger with any balloon flying experience
Mrs Susie Harris – Tom's wife, a primary school teacher who is 7 months pregnant with their second child
Dr Robert Lewis – a cancer research scientist, who believes he is on the brink of discovering a cure for most common types of cancer
Miss Heather Sloan – a 14 year old musical prodigy who is considered to be a "twenty-first century Mozart"

2.3 Procedure

The 32 participants, from the student population at the University of Gothenburg, were instructed via zoom on how to access the experiment using the Telegram app. Once logged in, they were told to discuss the task until they got a message from the server to stop, in order to ensure sufficient turns. The manipulation consisted of introducing 'spoof' turns into the dialogue, which appeared to the recipient to have originated from their dialogue partner.

---

[1] https://dialoguetoolkit.github.io/chattool/

The spoof turns, consisting of questions such as 'why?' and 'what do you mean?' (see table 1) were generated pseudo randomly from a limited list of probes by the DiET-server and triggered by particular words or phrases in the preceding turn (e.g., "the doctor", "the pregnant woman").

2.4 Annotation

The dialogues were examined to establish whether or not the spoof question was responded to and if so, whether the response was a direct response (1), an indirect response (if there was an intervening contribution before the response, as in (2)) or a clarification request (3).

(1) **1:** I bet the father would not want the child nor their partner to die
     2: why?   *[artificial turn]*
     1: Paternal instincts and all that

(2) **1:** But is it morally acceptable to throw out the girl if the pilot is needed?
     2: what do you mean?   *[artificial turn]*
     2: The child you mean?
     1: yeah

(3) **2:** Doctor can have botanical knowledge or whatever
     1: why?   *[artificial turn]*
     2: Why which part

3 Results and discussion

With the exception of one pair, who were discarded from further analysis, the debriefing showed that none of the subjects were aware of any experimental manipulations. Each pair was exposed to an average of 6.4 spoof questions generated by the server.

Spoof question	NR	DR	IR	CR	Total
???	4	6	1	0	11
how so?	4	6	3	0	13
what?	4	6	4	0	14
what do you mean?	5	16	14	2	37
why?	4	14	6	3	27
Total	21	48	28	5	102

Table 1: Type of response given by probe. NR/DR/IR – no/direct/indirect response. CR – clarification request.

A greater proportion of "???", "how so?" and "what?" questions received no response compared to "why?" and "what do you mean?" (32% vs 14% $\chi^2_1 = 4.47; p = 0.03$). One possible explanation for this finding is that in the former cases the participants may think that the ongoing dialogue shows that the question is already resolved so they do not feel obliged to answer. This could also be affected by the text based medium in which participants can type simultaneously with turns both interleaved and persistent (Healey et al., 2018).

Qualitative analysis of the data by the first author suggests that "why" and "how so" spoof questions are more capable of addressing reasoning (as in (1)) with the other three cues more open ended and available to interpretations ranging from semantic or orthographic ambiguities to reasoning. Nevertheless, each of the probes, even the more open ended ones, did produce some responses which, combined with a previous utterance, constitute enthymemes, showing that the interpretation of a question is not fixed. An example of this can be seen in (4).

(4) **1:** I think pregnant women are not supposed to fly actually
     2: what?   *[artificial turn]*
     1: There are safety regulations at least during the 7th month of pregnancy

This study shows that our approach is a useful method for eliciting enthymemes to collect resources for common sense reasoning in spoken dialogue systems. This is useful in task-oriented domains to argue about the decisions taken by the system, as well as in chit-chat dialogues – especially ones which are concerned with controversial topics or current issues, such as the climate crisis.

In the process of dialogue systems development enthymeme eliciatation can be a part of data collection based on a *distilling dialogue* process (Larsson et al., 2000; Jönsson and Dahlbäck, 2000). After enthymeme elicitation, we plan to take the following steps to collect enthymematic resources for a dialogue system:

1. dependency parsing and pattern-based extraction of enthymeme candidates based on their surface structure

2. annotation, whether or not the extracted structure is an enthymeme and annotation of the premise(s) and the consequence(s) of it.

3. enthymeme classification (for example, keywords like 'since' can relate to a time frame)

4. enthymeme parsing, that will lead to a semantic representation of an enthymeme

Extracted enthymemes can then be clustered to induce more general principles of reasoning, such as the Aristotelian topos of 'the more and the less'. The gist of this topos is that a small thing is contained in a large thing – for example, if you can build a castle you can build a cottage, or if you can run a marathon then you can run a half marathon.

Acknowledgments

References

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Brill, Leiden, The Netherlands.

Patrick G. T. Healey, Gregory Mills, Arash Eshghi, and Christine Howes. 2018. Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science*, 10(2):367–388.

Patrick G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Gregory Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, Massachusetts.

Arne Jönsson and Nils Dahlbäck. 2000. Distilling dialogues-a method using natural dialogue corpora for dialogue systems development. In *6th Applied Natural Language Processing Conference, Seattle, Washington, 2000*, pages 44–51. Association for Computational Linguistics Stroudsburg.

Staffan Larsson, Lena Santamarta, and Arne Jönsson. 2000. Using the process of distilling dialogues to understand dialogue systems. In *Proceedings of ICSLP 2000*, pages 374–377. Citeseer.

Julian Schlöder, Ellen Breitholtz, and Raquel Fernández. 2016. Why? In *Proceedings of JerSem*, pages 5–14.