Occam's Razor and Bender and Koller's Octopus

Michael Guerzhoy
University of Toronto
guerzhoy@cs.toronto.edu

Abstract

We discuss the teaching of the discussion surrounding Bender and Koller's prominent ACL 2020 paper, "Climbing toward NLU: on meaning form, and understanding in the age of data" (Bender and Koller, 2020).

We present what we understand to be the main contentions of the paper, and then recommend that the students engage with the natural counter-arguments to the claims in the paper.

We attach teaching materials that we use to facilitate teaching this topic to undergraduate students.

1 Introduction

The claim in Bender and Koller's argument in (Bender and Koller, 2020) is that a being that only has access to the form of the communication – e.g., an intelligent octopus that taps into only the submarine signals that encode accounts of the events above the sea that two people on land send each other – will not be able to "understand" what is happening above sea-level, lacking the semantics of the Morse code that was used to communicate the events transpiring above the seas. Koller and Bender argue that even if the octopus can send messages based on the patterns it sees that would be understood by the humans and the humans would be fooled into thinking they are reading messages from another human, the shallow understanding of the octopus would necessarily be revealed when trying to pretend to answer more complicated queries.

Implicit in the argument is that the intelligent octopus is analogous to a Large Language Model (LLM), akin to GPT-2 (Brown et al., 2020), GPT-4 [1], Claude 3 [2], or Meta Llama 3 [3], and that such LLMs would not be able to truly understand natural language in the same way that Bender and Koller's octopus will not.

In this paper, we present a lecture + activity that challenges Bender and Koller's argument. Students will engage with Bender and Koller's argument and with the counterargument, and come away with their own conclusions

2 Building theory form data

The scientific process itself can be analogized to a B&K octopus observing data they don't understand.

For example, astronomers observe and try to predict the motions of heavenly bodies, initially with no mechanistic understanding of why the stars appear to move the way they do. Historically, astronomers came up with multiple incorrect theories for why the heavenly bodies move the way they do (notably, the family of geocentric models). Astronomers used "epicycles" as a way to align predictions with their model, at the expense of parsimony (Duhem, 2015).

Historically, Copernicus' first models did not predict the data as well as the best epicycle-based geocentric models (Klein and Loftus, 2015).

Note that, unlike the astronomers, the octopus cannot *interact* with the world – he cannot influence what observations are made (at least before he starts communicating with the astronomers). This can influence how fast the octopus can "converge." Historically, much of the data used by Kepler was previously collected by Tycho Brahe.

2.1 Occam's razor

Occam's razor – the principle that, all things being equal, we should prefer the simpler theory – can help select the better scientific theory. For example, the B& K octopus might consider all possible theories of the world over the sea, and settle on the simplest one that explains the communications the

---

[1] https://openai.com/index/gpt-4-research/
[2] https://www.anthropic.com/news/claude-3-family
[3] https://llama.meta.com/llama3/

octopus decodes.

3 Can the B& K octopus learn science

Every student can make their own conclusions, but ours is that it's not in principle impossible for that to happen (or if it is impossible, we don't have a clear reason to think so). The success of LLMs on tasks that require some level of world-theory-building such as the addition of integers task (Lee et al., 2023), predicted to be impossible by Bender and Koller (2020) (see Appendix B), indicates that if there are barriers to learning world models from observational data, they are not well-understood. Our view is that the prediction by B&K that a pure LLM could not learn to do arithmetic is due to insufficiently accounting for the possibility of using inductive biases to build a model of the data that corresponds to the world that the data is describing.

4 Materials

We provide slides we used in class to follow up the class's reading of Bender and Koller (2020). We also provide the following guiding questions

1. If the octopus observes different content in messages when it's dark vs. when it's light, what can the octopus possibly conclude about language?

2. Describe how the octopus might use tides to infer words that have to do with tides

3. Describe how the octopus might decode conversations about physics based on the conversations about tides – perhaps building up from observations of tides, stars, etc.

4. If you assume no "cheating" such as jointly observing tides, might you imagine conversations that involve physical and mathematical constant like G and π playing a similar role?

5. Explain why without Occam's razor, the Octopus will have a practically infinite number of theories about what the two humans could be talking about

6. What might be some insurmountable challenges for the octopus in the quest to understand the meaning of the cable signals? How

7. Consider the claim from the original paper that arithmetic is not learnable by form alone: where might that argument have gone wrong?

5 Additional materials

Julian Michael, *To Dissect An Octopus* `https://julianmichael.org/blog/2020/07/23/to-dissect-an-octopus.html` provides an excellent overview.

6 Conclusion

Many students in NLP would be familiar with Bender and Koller's argument, but have probably not engaged in the critical analysis of the arguments. We provide materials for critically analyzing the arguments made by Bender and Koller. We focus on the counterarguments since the argument itself is ably presented by the original authors. We provide slides introducing the B&K argument to the best our ability as well.

Many (though not all) students are captivated by the debate. We find that the structure provided by the guiding questions helps in our lectures.

7 Teaching materials

Slides: `https://github.com/guerzh/octopus`
Video lecture: `https://youtu.be/6QVjGF_J7I0`

References

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pierre Duhem. 2015. *To save the phenomena: An essay on the idea of physical theory from Plato to Galileo*. University of Chicago Press.

Stanley B Klein and Judith Loftus. 2015. The mental representation of trait and autobiographical knowledge about the self. *The mental representation of trait and autobiographical knowledge about the self*, pages 1–49.

Nayoung Lee, Kartik Sreenivasan, Jason D Lee, Kangwook Lee, and Dimitris Papailiopoulos. 2023. Teaching arithmetic to small transformers. In *The Twelfth International Conference on Learning Representations*.