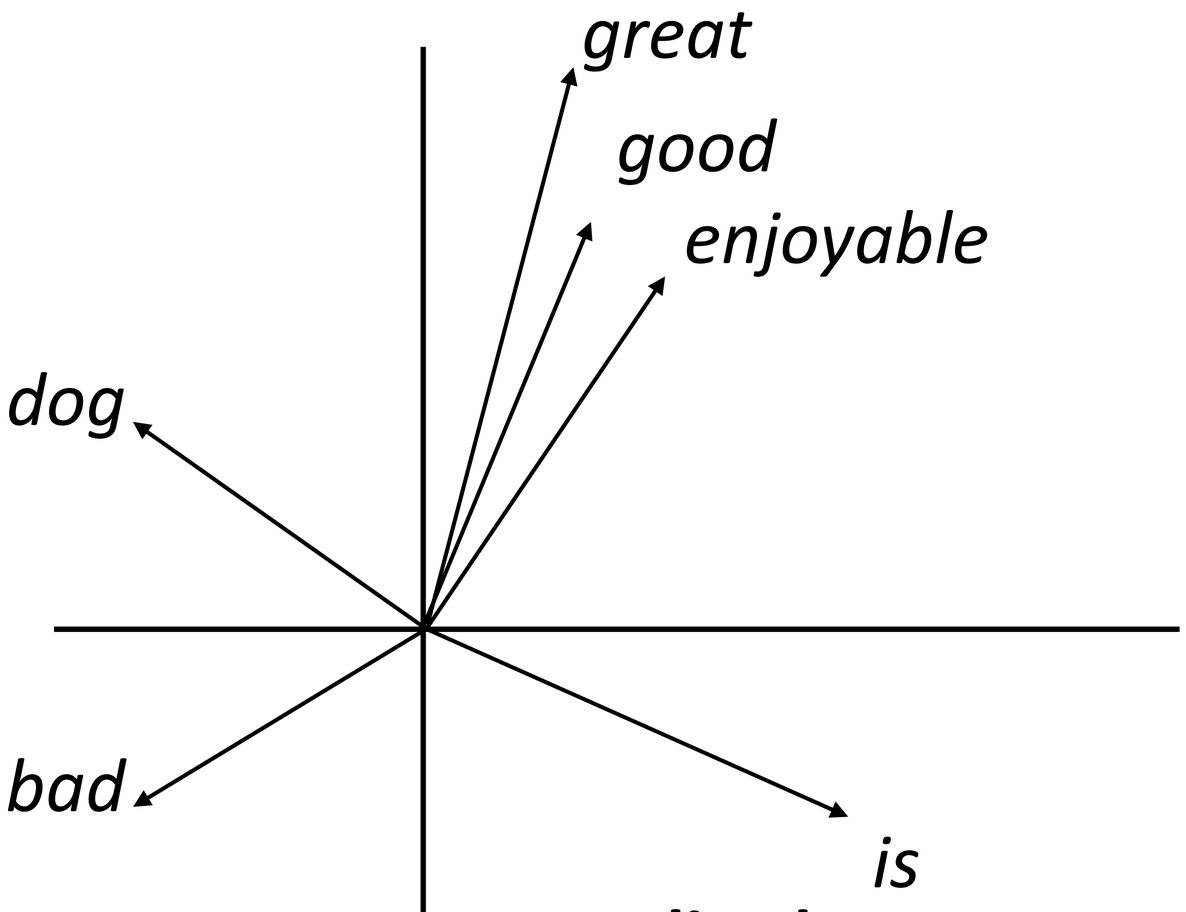
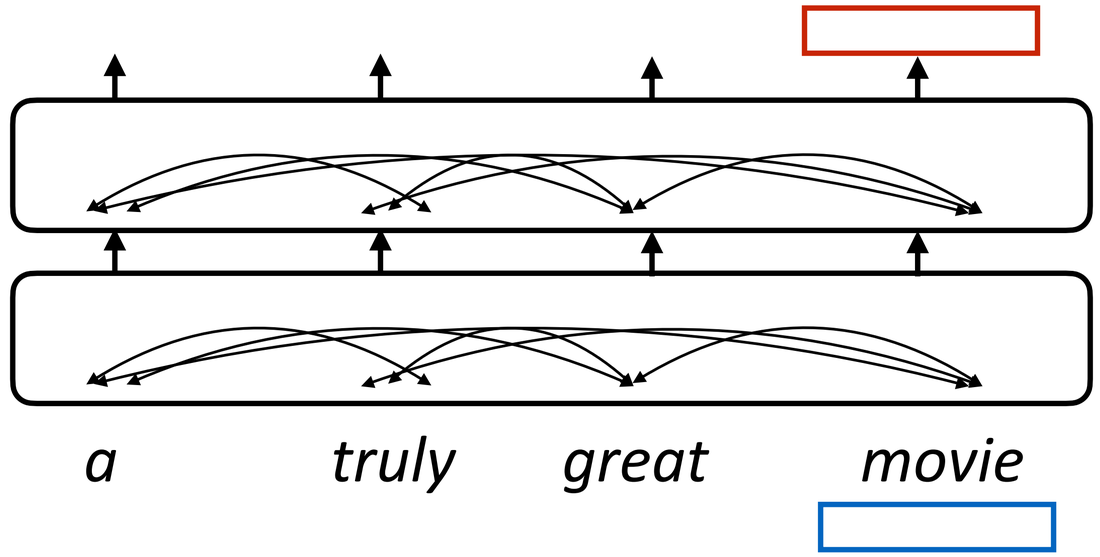
Language Grounding

- How do we represent language in our models?
- How did we learn these representations? What do the vectors “mean”?



non-contextualized



contextualized

Language Grounding

- Harnad defines a “symbol system”: we have symbols (e.g., strings) manipulated on the basis of rules, and these symbols ultimately have “semantic interpretation”
  - “Fodor (1980) and Pylyshyn (1980, 1984)…emphasize that the symbolic level (for them, the mental level) is a natural functional level of its own, with ruleful regularities that are independent of their specific physical realizations”
- Harnad challenges the idea that fully symbolic approaches can work well.
  - Argues that “horse” is something that should be understood bottom-up through grounding. “Zebra” = “horse” + “stripes” could emerge this way, but he claims it cannot through a top-down symbolic system
  - What does it mean to “understand” the symbols that get manipulated?

Harnad (1990) *The Symbol Grounding Problem*

Searle's Chinese Room

- Suppose we have someone in a room with a long list of rules, dictionaries, etc. for how to translate Chinese into English. A Chinese string is passed into the room and an English string comes out. The person is not a speaker of Chinese, but merely follows the rules and looks things up in the dictionaries to produce the translation.

- Does the person understand Chinese? Does the room? (the "system"?)

- Searle argues that (a) the room is like an AI system producing Chinese translations; (b) the operator in the room (the AI) does not "understand" Chinese. Harnad summarizes:

> *The interpretation will not be intrinsic to the symbol system itself: It will be parasitic on the fact that the symbols have meaning for us, in exactly the same way that the meanings of the symbols in a book are not intrinsic, but derive from the meanings in our heads.*

Searle (1980)

Language Grounding

- Bender and Koller separate form and meaning. Meaning = communicative intent. The role of the speaker/listener are crucial in language, LMs lack the underlying intent

- They propose the “octopus” experiment to show how form alone can fail.
  An octopus is eavesdropping on a conversation between A and B (using deep-sea communication cables). Suddenly, the octopus decides to cut the cable and impersonate B.

- A has an emergency and asks how to construct something with sticks to fend off a bear. The octopus can’t help because it can’t simulate this novel situation.

Bender and Koller (2020)
Climbing towards NLU

Counterarguments

- We can't necessarily learn semantics from predicting next characters alone without execution. Consider training on:

```
x = 2
y = x + 2
print(y)
```

- **However**, assertion statements are sufficient to teach us some semantics! (but this can still break down)

```
x = 2
y = x + 2
assert(y == 4)
```

- For language: similar argument. Assume people say true things. Consider saying a pair of sentences x_1, x_2; given enough examples, the fact that x_2 should not be contradicted by x_1 tells us something

Merrill et al. (2021) *Provable Limitations of Acquiring Meaning from Ungrounded Form*

Merrill et al. (2022) *Entailment Semantics can be Extracted from an Ideal Language Model*

Where are we?

- “Experience Grounds Language” (Bisk et al., 2020): Five levels of “world scope”: corpus, Internet, perception, embodiment, social

pure LM < LM fine-tuned on supervised data < vision+language LM < vision+language+manipulation LM < …

GPT-4 is here!

PaLM-E

- Unclear how quickly we’ll continue to climb this hierarchy: embodied/ social data is very hard to collect at scale!