Ling/CSE 472: Introduction to Computational Linguistics

4/27: Neural language models

Overview

- Neural nets and language processing: Some history
- XOR and “representations”
- Reading questions

Some history

- McCulloch-Pitts neuron: 1940s
- Perceptron: Rosenblatt 1958
- Single perceptron can't do XOR: Minsky & Paper 1969
- Error backpropagation: Rumelhart et al 1986
- 1980s: Neural nets as models of human cognition
- 1990s: early NLP applications (handwriting recognition: LeCun et al 1989, 1990; ASR: Morgan and Bourlard 1989, 1990)
- 2000s: "deep learning" (Hinton et al 2006, Bengio et al 2007)

A basic neural unit

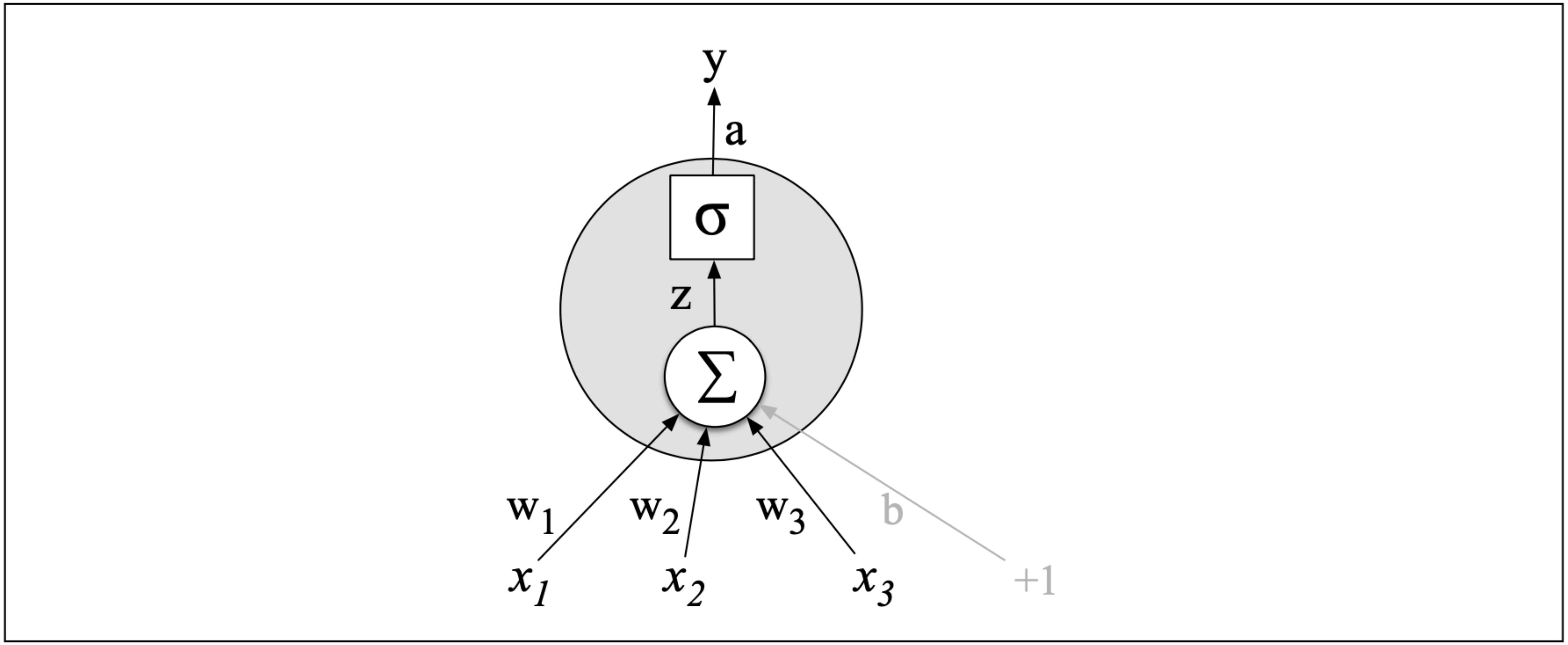
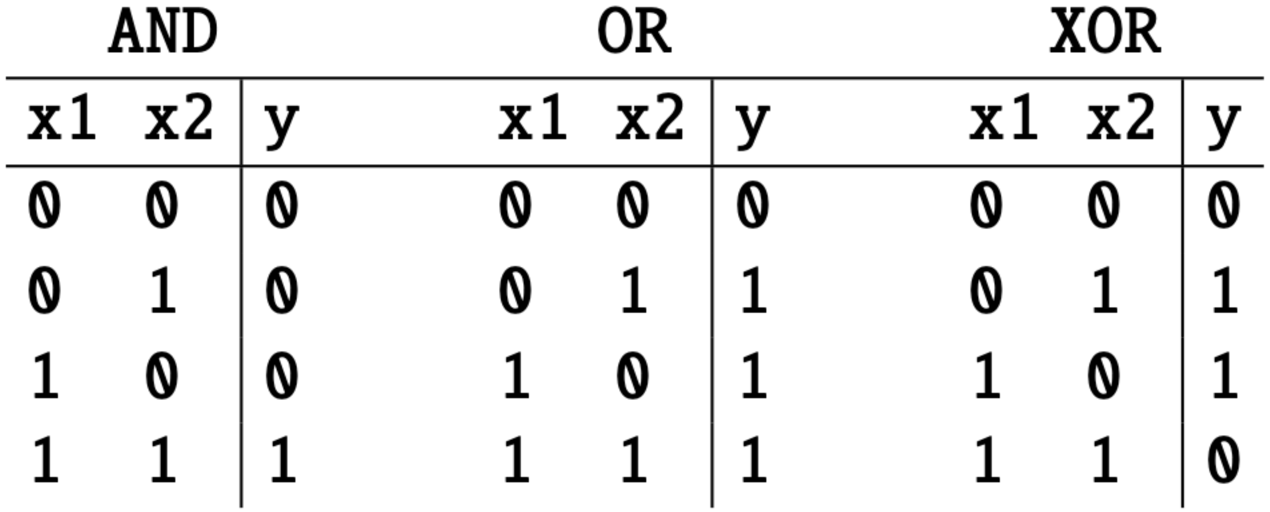


Figure 7.2 A neural unit, taking 3 inputs x_1, x_2, and x_3 (and a bias b that we represent as a weight for an input clamped at +1) and producing an output y. We include some convenient intermediate variables: the output of the summation, z, and the output of the sigmoid, a. In this case the output of the unit y is the same as a, but in deeper networks we'll reserve y to mean the final output of the entire network, leaving a as the activation of an individual node.

The XOR problem, with a single perceptron



AND			OR			XOR		
x1	x2	y	x1	x2	y	x1	x2	y
0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	1	1
1	0	0	1	0	1	1	0	1
1	1	1	1	1	1	1	1	0

$$y = \begin{cases} 0, & \text{if } w \cdot x + b \leq 0 \\ 1, & \text{if } w \cdot x + b > 0 \end{cases}$$

The XOR problem, with a single perceptron

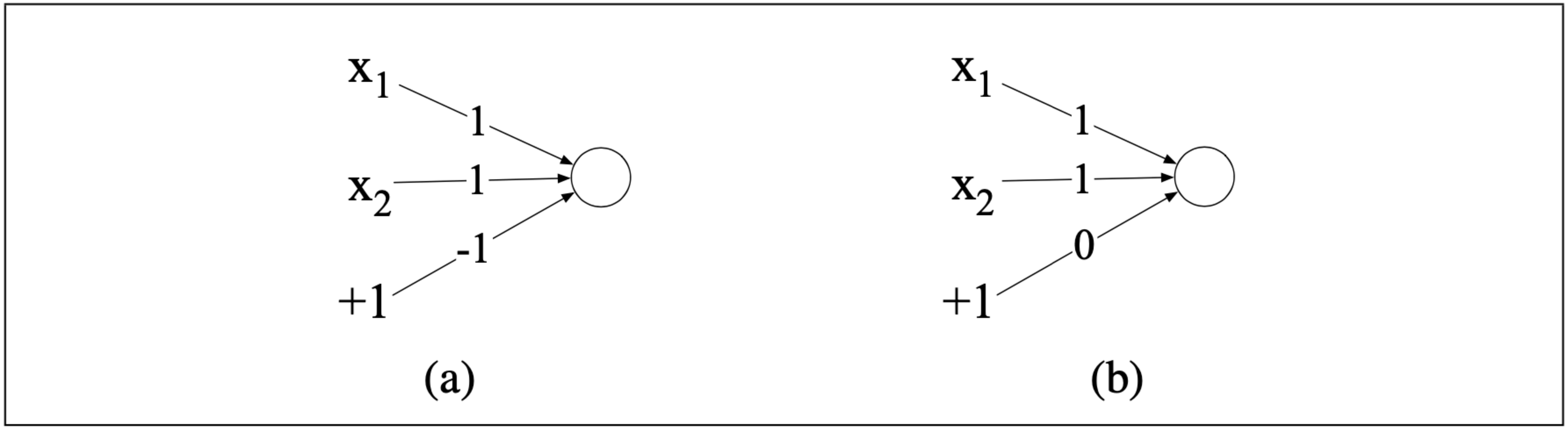
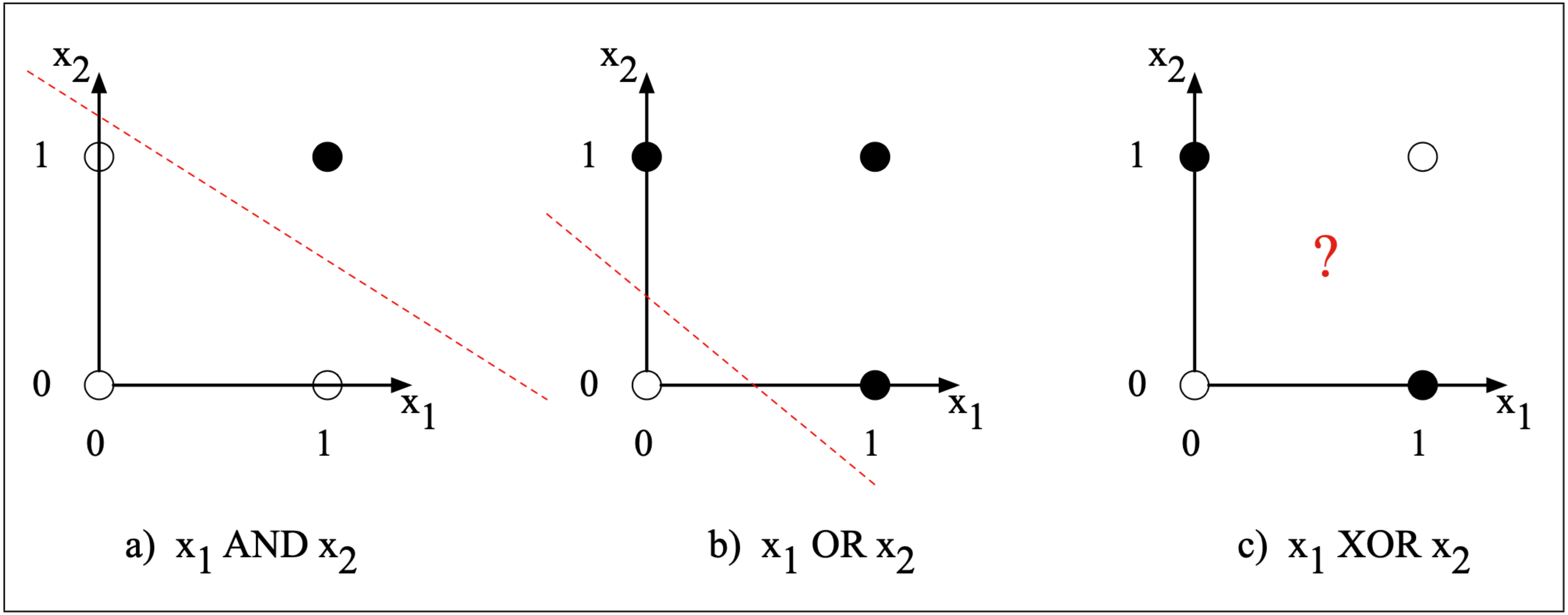


Figure 7.4 The weights w and bias b for perceptrons for computing logical functions. The inputs are shown as x_1 and x_2 and the bias as a special node with value $+1$ which is multiplied with the bias weight b. (a) logical AND, showing weights $w_1 = 1$ and $w_2 = 1$ and bias weight $b = -1$. (b) logical OR, showing weights $w_1 = 1$ and $w_2 = 1$ and bias weight $b = 0$. These weights/biases are just one from an infinite number of possible sets of weights and biases that would implement the functions.

The XOR problem, with a single perceptron



a) x_1 AND x_2

b) x_1 OR x_2

c) x_1 XOR x_2

Figure 7.5 The functions AND, OR, and XOR, represented with input x_1 on the x-axis and input x_2 on the y axis. Filled circles represent perceptron outputs of 1, and white circles perceptron outputs of 0. There is no way to draw a line that correctly separates the two categories for XOR. Figure styled after Russell and Norvig (2002).

XOR solution, with a hidden layer and ReLU units

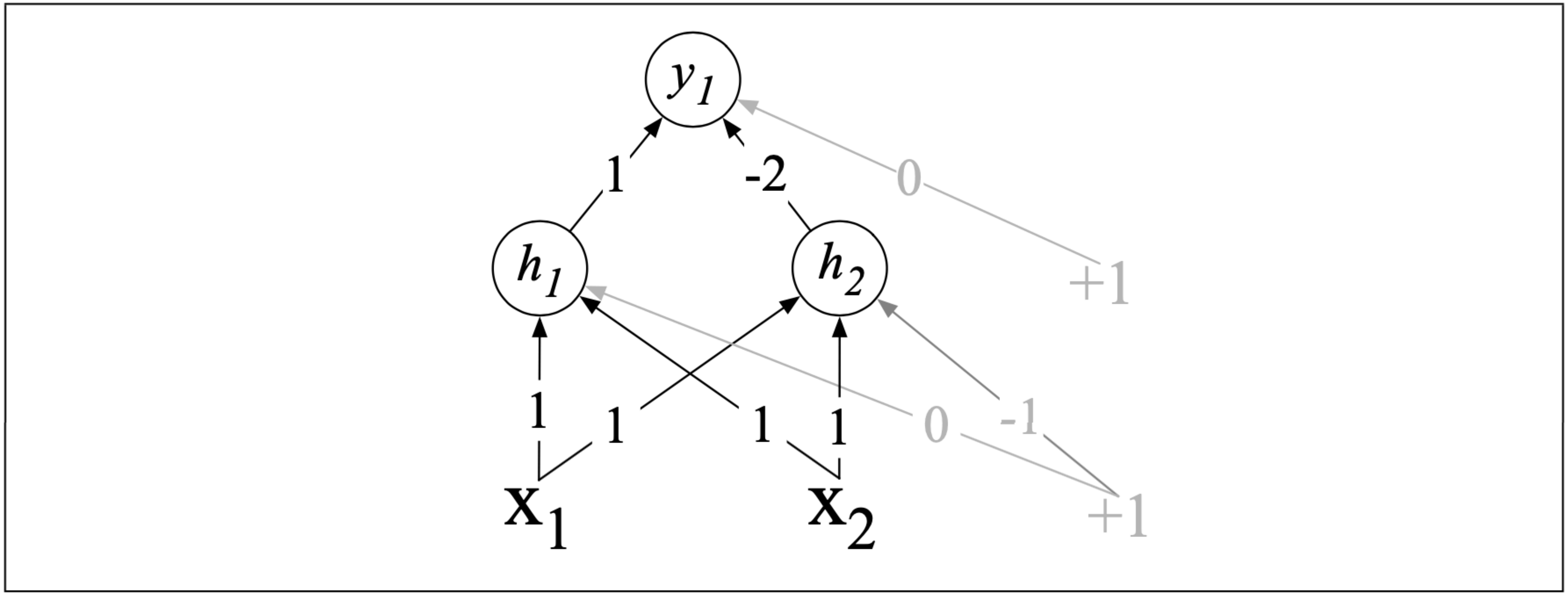
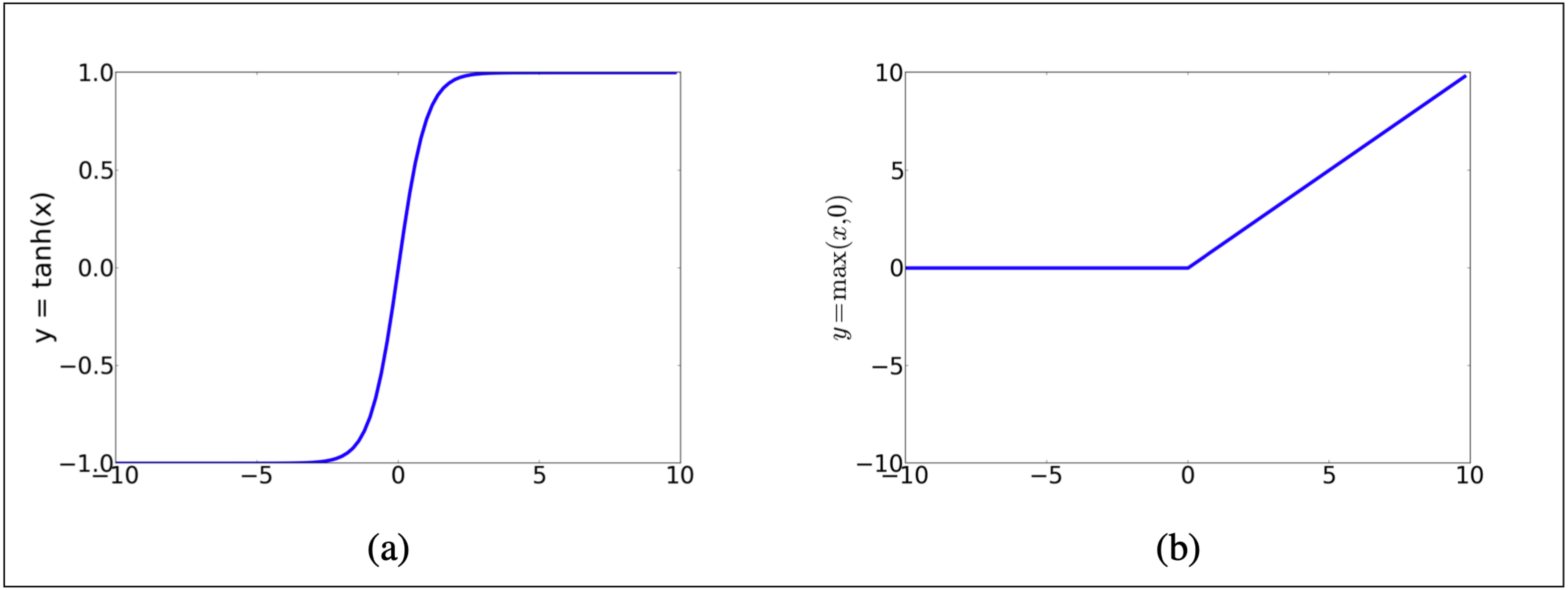


Figure 7.6 XOR solution after Goodfellow et al. (2016). There are three ReLU units, in two layers; we've called them h_1, h_2 (h for "hidden layer") and y_1. As before, the numbers on the arrows represent the weights w for each unit, and we represent the bias b as a weight on a unit clamped to +1, with the bias weights/units in gray.

tanh and ReLU activation functions



(a)

(b)

Figure 7.3 The tanh and ReLU activation functions.

Simple neural LM (J&M, Ch 7, p 16)

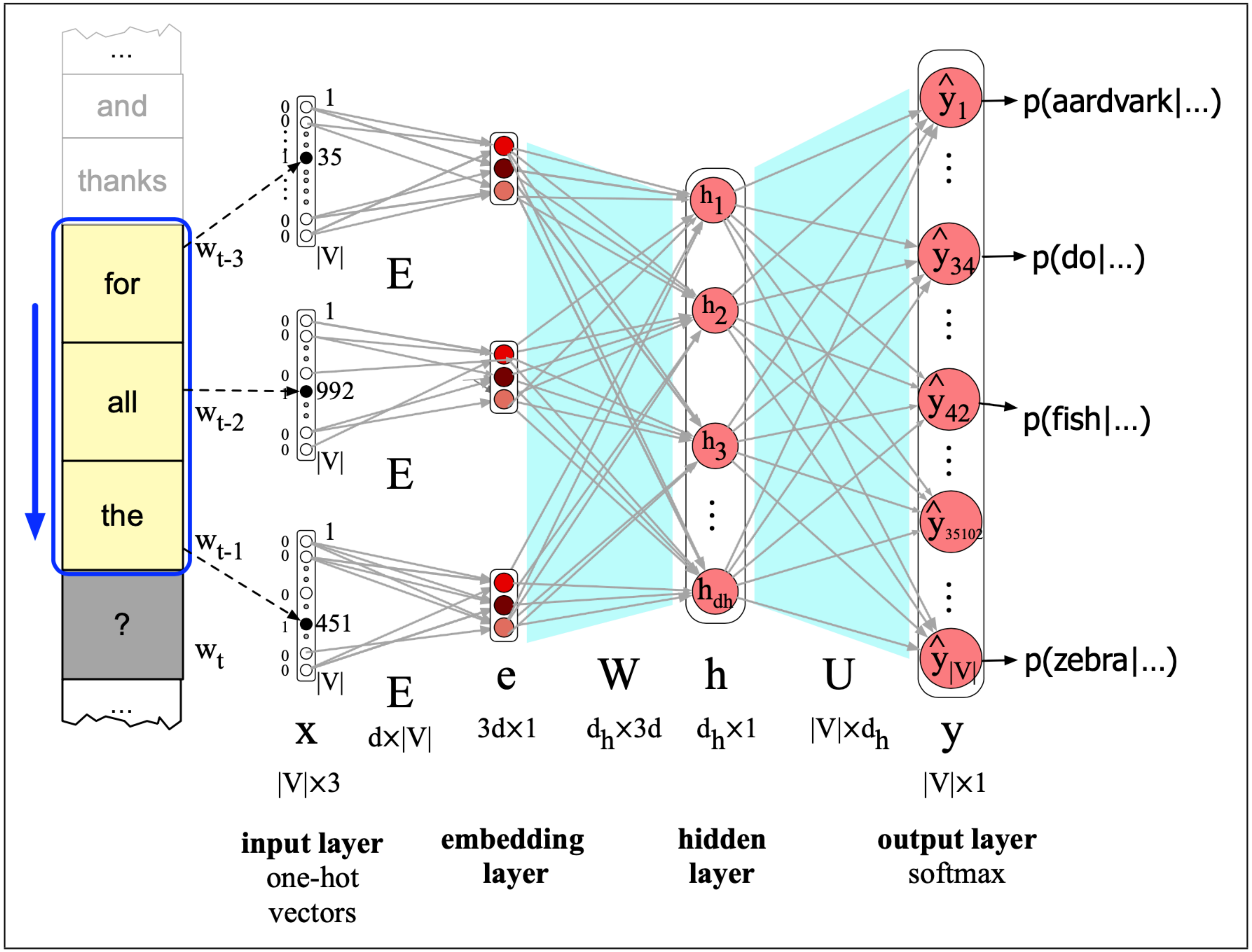


Figure 7.13 Forward inference in a feedforward neural language model. At each timestep

Back-propagation

- At each training step, compare output to “correct” answer
- Compute loss function
- Update weights & biases to reduce loss
- But the answer is only directly related to the last layer
    - => How do distribute the loss to the earlier layers?

Neural nets and linguistics

- What can word embeddings tell us about semantics?
- About syntax?
- Why do people think neural language models “understand”?
- What might be the evidence that they don’t?
    - => The Octopus paper

Reading questions

- The most confusing part of this reading was just how much new vocabulary was present in the reading and the diagrams could be kind of confusing with all of the arrows and words present. The way that the reading talks about feedforward neural language modeling sounds very much like how Dr. Bender has spoken about how a chatbot like ChatGPT generates responses. Is this accurate?

Reading questions

- The reading mentioned all of the ways that neural language models have many benefits over n-gram language models (which are very apparent) but is/are there still use(s) for n-gram models? I would imagine that n-gram language models might be easier to implement and take much less computing power.

- Section 7.5 says neural models are slower and take more energy to train than n-gram models; how big of a difference is this? I always imagine computing as being super fast these days, even for complex processes. How long might it take to train and execute a typical neural model?

Reading questions

- How do you embed words with things like their meanings, parts-of-speech, and their relationship to other words in an algorithm?

Reading questions

- The reading mentions that one way in which neural language models differ from n-gram models is that they represent words by their embeddings rather than just their word identity. What exactly constitutes a word embedding?

- I would like some more clarification on what exactly constitutes an embedding. How are embeddings 'decided' for words and how are they tweaked over time to more accurately predict following words or phrases?

Reading questions

- Is there much of a difference in the results of using pre-trained word embeddings or learning the embeddings while training the neural network? For instance, are there cases where pre-trained embeddings won't work well for the specific language task the neural LM is being used for?

Reading questions

- From what the chapter explained about neural networks and how they work to guess future words based on words that showed up previously, is all of this in the NLP setting just the system guessing what the next words will be and hoping for the best? I struggle to imagine how this system could be used in translation if that's the case then.

Reading questions

- As far as I can tell, language models like these output probabilities for every word they know based on the word history. For text prediction, it would make sense to choose the word with the highest probability. Are the probabilities ever used in any other way for different tasks? Or is selecting the word/phrase/etc. with the highest probability always the best way to go for NLP?

Reading questions

- As described in the paper, BERT is a discriminative model which tries to understand a piece of text. The two pre-training gave BERT something we could utilize. I am wondering will a trained model like BERT which can predict whether or not the next sentence is the "real" next sentence be used to train a generative model like chatGPT as a evaluation step?

Reading questions

- Transformers have been one of the most important and revolutionary advancements in machine learning, especially in the context of NLP. What aspects of the more basic feedforward perceptron model are carried over to transformer models and what advantages do these alterations provide?

- What are some active areas of research in further evolving neural language model architectures?

Climbing towards NLU:
On Meaning, Form, and Understanding
in the Age of Data

Emily M. Bender, University of Washington
Alexander Koller, Saarland University

ACL 2020

This position paper talk in a nutshell

- Human-analogous natural language understanding (NLU) is a grand challenge of AI

- While large neural language models (LMs) are undoubtedly useful, they are not nearly-there solutions to this grand challenge

  - Despite how they are advertised

- Any system trained only on linguistic form cannot in principle learn meaning

- Genuine progress in our field depends on maintaining clarity around big picture notions such as *meaning* and *understanding* in task design and reporting of experimental results.

What is meaning?

- Competent speakers easily conflate ‘form’ and ‘meaning’ because we can only rarely perceive one without the other

- As language scientists & technologists, it’s critical that we take a closer look

Working definitions

- **Form** : marks on a page, pixels or bytes, movements of the articulators
- **Meaning** : relationship between linguistic form and something external to language
  - $M \subseteq E \times I$: pairs of expressions and communicative intents
  - $C \subseteq E \times S$: pairs of expressions and their standing meanings
- **Understanding** : given an expression *e*, in a context, recover the communicative intent *i*

BERT fanclub

- “In order to train a model that understands sentence relationships, we pre-train for a binarized next sentence prediction task that can be trivially generated from any monolingual corpus.” (Devlin et al 2019)

- “Using BERT, a pretraining language model, has been successful for single-turn machine comprehension …” (Ohsugi et al 2019)

- “The surprisingly strong ability of these models to recall factual knowledge without any fine-tuning demonstrates their potential as unsupervised open-domain QA systems.” (Petroni et al 2019)

BERT fanclub

- “In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task that can be trivially generated from any monolingual corpus.” (Devlin et al 2019)
- “Using BERT, a pretraining language model, has been successful for single-turn machine **comprehension** …” (Ohsugi et al 2019)
- “The surprisingly strong ability of these models to **recall factual knowledge** without any fine-tuning demonstrates their potential as unsupervised open-domain QA systems.” (Petroni et al 2019)

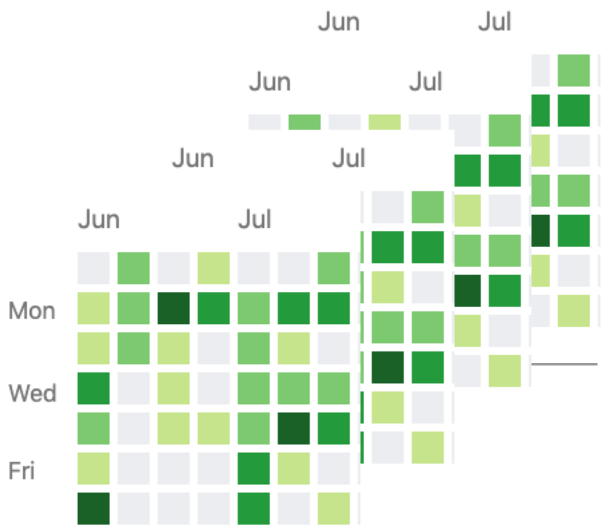
BERTology

- Strand 1: What are BERT and similar learning about language structure?
  - Distributional similarities between words (Lin et al 2015, Mikolov et al 2013)
  - Something analogous to dependency structure (Tenney et al 2019, Hewitt & Manning 2019)
- Strand 2: What information are the Transformers using to ‘beat’ the tasks?
  - Niven & Kao (2019): in ARCT, BERT is exploiting spurious artifacts
  - McCoy et al (2019): in NLI, BERT leans on lexical, subsequence, & constituent overlap heuristics
- Our contribution: Theoretical perspective on why models exposed only to form can never learn meaning

So how do babies learn language?

- Interaction is key: Exposure to a language via TV or radio alone is not sufficient (Snow et al 1976, Kuhl 2007)
- Interaction allows for joint attention: where child and caregiver are attending to the same thing and mutually aware of this fact (Baldwin 1995)
- Experimental evidence shows that more successful joint attention leads to faster vocabulary acquisition (Tomasello & Farrar 1986, Baldwin 1995, Brooks & Meltzoff 2005)
- Meaning isn’t in form; rather, languages are rich, dense ways of providing cues to communicative intent (Reddy 1979). Once we learn the systems, we can use them in the absence of co-situatedness.

Thought Experiment: Java



- Model: Any model type at all
  - For current purposes: BERT (Devlin et al 2019), GPT-2 (Radford et al 2019), or similar
- Training data: All well-formed Java code on GitHub
  - but only the text of the code; no output; no understanding of what unit tests mean
- Test input: A single Java program, possibly even from the training data
- Expected output: Result of executing that program

That’s not fair!

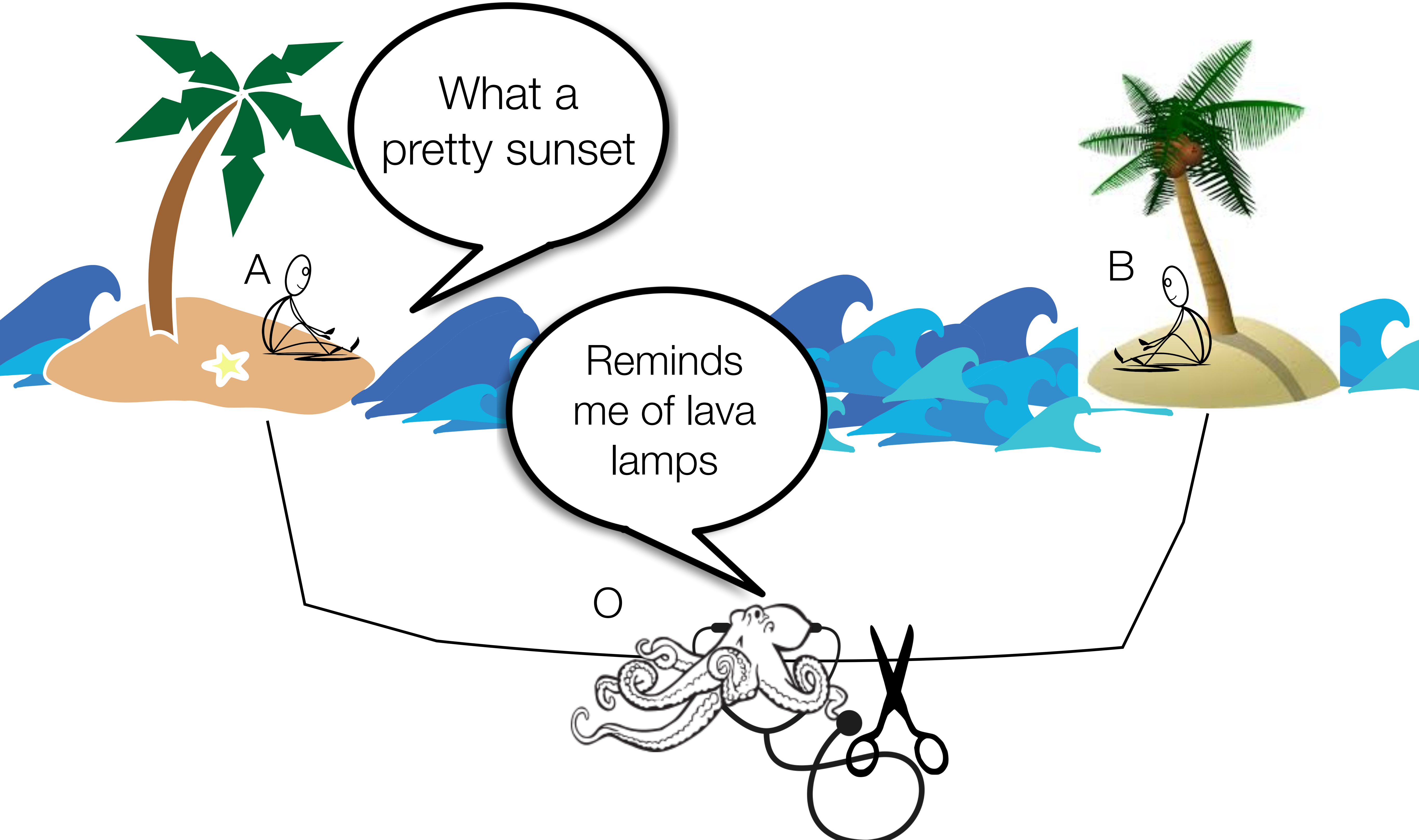
- Of course not! What’s interesting about this thought experiment is what makes the test unfair

- It’s unfair because the training data is insufficient for the task

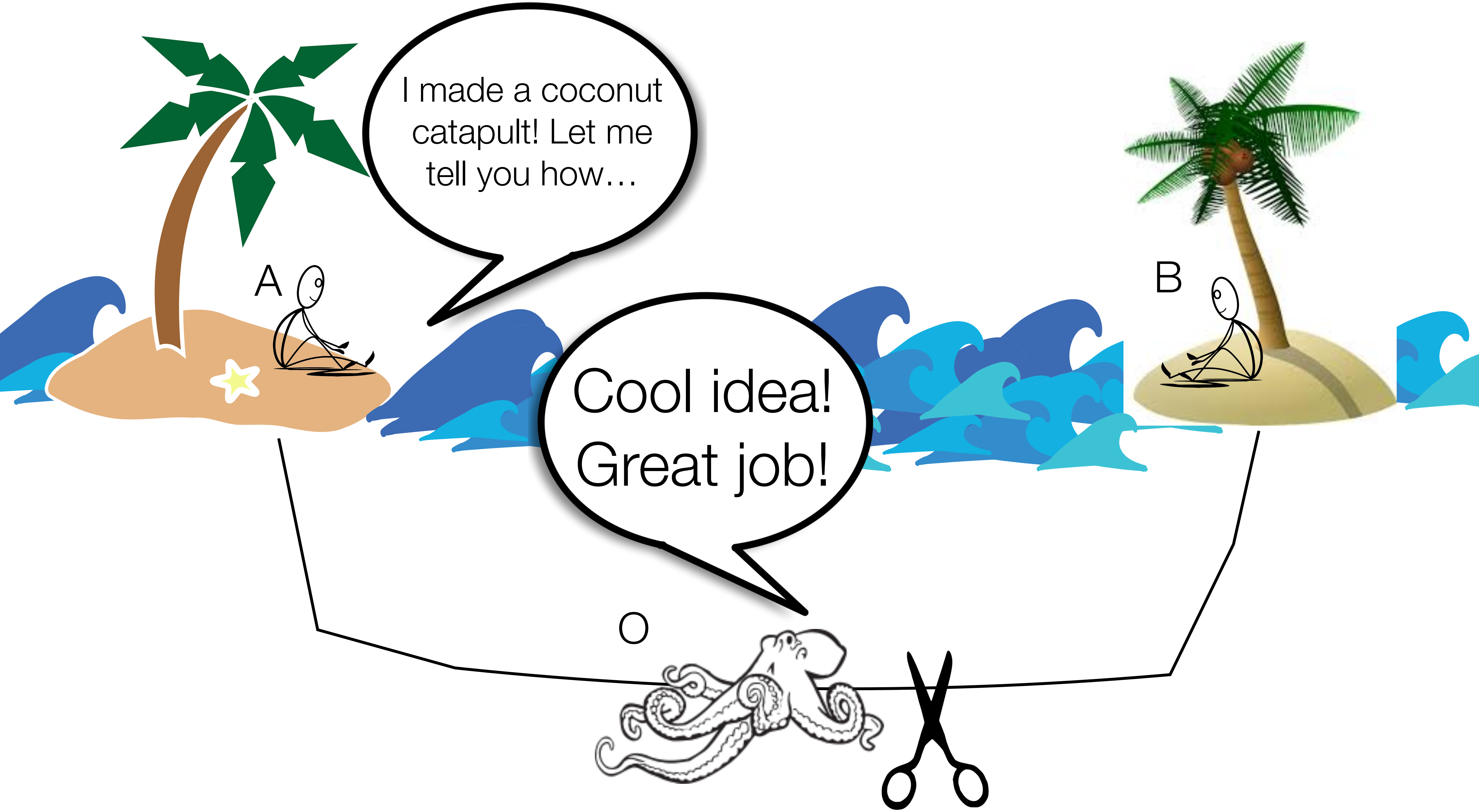
- What’s missing: Meaning — in the case of Java, what the machine is supposed to do, given the code

- What would happen with a more intelligent and motivated learner?

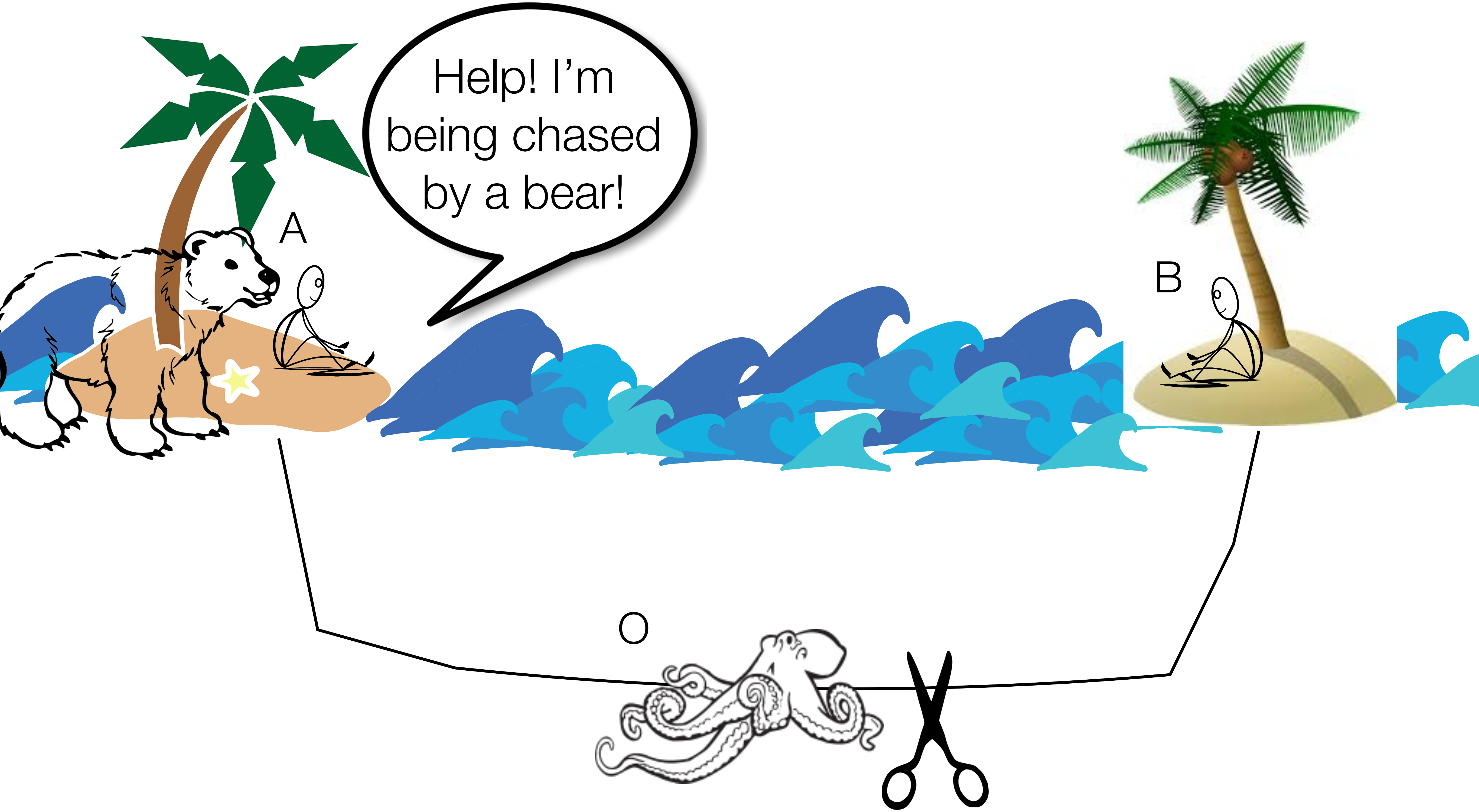
Thought experiment: Meaning from form alone



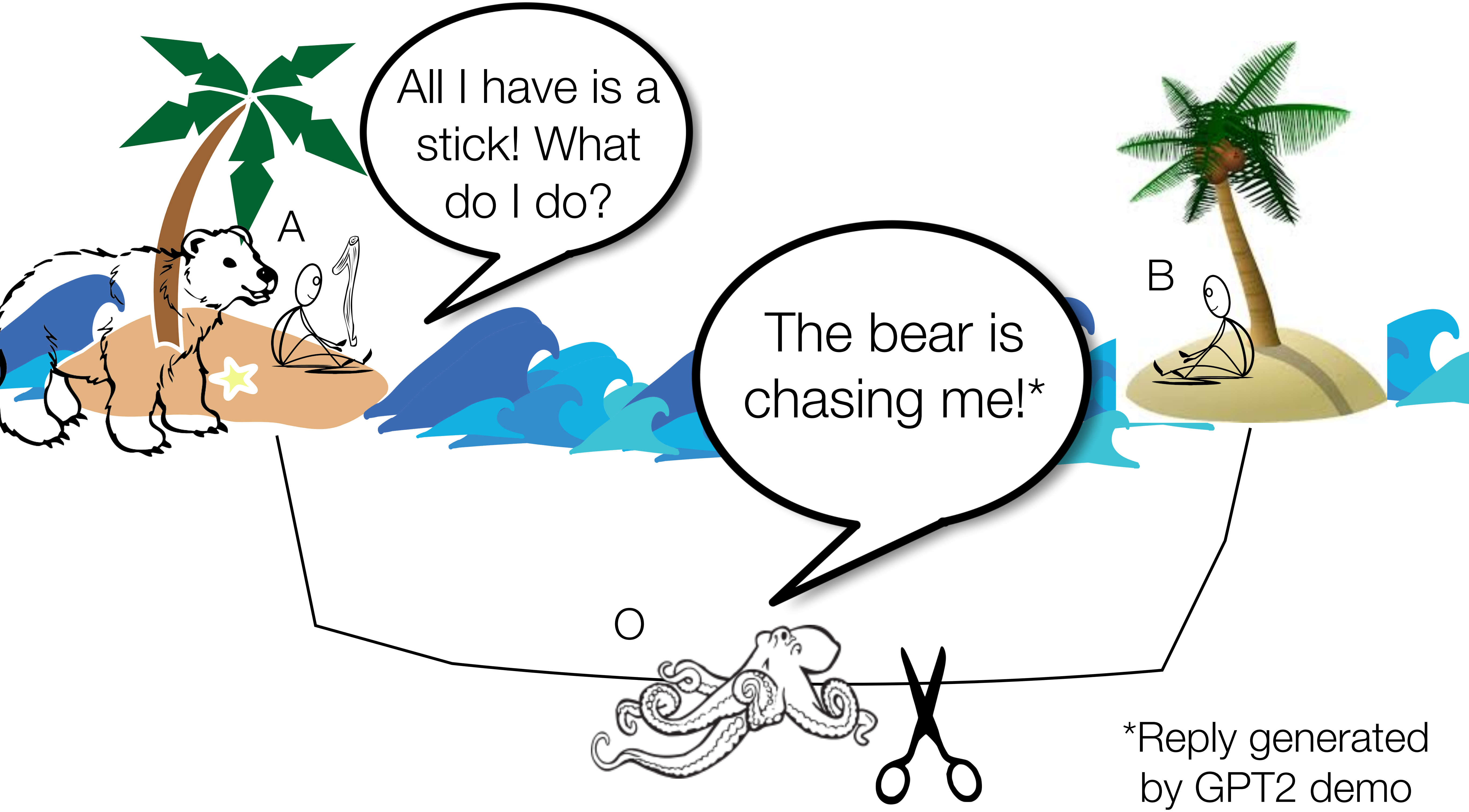
Thought experiment: Meaning from form alone



Thought experiment: Meaning from form alone

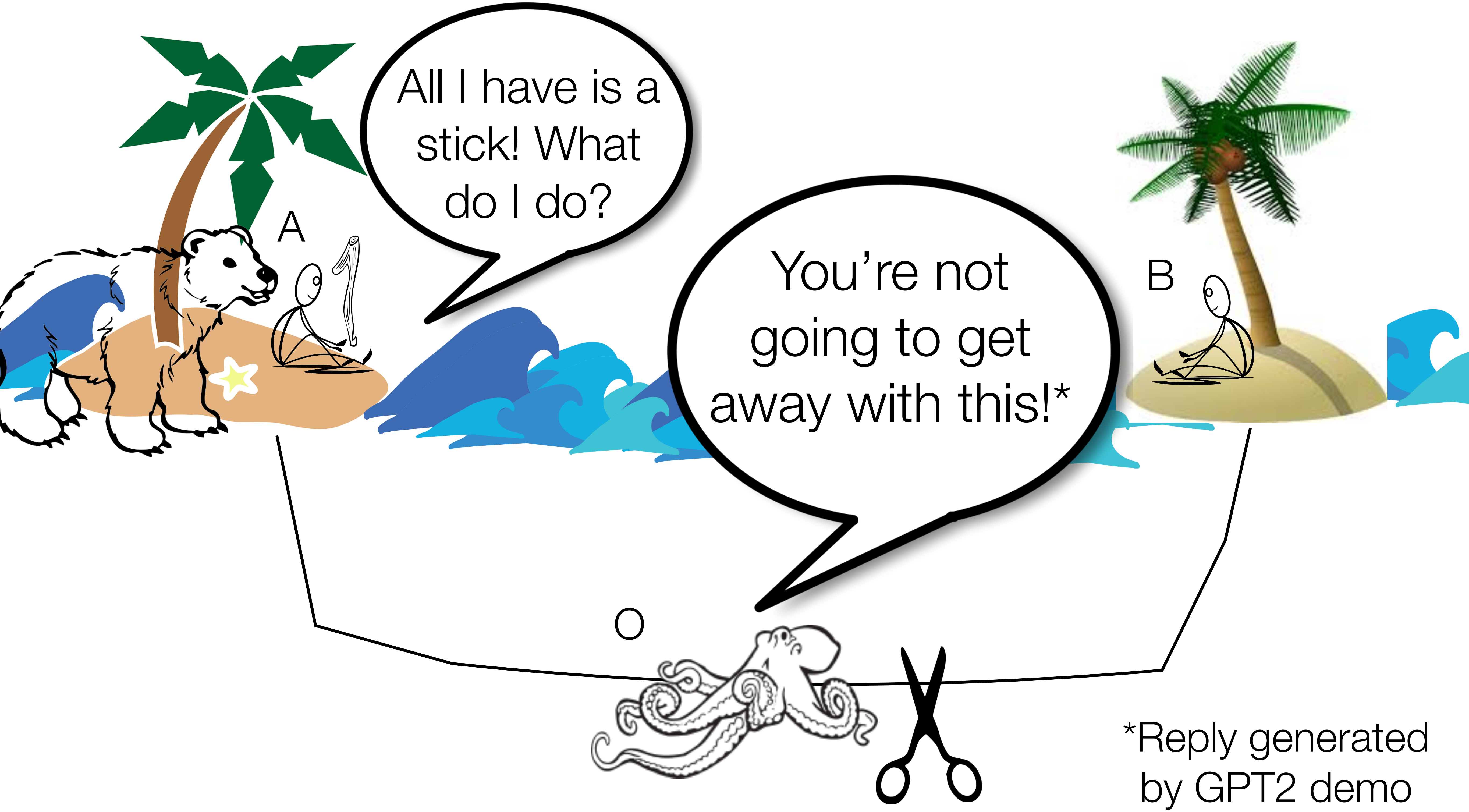


Thought experiment: Meaning from form alone



*Reply generated by GPT2 demo

Thought experiment: Meaning from form alone


A
All I have is a stick! What do I do?
You’re not going to get away with this!*
B
O

*Reply generated by GPT2 demo

Octopus Test: Analysis

- O did not learn to communicate successfully, and the reason is that
  O did not learn meaning.

- This is because O could only observe forms,
  and meaning can’t be learned from form alone.

  Learning the meaning relation requires access to the outside world
  so communicative intents can be hypothesized and tested.

- To the extent that A finds O’s utterances meaningful,
  it was not because O’s utterances made sense;
  it is because A, as a human active listener, could make sense of them.

Broader point

- The field of computational linguistics is making rapid progress, but we have made rapid progress before (grammar-based; statistical; …).

  How do we know this time it's different?

- One can look at progress in a field of science from two perspectives: top-down and bottom-up.

Top-down progress

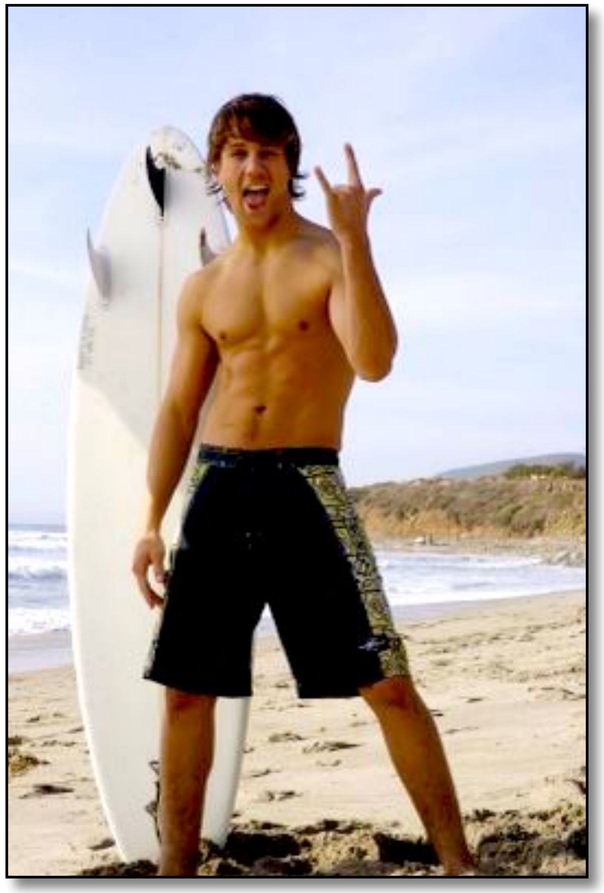
> "Semantics with no treatment of truth-conditions is not semantics."
>
> - Lewis 1972

We have not succeeded until we have succeeded completely.
Are we making progress towards our end goal?

Bottom-up progress

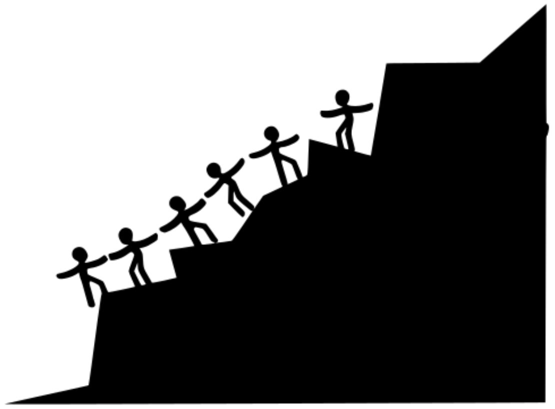
> "Using BERT … has been successful
> for single-turn machine comprehension."
>
> - Ohsugi et al. 2019

So much winning! And there will be more winning! Yeah!



We need thoughtful balance of bottom-up (rapid, fun hillclimbing) and top-down (climbing the right hill?).

Onwards!



- Value both error analysis and success analysis:
When a system does well on natural language “understanding” tasks, does it do that in a way which leads towards the end goal?
(Don’t allow the octopus to game the system.)

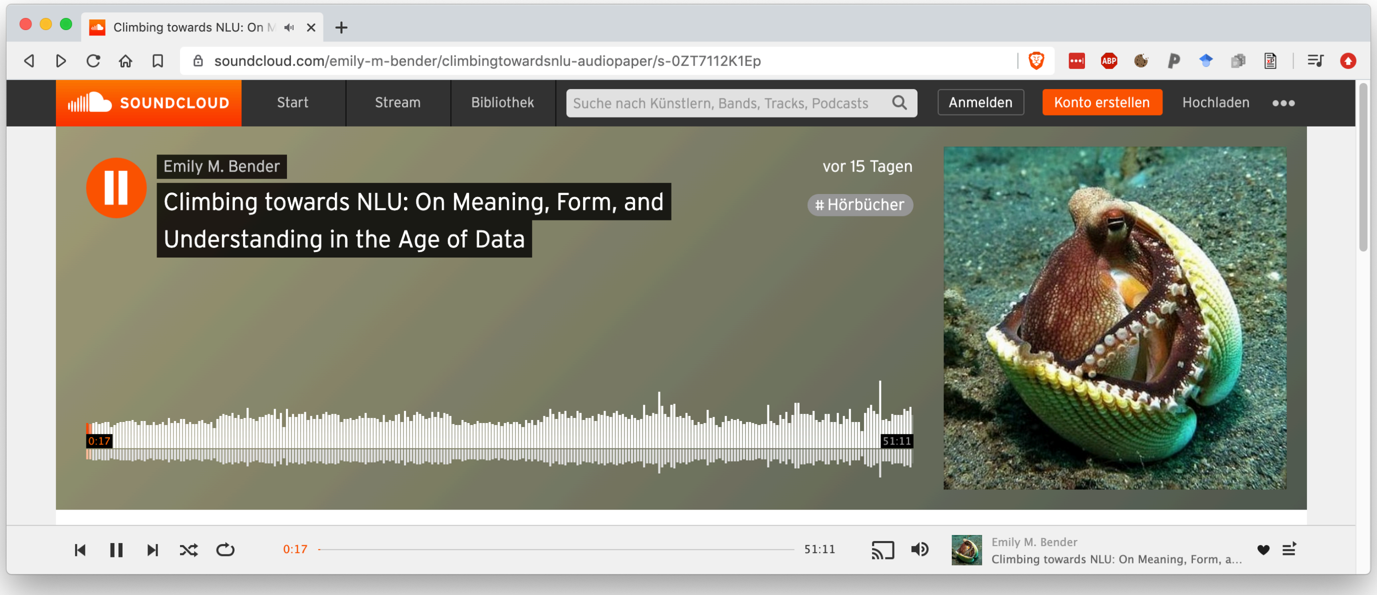
- Create tasks and datasets which ground language in reality/interaction.
Models trained on these don’t have to learn from form alone.

- Science over marketing: Let’s be careful with terms like ‘understanding’, ‘meaning’, and ‘comprehension’.

Come talk to us!

Q&A Sessions at ACL 2020
9A THEME-1: Tue July 7, 17:00 UTC+0
10A THEME-2: Tue July 7, 20:00 UTC+0

We also invite you to listen to our audiopaper:
https://soundcloud.com/emily-m-bender/climbingtowardsnlu-audiopaper/s-0ZT7112K1Ep



Climbing towards NLU: On M

soundcloud.com/emily-m-bender/climbingtowardsnlu-audiopaper/s-0ZT7112K1Ep

Emily M. Bender

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

vor 15 Tagen

Hörbücher

0:17

51:11

Emily M. Bender
Climbing towards NLU: On Meaning, Form, a...

NLP/Compling in the news

- Joint statement from the FTC, DOJ, EEOC and CFPB:
  - https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai