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?

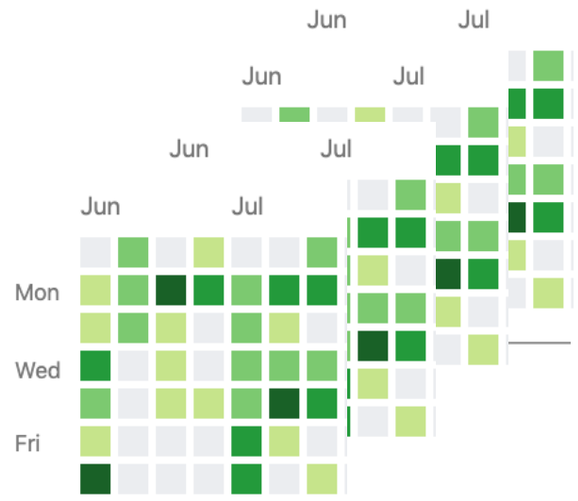
- Interaction is key: Exposure to a language via TV or radio alone is not sufficient (Snow et al 1976, Kuhl 2007)

- 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!

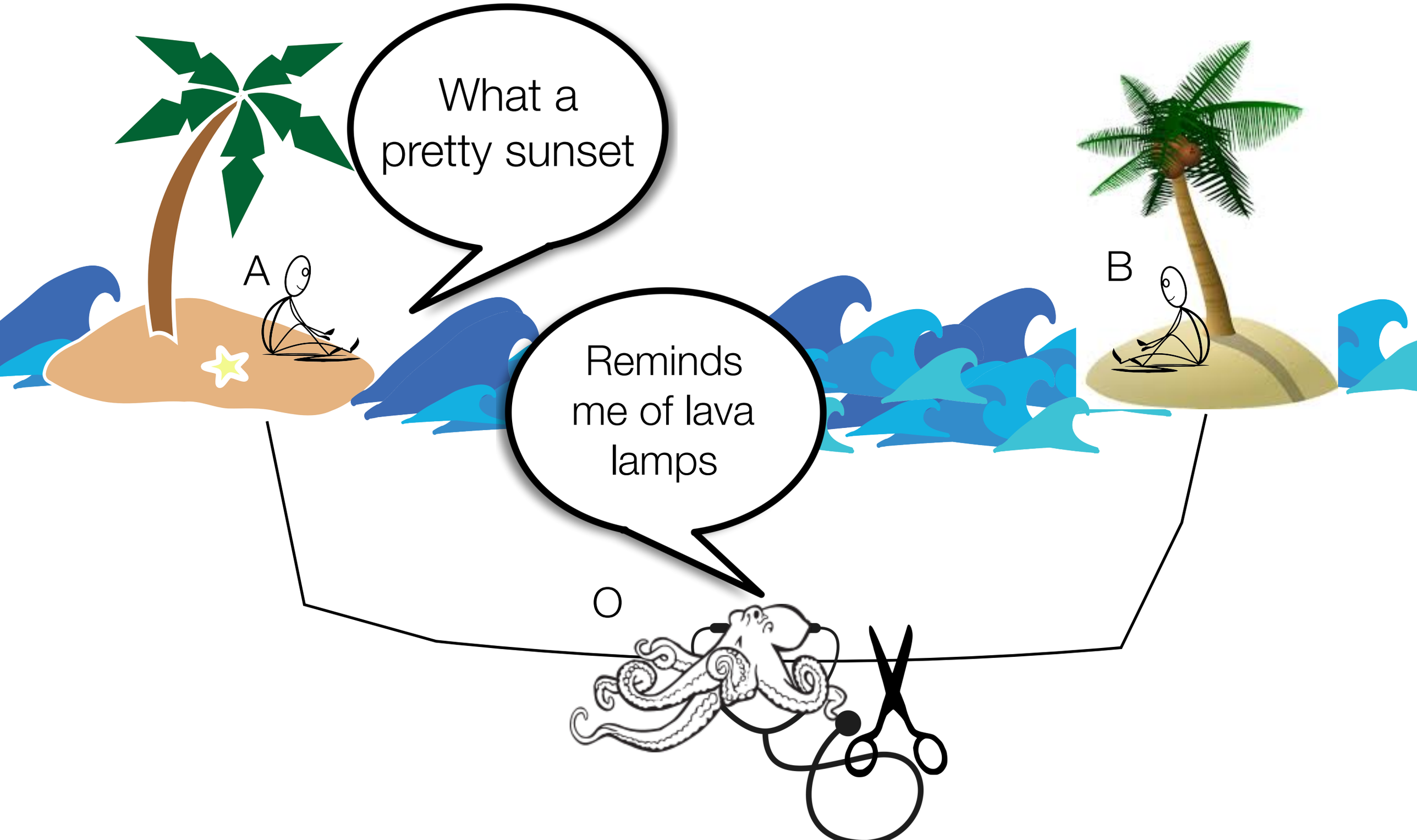
- Of course not! What’s interesting about this thought experiment is what makes the test unfair

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

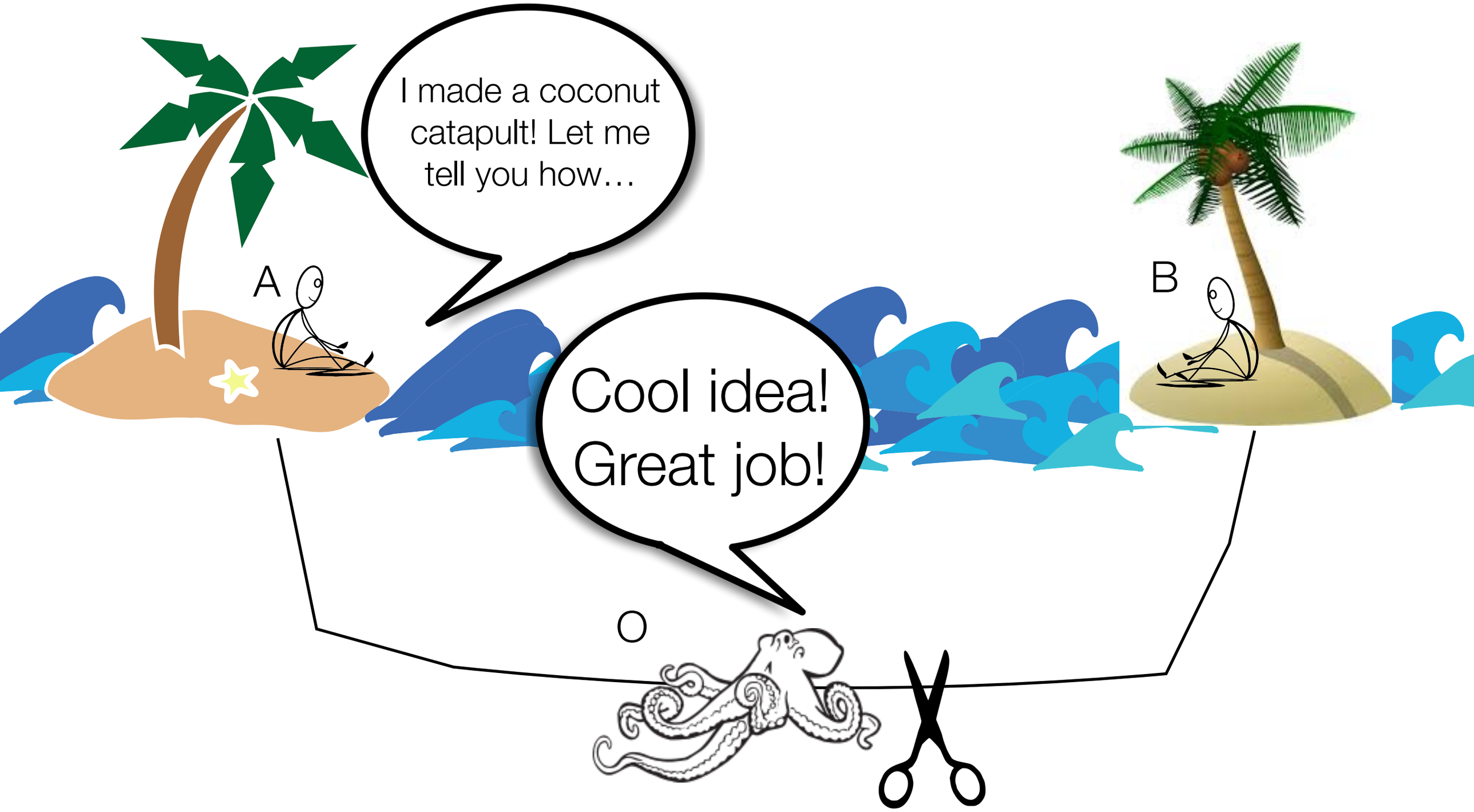
- What’s missing: Meaning — in the case of Java, what the machine is supposed to do, given the code

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

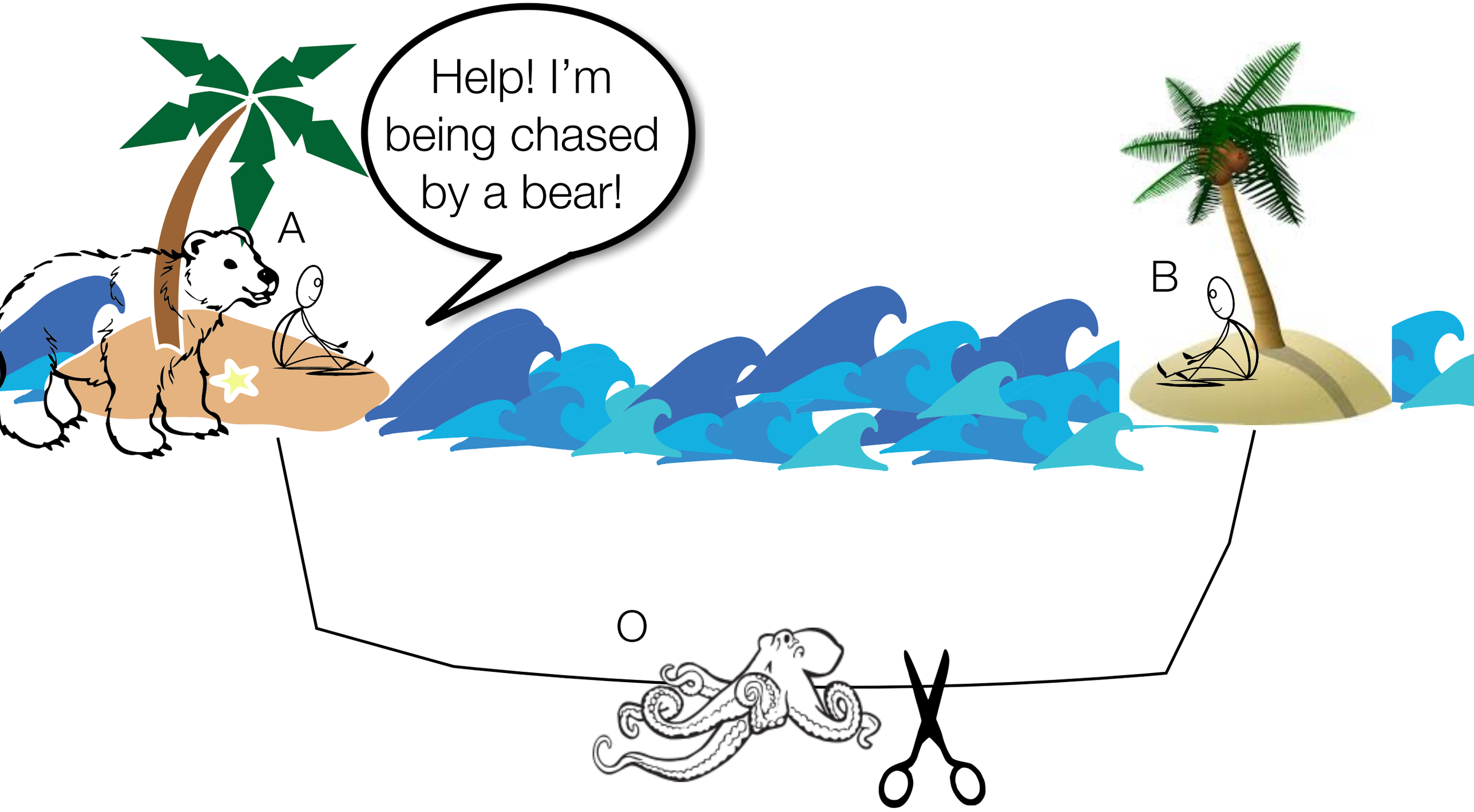
Thought experiment: Meaning from form alone



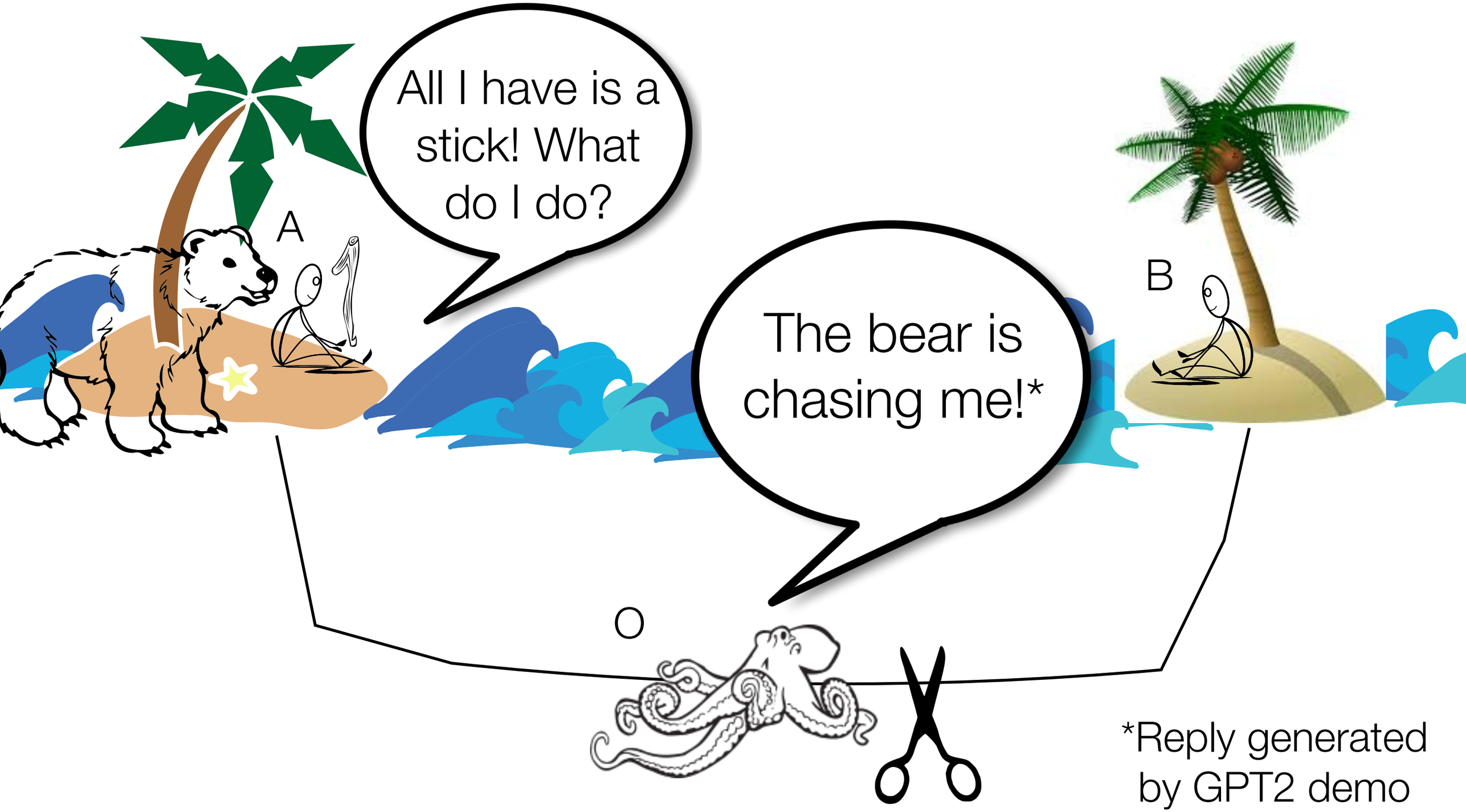
Thought experiment: Meaning from form alone



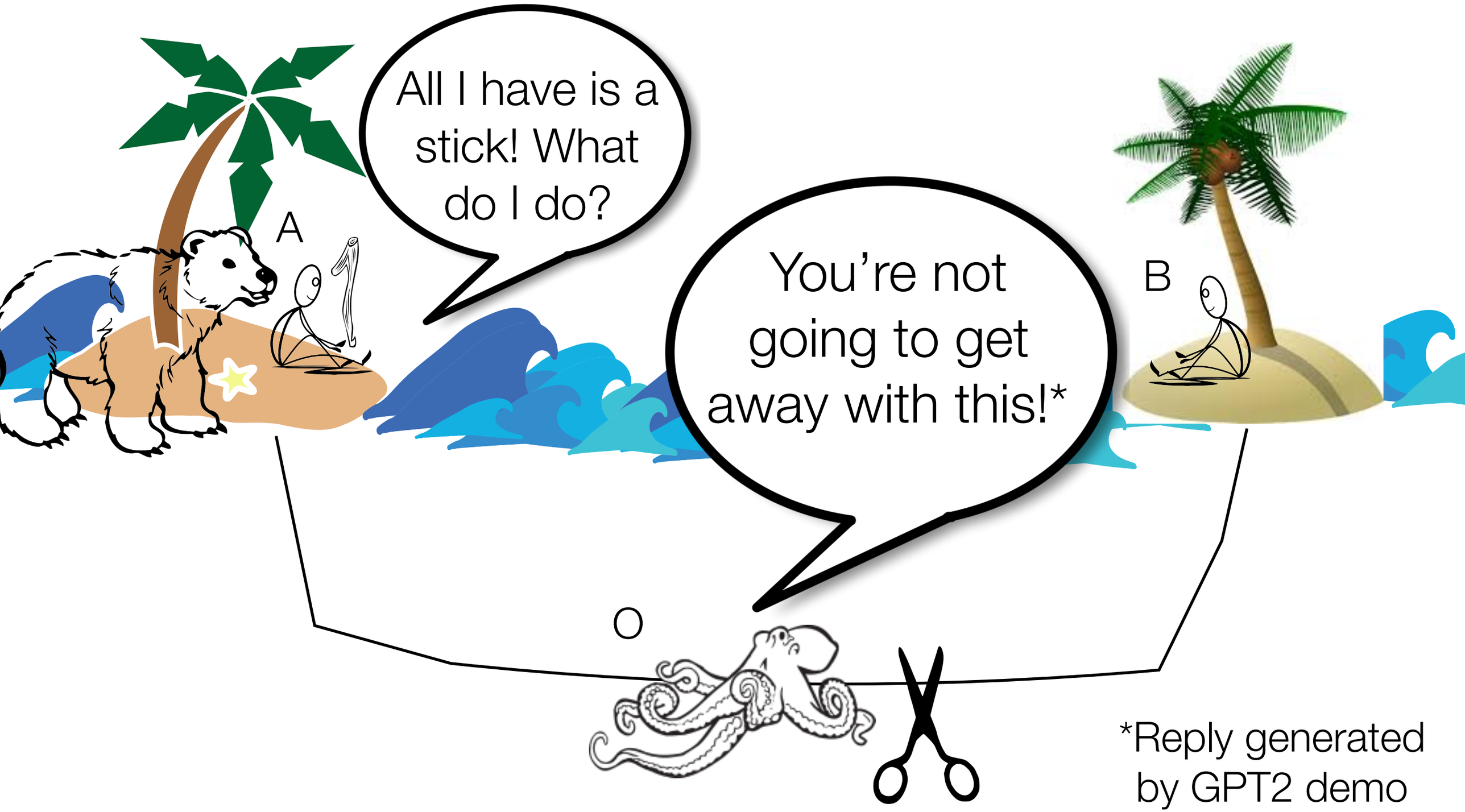
Thought experiment: Meaning from form alone



Thought experiment: Meaning from form alone



Thought experiment: Meaning from form alone



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?

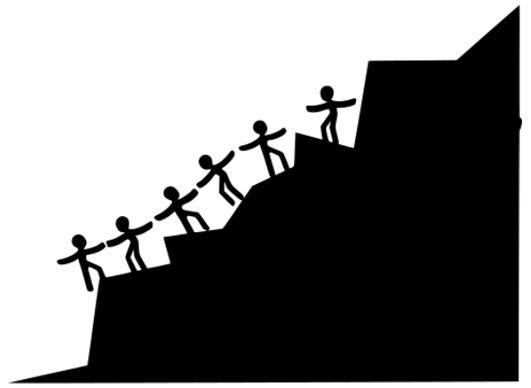
Bottom-up progress

> “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