Ling/CSE 472: Introduction to Computational Linguistics

5/23: Linguistic semantics and NLP

Overview

- English Resource Semantics
- Form & meaning (& octopusses & parrots)

MRS & ERS

- Minimal Recursion Semantics (Copestake et al 2005): A formalism for underspecified logical forms

- English Resource Semantics (Flickinger et al 2014): MRS representations for English sentences, including many design decisions about specific semantic phenomena

- ERG Semantic Documentation: An attempt to explain those representations for consumers of them (people who use the grammar in parsing or generation)

What's in an ERS?

- ERSes:
  - make explicit the connections between the semantic predicates introduced by the words
  - make explicit semantic predicates introduced by syntactic constructions
  - make explicit morphosemantic features such as person/number, tense/aspect, and sentential force

ERS examples: Predicate-argument structure

- The cheerful children wanted to sing and dance

$$\langle h_1, e_3,$$
$$\left| \begin{array}{l} h_4{:}\text{\_the\_q}(x_6, h_7, \_), h_8{:}\text{\_cheerful\_a\_1}(\_, x_6), h_8{:}\text{\_child\_n\_1}(x_6), \\ h_2{:}\text{\_want\_v\_1}(e_3, x_6, h_{10}), \\ h_{14}{:}\text{\_and\_c}(\_, h_{11}, h_{16}), h_{11}{:}\text{\_sing\_v\_1}(e_{12}, x_6, \_), h_{16}{:}\text{\_dance\_v\_1}(e_{17}, x_6, \_) \end{array} \right|$$
$$\{ h_1 =_q h_2, h_7 =_q h_8, h_{10} =_q h_{14} \} \rangle$$

- This technique is impossible to apply

$$\langle h_1, e_3,$$
$$\left| \begin{array}{l} h_4{:}\text{\_this\_q\_dem}(x_6, h_7, \_), h_8{:}\text{\_technique\_n\_1}(x_6), \\ h_2{:}\text{\_impossible\_a\_for}(e_3, h_9, \_), h_{11}{:}\text{\_apply\_v\_2}(e_{12}, \_, x_6) \end{array} \right|$$
$$\{ h_1 =_q h_2, h_7 =_q h_8, h_9 =_q h_{11} \} \rangle$$

ERS examples: Quantifiers

- All short jokes are funny v. All funny jokes are short

$$\langle h_1, e_3, \left| \begin{array}{l} h_4{:}\text{\_all\_q}(x_5, h_6, \_), \\ h_8{:}\text{\_short\_a\_of}(\_, x_5, \_), \\ h_8{:}\text{\_joke\_n\_1}(x_5), \\ h_2{:}\text{\_funny\_a\_1}(e_3, x_5) \end{array} \right| \{ h_1 =_q h_2, h_6 =_q h_8 \} \rangle$$

$$\langle h_1, e_3, \left| \begin{array}{l} h_4{:}\text{\_all\_q}(x_5, h_6, \_), \\ h_8{:}\text{\_funny\_a\_1}(\_, x_5), \\ h_8{:}\text{\_joke\_n\_1}(x_5), \\ h_2{:}\text{\_short\_a\_of}(e_3, x_5, \_) \end{array} \right| \{ h_1 =_q h_2, h_6 =_q h_8 \} \rangle$$

ERS examples: Scopal operators

- The meteorologist says it probably won't rain

$$
\langle\, h_1, e_3,
\left|\begin{array}{l}
h_4\text{:\_the\_q}(x_6, h_7, \_), h_8\text{:\_meteorologist\_n\_1}(x_6), \\
h_2\text{:\_say\_v\_to}(e_3, x_6, h_{10}, \_), \\
h_{11}\text{:\_probable\_a\_1}(\_, h_{13}), \\
h_{14}\text{:neg}(\_, h_{15}), \\
h_{17}\text{:\_rain\_v\_1}(e_{18})
\end{array}\right|
\left\{\begin{array}{l}
h_1 =_q h_2, h_7 =_q h_8, \\
h_{10} =_q h_{11}, h_{13} =_q h_{14}, h_{15} =_q h_{17}
\end{array}\right\} \rangle
$$

ERS examples: Multi-word expressions

- Kim looked up the answer

$$\langle h_1, e_3, \left| \begin{array}{l} h_4{:}\text{proper\_q}(x_6, h_5, \_), h_8{:}\text{named}(x_6, \textit{Kim}), \\ h_2{:}\text{\_look\_v\_up}(e_3, x_6, x_9), \\ h_{10}{:}\text{\_the\_q}(x_9, h_{12}, \_), h_{13}{:}\text{\_answer\_n\_to}(x_9, \_) \end{array} \right| \{ h_1 =_q h_2, h_5 =_q h_8, h_{12} =_q h_{13} \} \rangle$$

- Kim looked up the chimney

$$\langle h_1, e_3, \left| \begin{array}{l} h_4{:}\text{proper\_q}(x_6, h_5, \_), h_8{:}\text{named}(x_6, \textit{Kim}), \\ h_2{:}\text{\_look\_v\_1}(e_3, x_6), h_2{:}\text{\_up\_p\_dir}(\_, e_3, x_{10}), \\ h_{11}{:}\text{\_the\_q}(x_{10}, h_{13}, h_{12}), h_{14}{:}\text{\_chimney\_n\_1}(x_{10}) \end{array} \right| \{ h_1 =_q h_2, h_5 =_q h_8, h_{13} =_q h_{14} \} \rangle$$

Where do ERSes come from?

- Implementations of analyses of specific constructions in the English Resource Grammar

- At parse time, these various analyses interact to produce syntactico-semantic structures for input sentences

Are ERSes ‘meanings’?

- More accurately: ‘meaning representations’
- Need to be paired with a model theory/interpretation function
- Include information that goes beyond any theory logic developed to date
- For the subset that is covered by e.g. predicate logic, compatible

What are ‘fingerprints’?

- Hypothesis: recurring subparts of ERSes that can be attributed to specific grammar entities (phrase structure rules, lexical rules, lexical types) are interesting candidates for ‘semantic phenomena’

- Fingerprints are schematized ERS pieces that should match the ERS for any sentence evincing the phenomenon they illustrate

- In principle, fingerprints can be used to search sembanks of sentences annotated with ERSes

- We hope that an explanation of ERG semantic analyses centered on fingerprints will make the representations more interpretable to non-grammar developers

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?

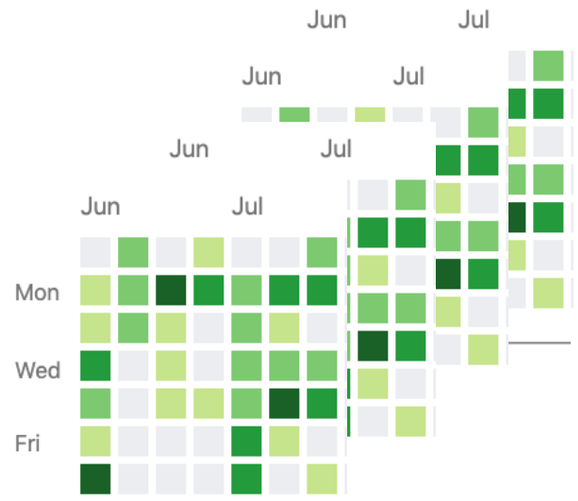
- Interaction is key: Exposure to a language via TV or radio alone is not sufficient (Snow et al 1976, Kuhl 2007)

- 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!

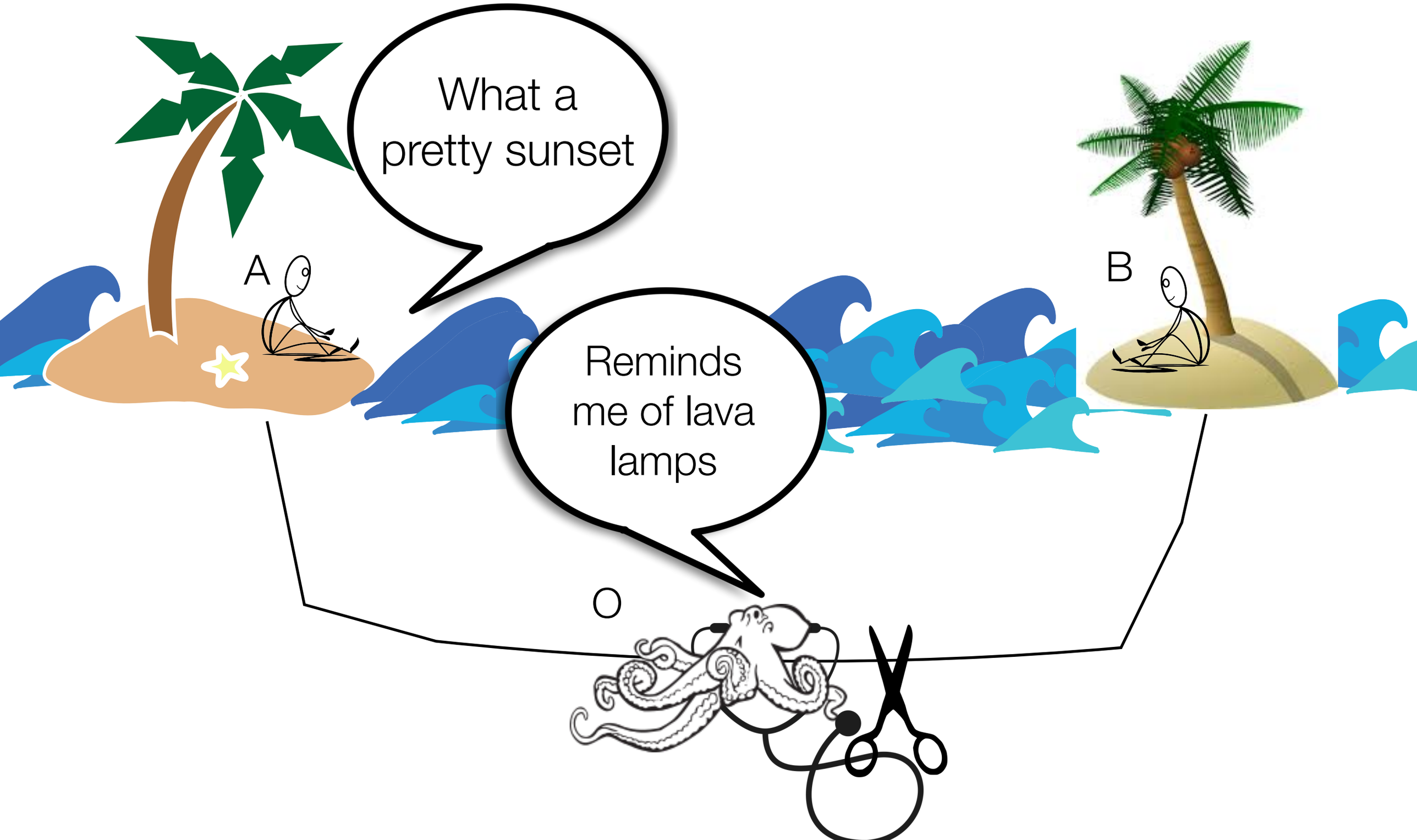
- Of course not! What’s interesting about this thought experiment is what makes the test unfair

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

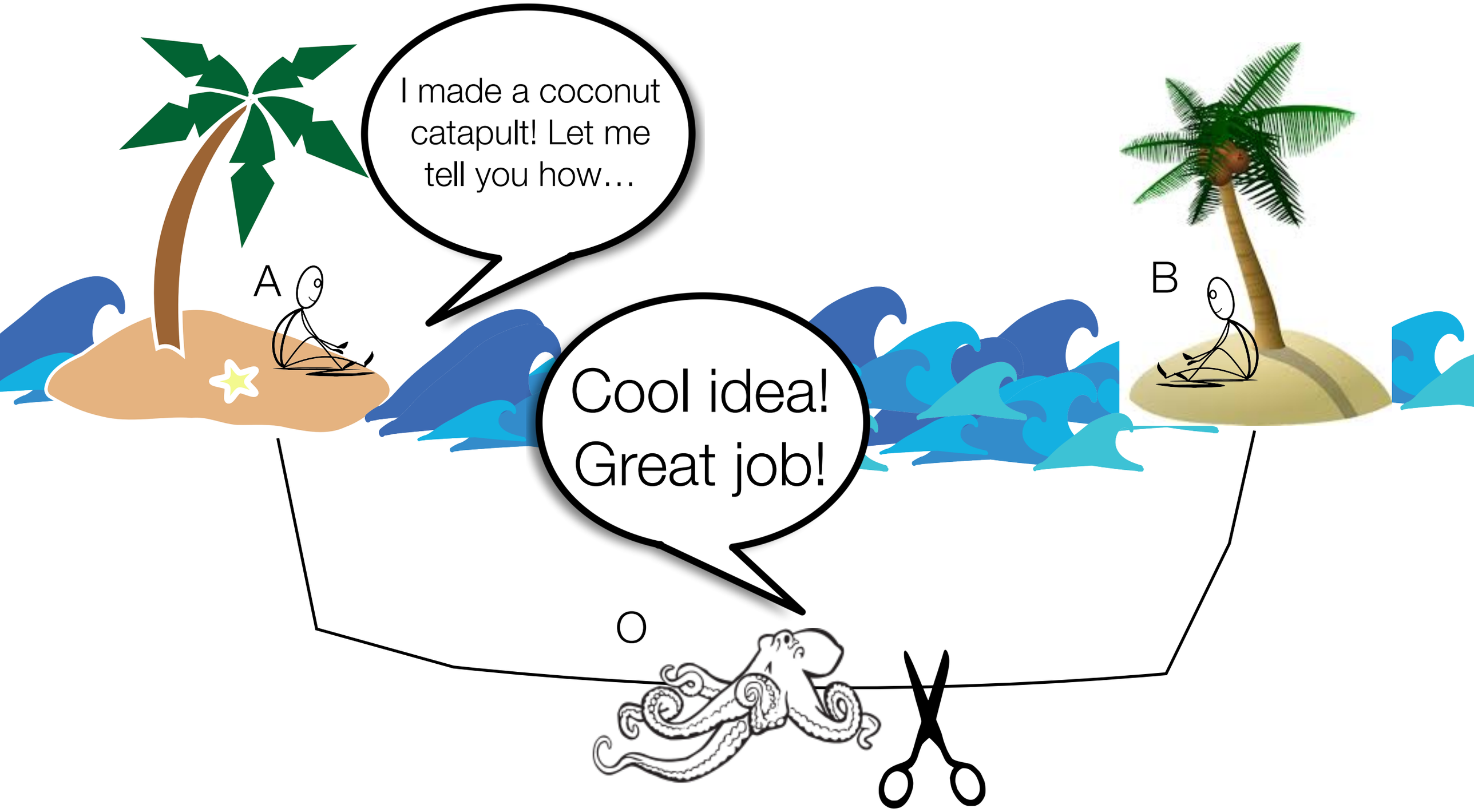
- What’s missing: Meaning — in the case of Java, what the machine is supposed to do, given the code

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

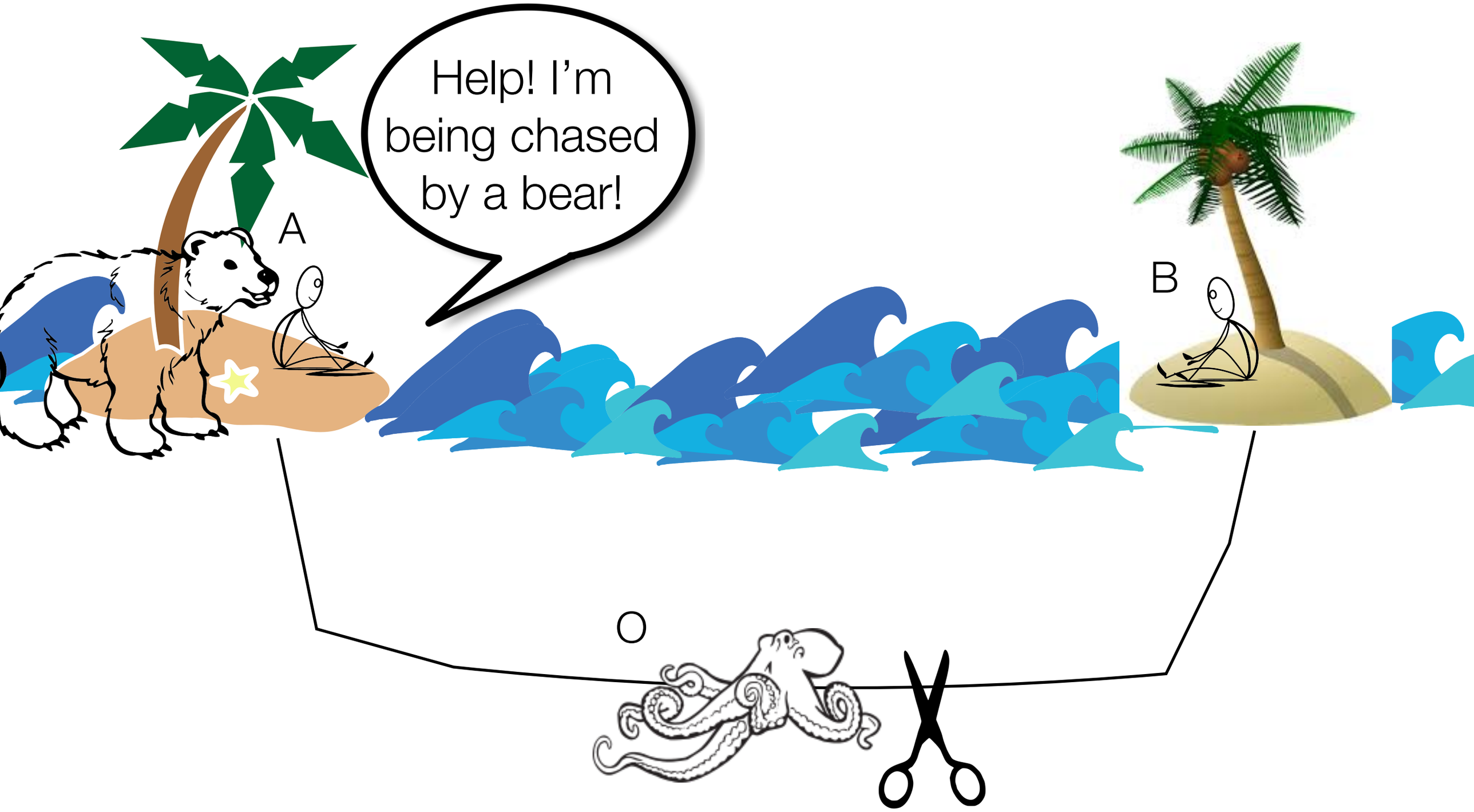
Thought experiment: Meaning from form alone



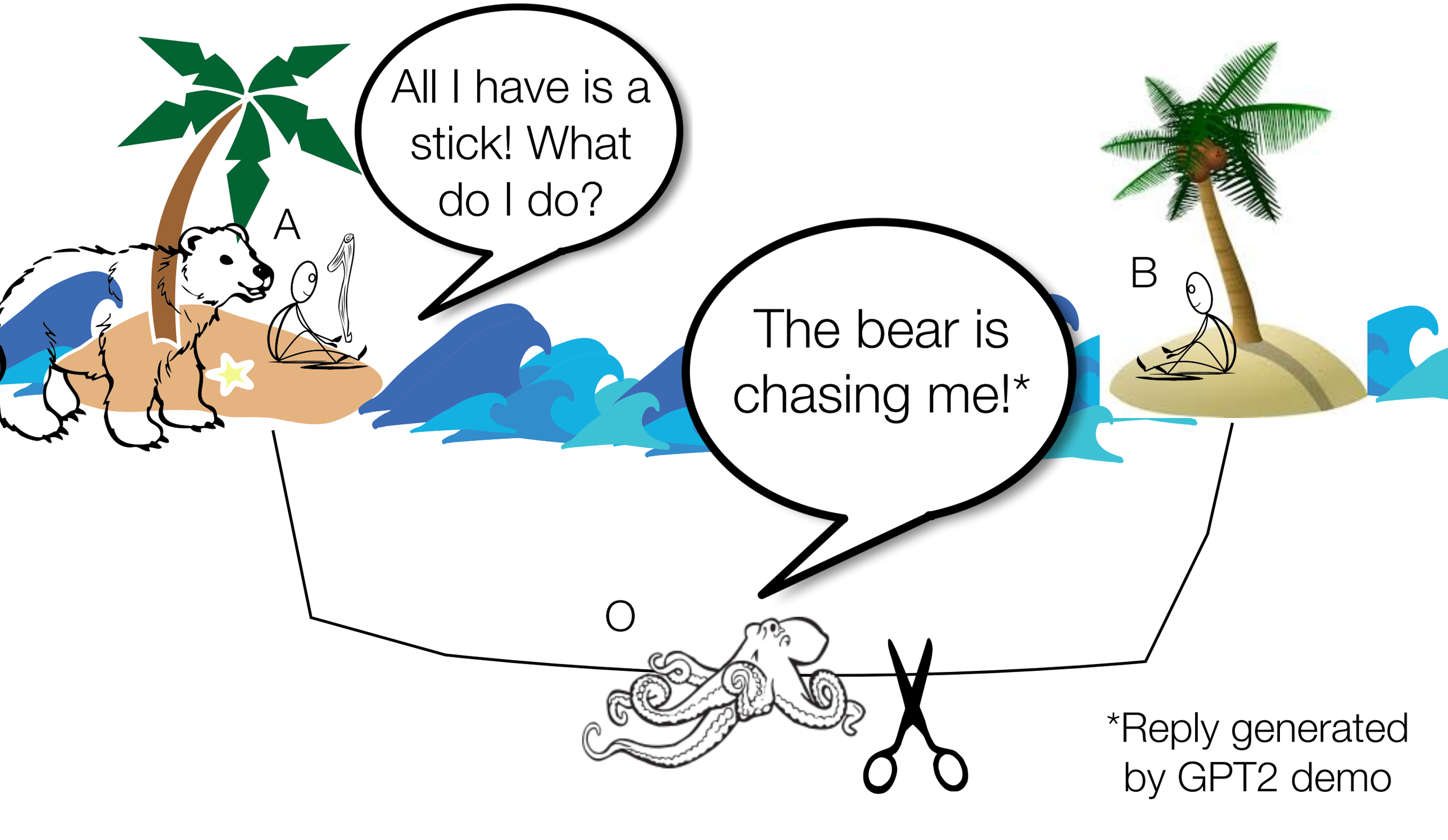
Thought experiment: Meaning from form alone



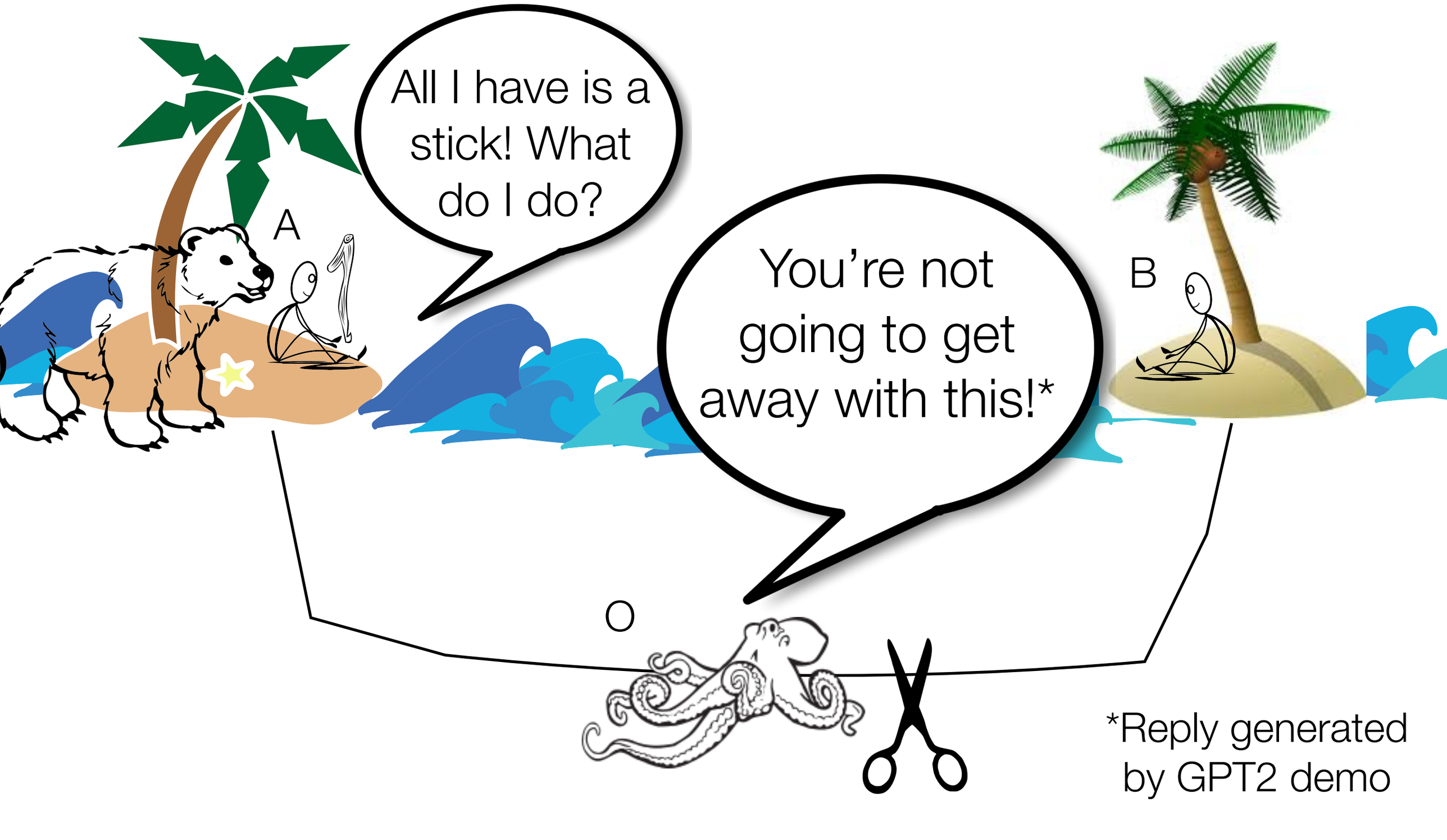
Thought experiment: Meaning from form alone



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?

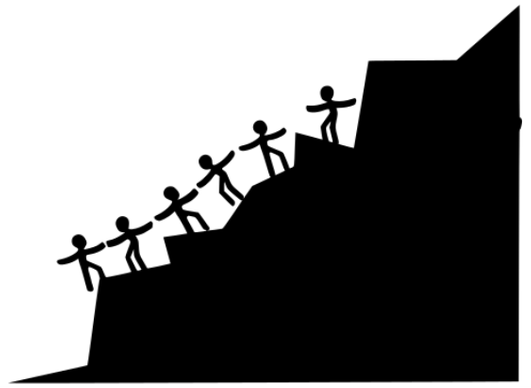
Bottom-up progress

> “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’.

Reading questions

- How do things like Vector Semantics play into this topic? The model may not "understand" the words it is being fed, but would be given a definition (or make a calculation) of how they relate to each other in the real world.

Reading questions

- How might that analysis change/develop with the advancement of LLMs? GPT-2 was used as an example, but what about GPT-4, which definitely outputs more appropriate responses for the bear situation. Also, looking back on this paper, would you say that the field of computational linguistics has started to sort of follow the guidelines/suggestions outlined in the paper, or is "climbing the wrong hill"?

- How might you update this paper to reflect the developments in LMs in the few years since publication? If anything, it seems like the hype has gotten even bigger and more inaccurate with the recent popularity of GPT 3 and 4. Would you change or add to any sections of the paper to address current themes behind these newer models?

Reading questions

- The idea of training distributional models on corpora alongside perceptual data like photos sounds like it could provide some form of context that enhances a language model's ability to produce more coherent (or "meaningful") text. Is this something that is already being incorporated in language models today? (If not, why is it not widely used?)

Reading questions

- In the paper, it described the thought that the octopus cannot deal with the message of “danger, bear” as it has never heard of, it cannot make a responses to it. However, in the case of new inventions, it was able to spin up some filler words. This has me relates to my user experience of chatGPT where it will always produces a response. It looks very long, well formatted, but almost meaningless nonsense. It is similar to the behavior that a human guess the meaning of a new word without actually understanding it? Is this a phenomenon that is an inherent prove that machines cannot understand language?

Reading questions

- A topic referenced in this paper distinguishes between 'learning meaning' and 'reflecting meaning', and how current language models can reflect but not truly understand the content it absorbs. Today we are clearly very far from AGI. However, LLMs are continuing to advance and new architectures are always being developed - is there in theory a fundamental point or threshold at which a machine/model could reach this point of truly understanding something?

- Thought experiment: Imagine we have an advanced machine that simulates every neuron and complex chemical interaction within a human brain, receiving the same inputs and producing the same outputs as the human. If such a machine were possible, would it not be capable of 'learning meaning' in the same way a human does? My perspective is that as human beings, we represent some physical phenomena that follow the laws of the universe. What, in your view, is the fundamental difference between a biological being and an artificial system that might prevent the latter from truly understanding meaning?

Reading questions

- At the moment, we can't really say that LLMs like Chat-GPT or GPT-4 really understand language or the meaning of phrases, but it sure feels like it does at times when it gives accurate responses. Since these LLMs are just like the octopus, they learn statistically what word should best come next and thus claims by the media of NLU are not accurate. However, what does the research into AI having human-analogous NLU look like? Is it even possible for AI to be able to understand the meaning of words and not just statistically what word should best follow the previous word? I have no idea how this could be done but I am sure that people are working on it.

Reading questions

- What would "grounding meaning in the real world" look like? Would this not be possible at a small scale for a language model in a very specific domain?

- Is there a difference between a language model that takes meaning into account and a language model that "understands" meaning? Or should we consider these the same thing?

Reading questions

- A focus of the paper is that to understand meaning, there must be communicative intent, and for communicative intent, the speaker and listener must be grounded in the real world and understand what their speech refers to. This works well to explain the lack of meaning/understanding in the text produced by chatbots, but how would you extend this argument to devices such as Alexa which can be used to make purchases, which affects the real world? Would you say they have some limited form of "understanding"?

Reading questions

- Are there ethical reasons for why it's better to keep LMs as purely form-based rather than developing LMs that "understand" the meanings of phrases and can express communicative intent? Are there people out there who are speaking against that sort of development? What are some risks involved?

On the dangers of stochastic parrots
Can language models be too big? 🦜

Emily M. Bender
University of Washington
@emilymbender

AI Sweden & RISE NLP
Sept 14, 2022
Originally presented at FAccT 2021

Slides:
https://bit.ly/ParrotsSept2022

- Joint work with: Timnit Gebru, Angelina McMillan-Major, Margaret Mitchell, Vinodkumar Prabhakaran, Mark Díaz, and Ben Hutchinson

- *Prabhakaran*: Prabhakaran et al 2012, Prabhakaran & Rambow 2017, Hutchison et al 2020

- *Hutchinson*: Hutchinson 2005, Hutchison et al 2019, 2020, 2021

- *Díaz*: Lazar et al 2017, Díaz et al 2018

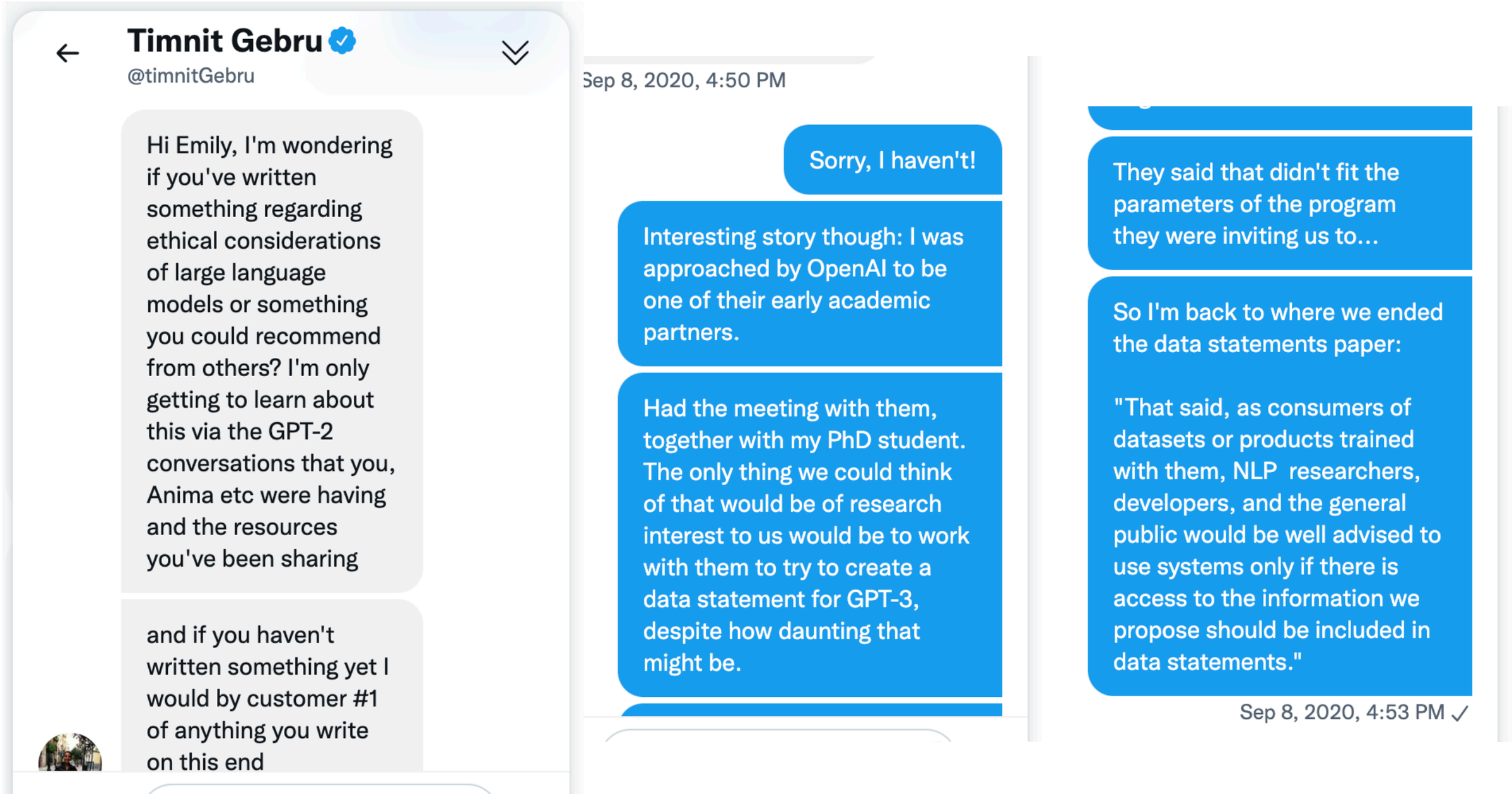
Slides: https://bit.ly/ParrotsSept2022






History & context of this paper

- Started off as a Twitter DM conversation, with Dr. Timnit Gebru:



History & context of this paper

I can think of three other angles around ethical implications of GPT-3 and the like:

1) Carbon cost of creating the damn things (see Strubell et al at ACL 2019)
2) AI hype/people claiming it's understanding when it isn't (Bender & Koller at ACL 2020)
3) Deepfakes/random generated text that no one is accountable for but which is interpreted as meaningful.

Sep 8, 2020, 4:54 PM ✓

This is somethign I'm trying to advocate for at Google

sent them some of your tweets as well haha

so many converstations about how we're not leading in large language models and should

GPT-3 is so impressive etc

and each time I'm like ANNDD see what Emily has to say

big data energy...

Sep 8, 2020, 4:55 PM ✓

I'm trying to advocate for documentation

first off

and second off interventionist data collection

Sep 8, 2020, 4:55 PM

See Gebru et al 2018 :)

What is interventionist data collection?

well saying instead of only depending

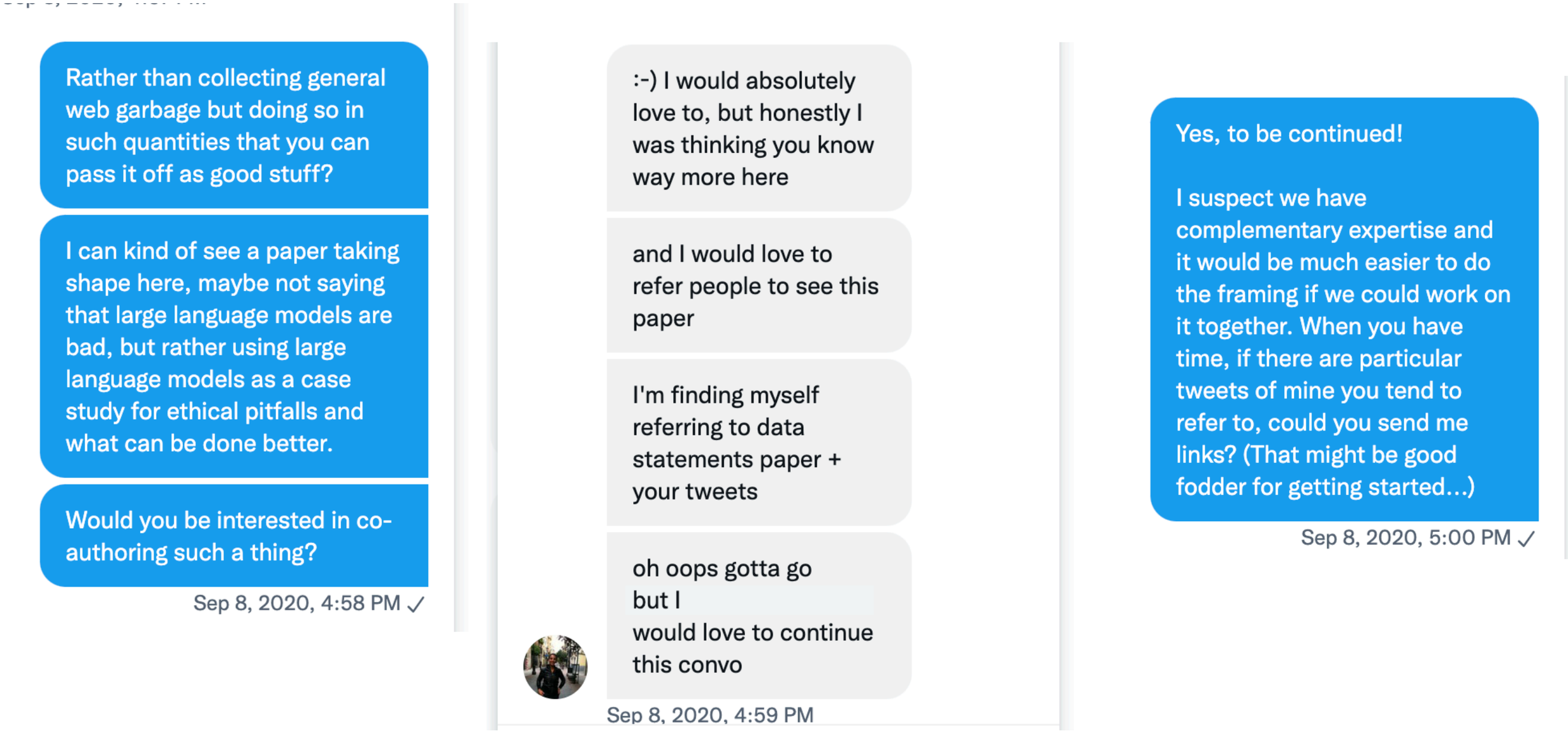
on whats available

like the internet

also think of ways to get other data, and also curate more. This is Eun Seo's term :)

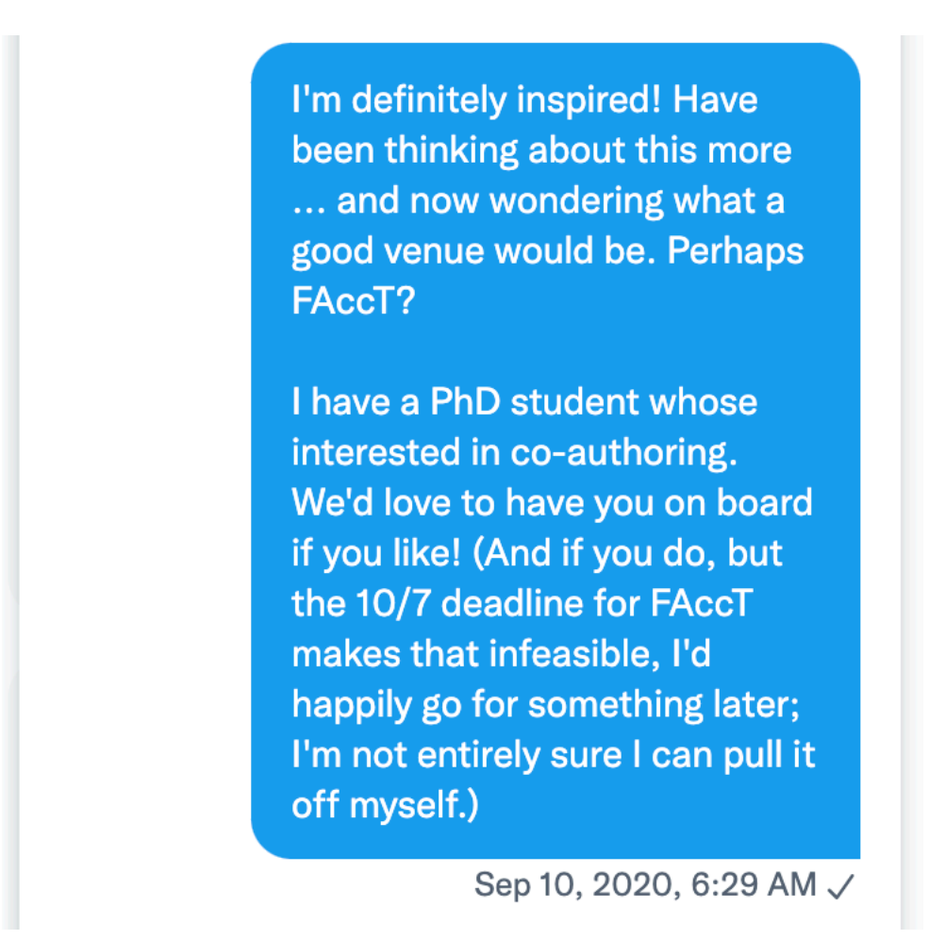
she's a historian and we talk about how in our view, archives are more interventionist and anything ML on the other end of the spectrum

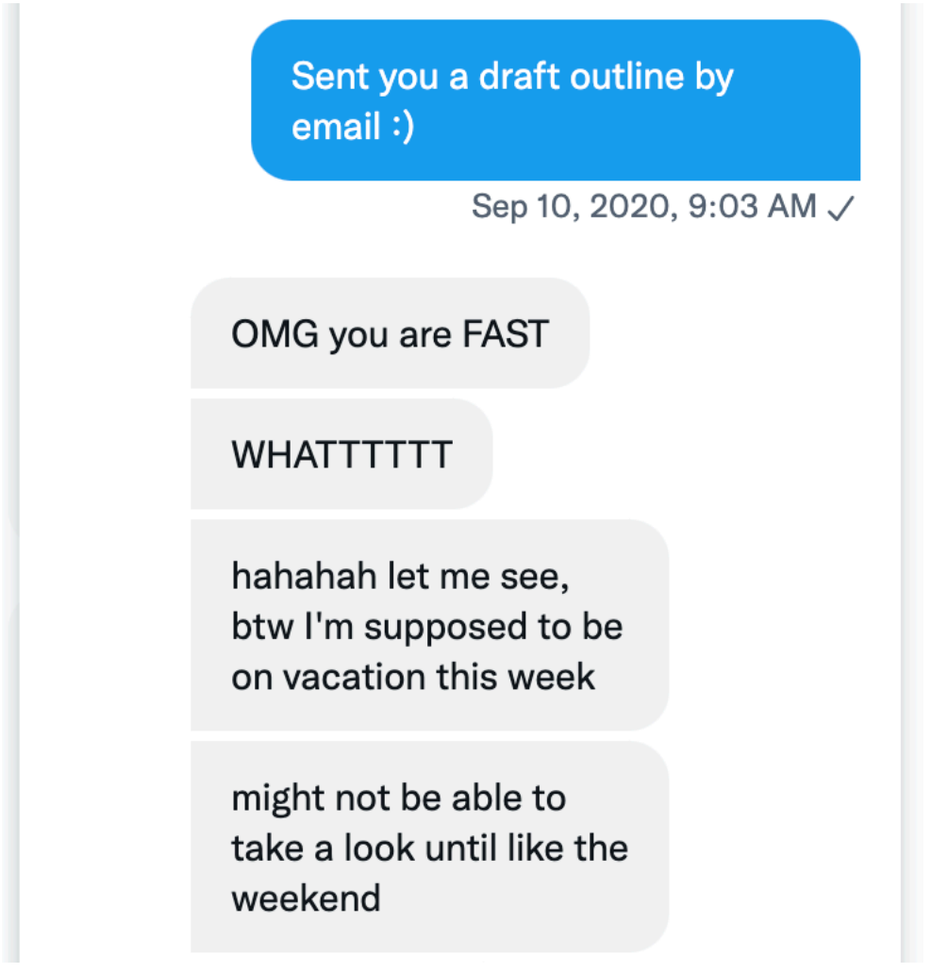
History & context of this paper



History & context of this paper

- Two days later:





History & context of this paper

- Four more co-authors joined from Google
- Pooled knowledge of 7 co-authors made it possible to pull the paper together by the Oct 7 deadline for FAccT
    - Survey/position paper: No new data analysis or experiments
- Cleared Google's "pub-approve" process before submission
- Sent it out to 30+ people for feedback in parallel to the peer review process

History & context of this paper

- Sharing all of this to help situate this research in its context
    - I’d like the audience to understand what our goals were for the paper as a piece of scholarship
    - You probably knew of this paper from the news before reading it, quite different to how one normally approaches research papers
- I think it’s also interesting to reflect on the processes of scholarship

History & context of this paper

- Late November: Google asks Dr. Gebru to either retract the paper or remove the Google co-authors' names from it

- Dr. Gebru pushes back, asking for information on what exactly was being objected to and objecting to how she & her team were being treated

- December 2, 2020: Google fires Dr. Gebru

- Dr. Margaret Mitchell starts documenting what happened to Dr. Gebru and calling on people within Google to apologize & fix systems

- February 19, 2021: Google fires Dr. Mitchell

History & context of this paper

- Google's actions led to intense media interest, both about their treatment of Dr. Gebru (and eventually Dr. Mitchell) and about our research
    - Selected media coverage
- Someone leaked the "pub-approve" version of the paper to Reddit
- Meanwhile…

History & context of this paper

- FAccT 2021 primary reviewers were complete before the media story broke (preserving anonymous review)

- FAccT 2021 acceptances announced on December 22, 2020

- Camera ready due January 22, 2021

- Still not allowed to include co-authors with Google affiliations

- Joint work with: Timnit Gebru, Angelina McMillan-Major, Margaret Mitchell, Vinodkumar Prabhakaran, Mark Díaz, and Ben Hutchinson

- *Prabhakaran*: Prabhakaran et al 2012, Prabhakaran & Rambow 2017, Hutchison et al 2020

- *Hutchinson*: Hutchinson 2005, Hutchison et al 2019, 2020, 2021

- *Díaz*: Lazar et al 2017, Díaz et al 2018

Slides: https://bit.ly/ParrotsSept2022






We would like you to consider

- Are ever larger language models (LMs) inevitable or necessary?

- What costs are associated with this research direction and what should we consider before pursuing it?

- Do the field of natural language processing or the public that it serves in fact need larger LMs?

- If so, how can we pursue this research direction while mitigating its associated risks?

- If not, what do we need instead?

Overview

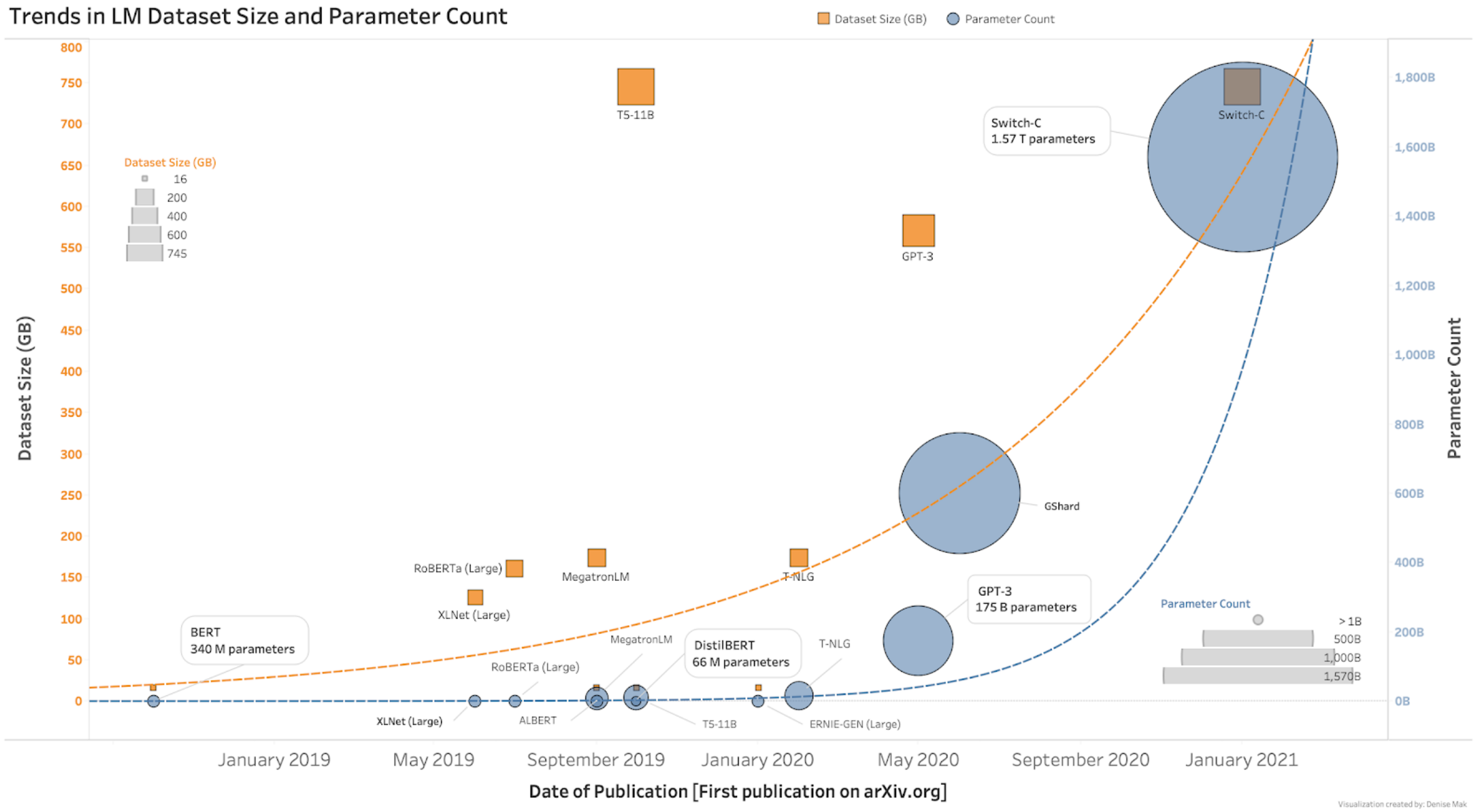
- History of Language Models (LMs)
- Risks
    - Environmental and financial costs
    - Unmanageable training data
    - Research trajectories
    - Potential harms of synthetic language
- Risk Mitigation Strategies

Brief history of language models (LMs)

- LM: A system trained to do string prediction
    - *What word comes ___? What word* [MASK] *here?*
- Proposed by Shannon in 1949, but implemented for ASR, MT, etc. in early 80's
    - N-grams and various neural architectures through Transformers
- Big takeaways
    - Better scores through more data and bigger models until scores don't improve, then move to new architecture
    - Multilingual models up to ~100 languages
    - Model-size reduction strategies
    - Growth of models $\propto$ range of application of models

How big is big?

[Special thanks to Denise Mak for graph design]



Updates since early 2021 (non-exhaustive)

Model	Source	Date	Parameters	Tokens	Citation
MT-NLG	Microsoft + NVIDIA	Oct 2021	530B	270B	
Gopher	DeepMind	Dec 2021	280B	300B ~1.3TB	(Rae et al 2021)
LaMDA	Google	Jan 2022	137B	1.56T	(Thoppilan et al 2022)
PaLM	Google	Apr 2022	540B	780B	(Chowdhery et al 2022)
BLOOM	BigScience	July 2022	176B	366B	

What are the risks?

Environmental costs & financial inaccessibility

Environmental and financial costs

- Average human across the globe responsible for 5t of CO2 emissions per year*
- Strubell et al. (2019)
    - Transformer model training procedure on GPUs 284t of CO2 emissions
    - 0.1 BLUE score increase en-de results in increase of ~$150,000 in compute cost
    - Encourage reporting training time and sensitivity to hyperparameters
    - Suggest more equitable access to compute clouds through government investment
- Which researchers and which languages get to ‘play’ in this space and who is cut out?

*Source: Our World In Data

Current mitigation efforts

- Renewable energy sources
    - Still incur a cost on the environment & take away from other potential uses of green energy
- Prioritize computationally efficient hardware
    - SustainNLP workshop
    - Green AI and promoting efficiency as evaluation metric (Schwartz et al 2020)
- Document energy and carbon metrics
    - Energy Usage Reports (Lottick et al 2019)
    - Experiment-impact-tracker (Henderson et al 2020)

Costs and risks to whom?

- Large LMs, particularly those in English and other high-resource languages, benefit those who have the most in society

- Marginalized communities around the world impacted most by climate change
  - Maldives threatened by rising sea levels (Anthoff et al 2010)
  - 800,000 residents of Sudan affected by flooding (7/2020-10/2020)*

- But these communities are rarely able to see benefits of language technology because LLMs aren't built for their languages, Dhivehi and Sudanese Arabic

*Source: https://www.aljazeera.com/news/2020/9/25/over-800000-affected-in-sudan-flooding-un

What are the risks?

Unmanageable training data

A large dataset is not necessarily diverse

- Who has access to the Internet and is contributing?
    - Younger people and those from developed countries
- Who is being subject to moderation?
    - Twitter - accounts receiving death threats more likely to be suspended than those issuing threats (see also Marshall 2021)
- What parts of the Internet are being scraped?
    - Reddit - US users 67% men and 64% are ages 18-29 (Pew)
    - Wikipedia - only 8.8-15% are women or girls
    - Not sites with fewer incoming and outgoing links, like blogs
- Who is being filtered out?
    - Filtering lists primarily target words referencing sex, likely also filtering LGBTQ online spaces (see also Dodge et al 2021)

Static data/Changing social views

- LMs run the risk of ‘value lock’, reifying older, less-inclusive understandings

- BLM movement lead to increased number of articles on shootings of Black people and past events were also documented and updated (Twyman et al 2017)

    - But media also doesn’t cover all events and tend to focus on more dramatic content

- LMs encode hegemonic views; retraining/fine-tuning would require thoughtful curation (see Solaiman and Dennison 2021 for partial proof of concept)

- See also Birhane et al 2021: ML applied as prediction is inherently conservative

Bias

- Research in probing LMs for bias has provided a wealth of examples of bias
    - See Blodgett et al 2020 for a critical overview
- Documentation of the problem is an important first step, but not a solution
- Automated processing steps may themselves be unreliable
- Probing requires knowing what social categories the LM may be biased against
    - Need for local input before deployment

Curation, documentation, accountability

- *How big is too big?*
    - Budget for documentation and only collect as much data as can be documented
    - Documentation: understand sources of bias & potential mitigating strategies
    - No documentation: potential for harm without recourse
- *Documentation debt*: datasets both undocumented and too big to document post-hoc

What are the risks?

Research trajectories

Research time is a valuable resource

- Focus on LMs and achieving new SOTA on leaderboards, particularly NLU

- But LMs have been shown to excel due to spurious dataset artifacts (Niven & Kao 2019, Bras et al 2020)

- LMs trained only on linguistic form don’t have access to meaning (Bender & Koller 2020)

- Are we actually learning about machine language understanding?

What are the risks?

Potential harms of synthetic language

We can’t help ourselves

- Human-human interaction is co-constructed and leads to a shared model of the world (Reddy 1979, Clark 1996)

- Text generated by an LM is not grounded in any communicative intent, model of the world, or model of the reader’s state of mind

- Counter-intuitive, given the increasing fluency of text synthesis machines, but:

  - Have to account for our predisposition to interpret locutionary artifacts as conveying coherent meaning & intent (Weizenbaum 1976, Nass et al 1994)

Stochastic 🦜

- An LM is a system for haphazardly stitching together linguistic forms from its vast training data, without any reference to meaning: a *stochastic parrot*.

- Nonetheless, humans encountering synthetic text make sense of it

- Coherence is in the eye of the beholder

Potential harms

- Denigration, stereotype threat, hate speech: harms to reader, harms to bystanders

- Cheap synthetic text can boost extremist recruiting (McGuffie & Newhouse 2020)

- LM errors attributed to human author in MT

- LMs can be probed to replicate training data for PII (Carlini et al 2020)

- LMs as hidden components can influence query expansion & results (Noble 2018)

Potential harms

- These harms largely stem from the interaction of the ersatz fluency of today's language models + human tendency to attribute meaning to text

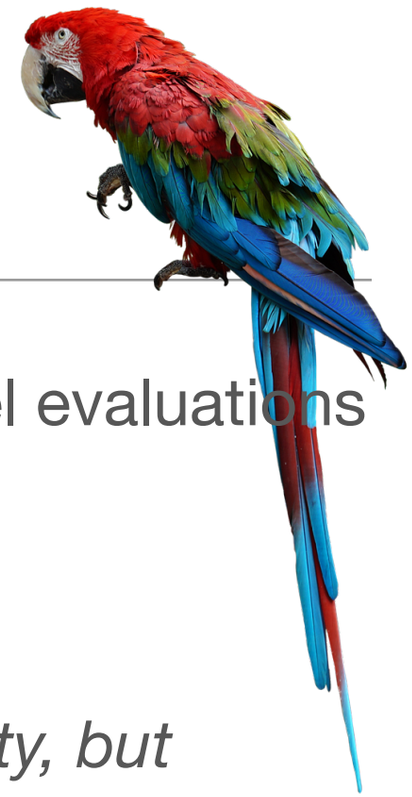
- Deeply connected to issue of accountability:
    - Synthetic text can enter conversations without anyone being accountable for it

- Accountability key to responsibility for truthfulness and to situating meaning

- Maggie Nelson (2015): "Words change depending on who speaks them; there is no cure."

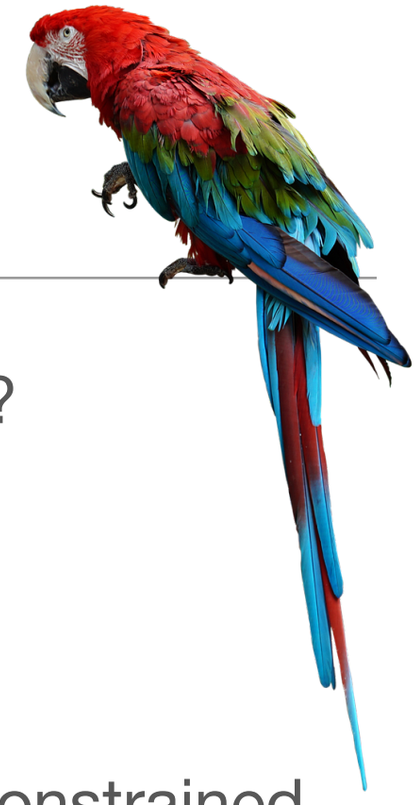
Risk management strategies

Allocate valuable research time carefully



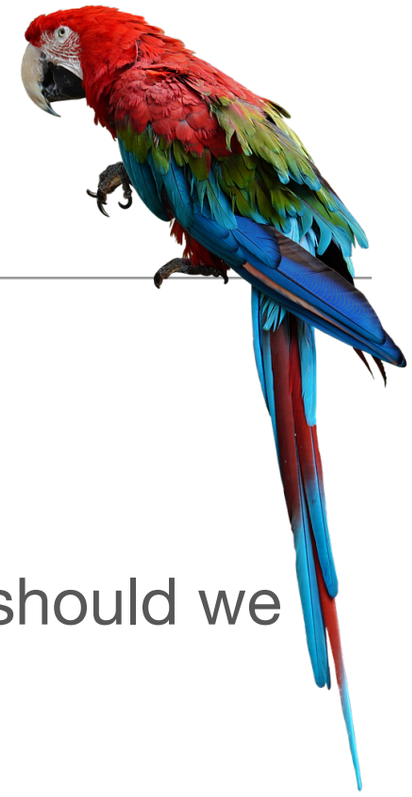
- Incorporate energy and compute efficiency in planning and model evaluations
- Select datasets intentionally
    - ‘*Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.*’ (Birhane and Prabhu 2021, after Ruha Benjamin)
- Document process, data, motivations, and note potential users and stakeholders
- Pre-mortem analyses: consider worst cases and unanticipated causes
- Value sensitive design: identify stakeholders and design to support their values

Risks of backing off from LLMs?



- What about benefits of large LMs, like improved auto-captioning?
  - Are LLMs in fact the only way to get these benefits?
  - What about for lower resource languages & time/processing constrained applications?
- Are there other ways the risks could be mitigated to support the use of LMs?
  - Watermarking synthetic text?
- Are there policy approaches that could effectively regulate the use of LLMs?

We would like you to consider



- Are ever larger language models (LMs) inevitable or necessary?
- What costs are associated with this research direction and what should we consider before pursuing it?
- Do the field of natural language processing or the public that it serves in fact need larger LMs?
- If so, how can we pursue this research direction while mitigating its associated risks?
- If not, what do we need instead?

The view from 2022

- Has the development of LLMs / tech based on LLMs slowed down? (No)
- Has data and model documentation become more mainstream? (Yes, but…)
- Have people become more aware of the risks of this technology? (Yes, but…)

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

Helping you when there isn't a simple answer

MUM has the potential to transform how Google helps you with complex tasks. MUM uses the T5 text-to-text framework and is 1,000 times more powerful than BERT. MUM not only understands language, but also generates it. It's trained across 75 different languages and many different tasks at once, allowing it to develop a more comprehensive understanding of information and world knowledge than previous models. And MUM is multimodal, so it understands information across text and images and, in the future, can expand to more modalities like video and audio.

Take the question about hiking Mt. Fuji: MUM could understand you're comparing two mountains, so elevation and trail information may be relevant. It could also understand that, in the context of hiking, to "prepare" could include things like fitness training as well as finding the right gear.

https://blog.google/products/search/introducing-mum/

See Shah & Bender 2022

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

Helping you when ther answer

MUM has the potential to transform how Goo tasks. MUM uses the T5 text-to-text framewor powerful than BERT. MUM not only understand it. It's trained across 75 different languages and many different tasks at once, allowing it to develop a more comprehensive understanding of information and world knowledge than previous models. And MUM is multimodal, so it understands information across text and images and, in the future, can expand to more modalities like video and audio.

Take the question about hiking Mt. Fuji: MUM could understand you're comparing two mountains, so elevation and trail information may be relevant. It could also understand that, in the context of hiking, to "prepare" could include things like fitness training as well as finding the right gear.

https://blog.google/products/search/introducing-mum/

Start now
Our platform can be plugged into any library, making it possible for NLP to be integrated into every build.

Large language models
Our models have been trained on billions of words, allowing them to learn nuance and context.

https://cohere.ai/

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

Helping you when ther… answer

MUM has the potential to transform how Goog… tasks. MUM uses the T5 text-to-text framewor… powerful than BERT. MUM not only understands l… it. It's trained across 75 different languages and m… allowing it to develop a more comprehensive unde… and world knowledge than previous models. And … understands information across text and images … expand to more modalities like video and audio.

Take the question about hiking Mt. Fuji: MUM cou… comparing two mountains, so elevation and trail i… It could also understand that, in the context of hik… include things like fitness training as well as findin…

https://blog.google/products/search/in…

Start now
Our platform can be plugged into any …

Large language models
Our models have been trained on …ving them to …text.

https://cohere.ai/

Is it real?

This is just an experiment with AI technology. We wanted to pay homage to a great thinker and leader with a fun digital experience. It is important to remember that AI in general, and language models specifically, still have limitations. The model can sometimes give inaccurate or inappropriate responses, so you should take any information given with a grain of salt.

https://ask-rbg.ai/

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

- Have people at large become better at critically analyzing claims of “understanding language”?

Thank you!

- Slides: https://bit.ly/ParrotsSept2022
- Twitter: @emilymbender

Overview

- English Resource Semantics
- Form & meaning (& octopusses)
- Next time: Catch-up/review