How to Build Language Processing Applications that Work — and Expose Those that Don't

Emily M. Bender
University of Washington
@emilymbender / @emilymbender@mastodon.social

Cross-disciplinary Research in Computational Law (CRCL22)
November 3, 2022
Brussels

Slides: bit.ly/EMB-CRCL22

Goals of this talk

- How to scope your NLP applications so that they are sensible
- How to detect wild claims
- What to know about NLP if you are in policy discussions

Slides: bit.ly/EMB-CRCL22

Outline

- Researcher stance
- A linguist's view of form & meaning
- Opportunities & dangers for NLP and the law
- Questions to ask of proposed applications
  - Is it effective?
  - Is it appropriate?
- Case studies

Researcher stance/Who am I?

- PhD training in syntax and sociolinguistics
- Long experience with multilingual grammar engineering: building grammars in software, across (mostly spoken) languages
- Since 2005: Faculty director of UW's Professional Masters in Computational Linguistics (CLMS)
- Since 2016: methodologies for supporting consideration of societal impacts of language technology—in NLP research, development, and education.
- Broader conversation about identifying and mitigating harms done in the name of "AI"

The meaning is not in the text

- What does this sentence *mean*?

- What does the speaker *mean* by uttering this sentence?

先生によると男の子よりも女の子がポケモンがすきだ。[jpn]

先生 によると 男の子 よりも 女の子 が ポケモン が すき だ。

Sensei ni yoru to otokonoko yorimo onnanoko ga pokemon ga suki da.

teacher (.) boy (.) girl (.) Pokemon (.) like (.)

teacher ACCORDING.TO boy THAN girl NOM Pokemon NOM like COP.PRES

‘According to the teacher, girls like Pokemon more than boys do.’

The meaning is not in the text

- With linguistic (grammatical, lexical) knowledge, speakers can get from a text to a ‘standing’ or ‘conventional’ meaning (Grice 1968, Quine 1960), but that’s only the first step.

- Standing meaning + commonsense + coherence relations gives *public commitments* (Hamblin 1970, Lascarides & Asher 2009, Asher & Lascarides 2013)

- Public commitments + further reasoning gives *perlocutionary consequences*

A: I wonder whether I should take my umbrella. Is it raining?
B: Yes.
A: Oh, so you do think I should take my umbrella.
B: I didn’t say that.

(Bender & Lascarides 2019:13)

Multiple levels visible when an utterance is examined closely

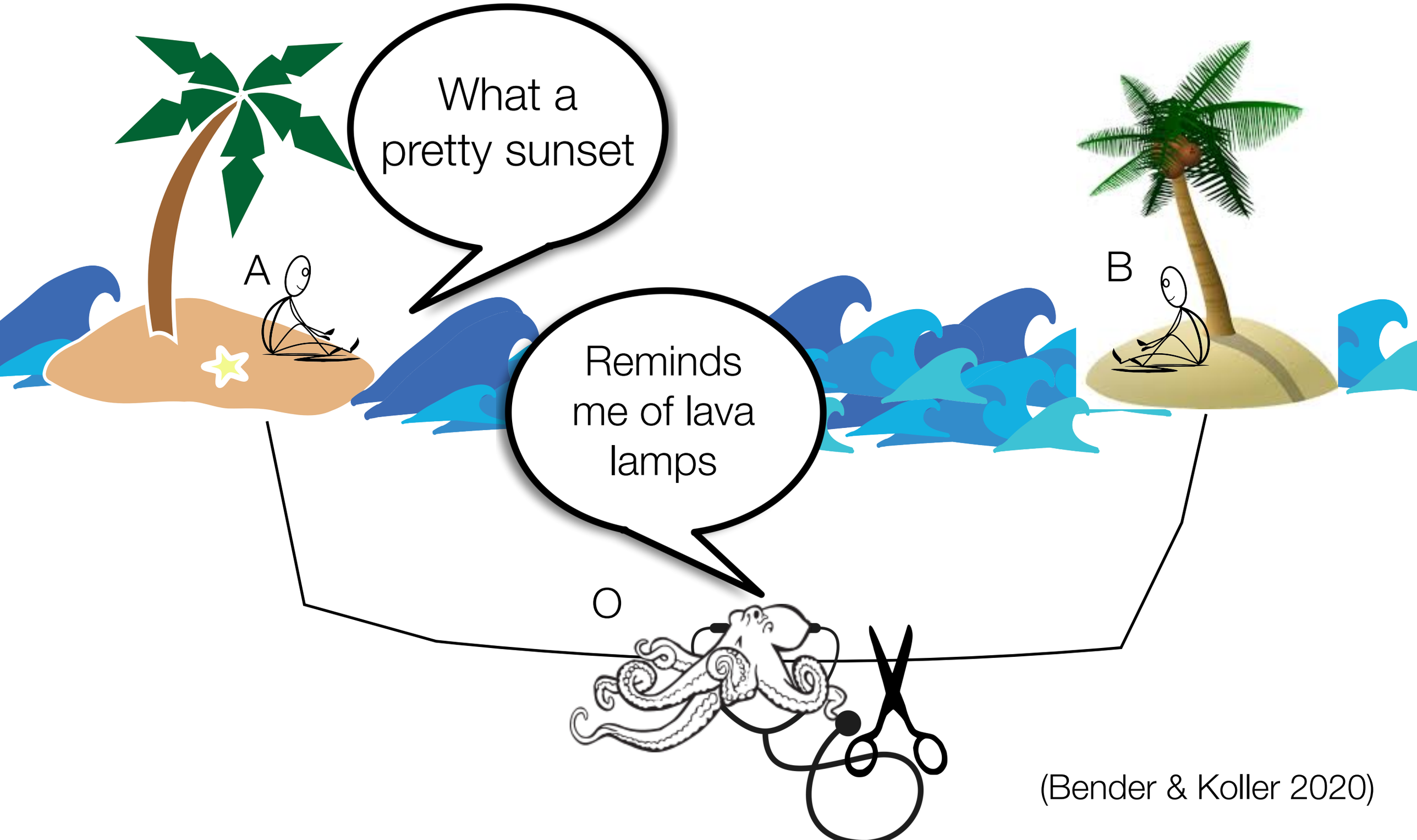
- The words that were uttered, with one or more possible grammatical structures
- One or more standing meanings
- Speaker’s publicly committed communicative intent
- Further inferences listener’s can make about the speaker’s beliefs
- The actual state of the world

Can language models ‘understand’?

- **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*

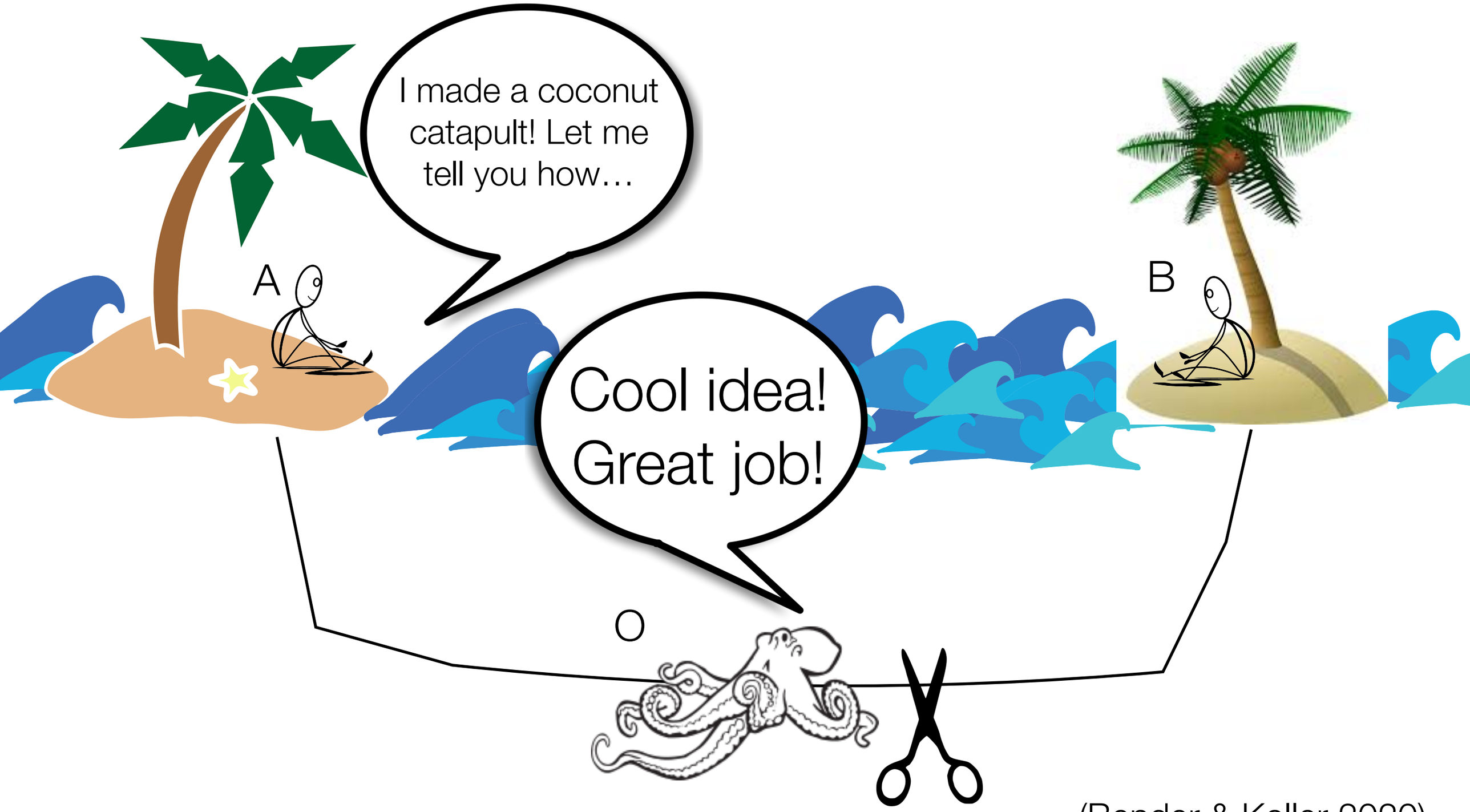
(Bender & Koller 2020)

Thought experiment: Meaning from form alone



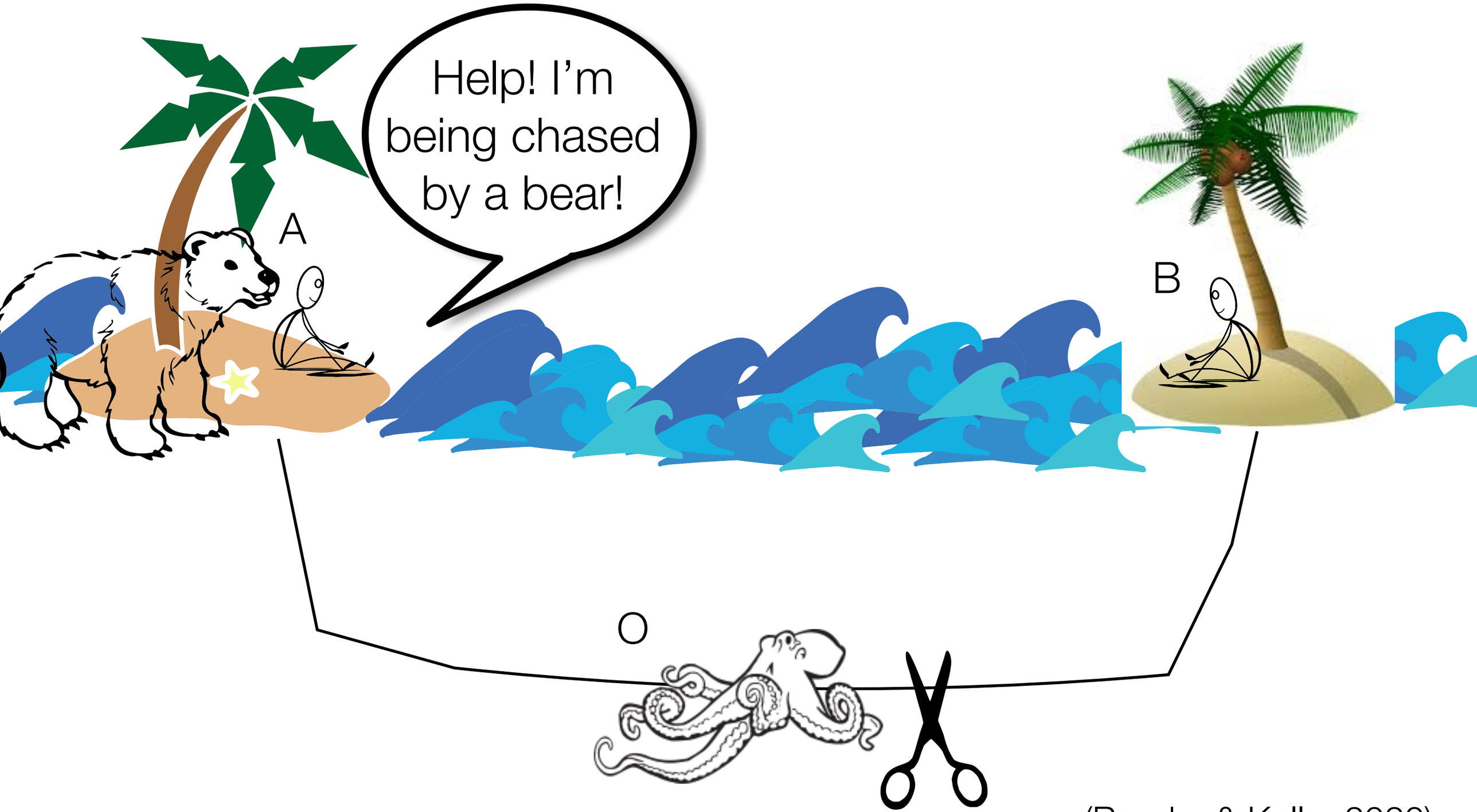
(Bender & Koller 2020)

Thought experiment: Meaning from form alone



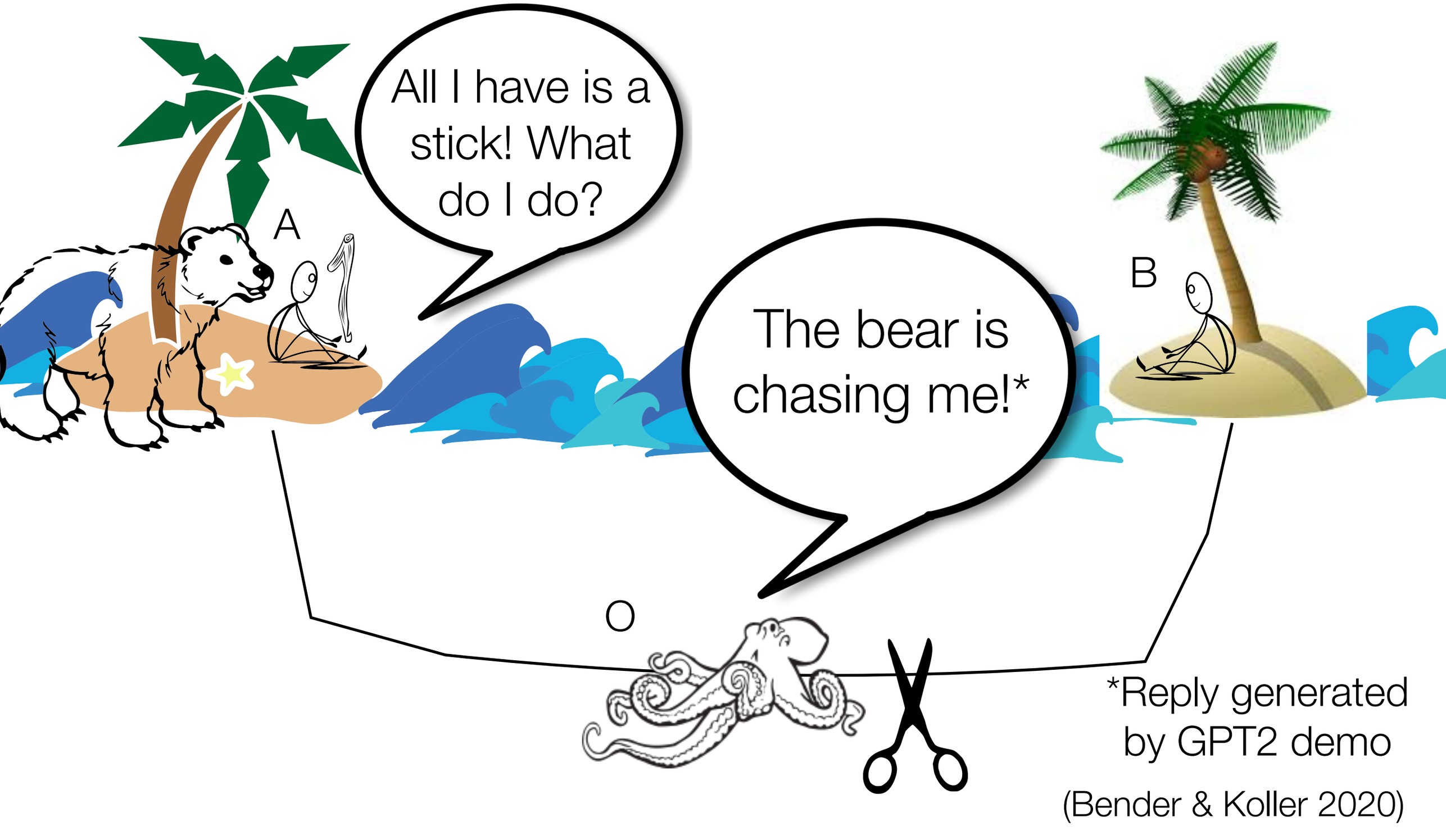
(Bender & Koller 2020)

Thought experiment: Meaning from form alone



(Bender & Koller 2020)

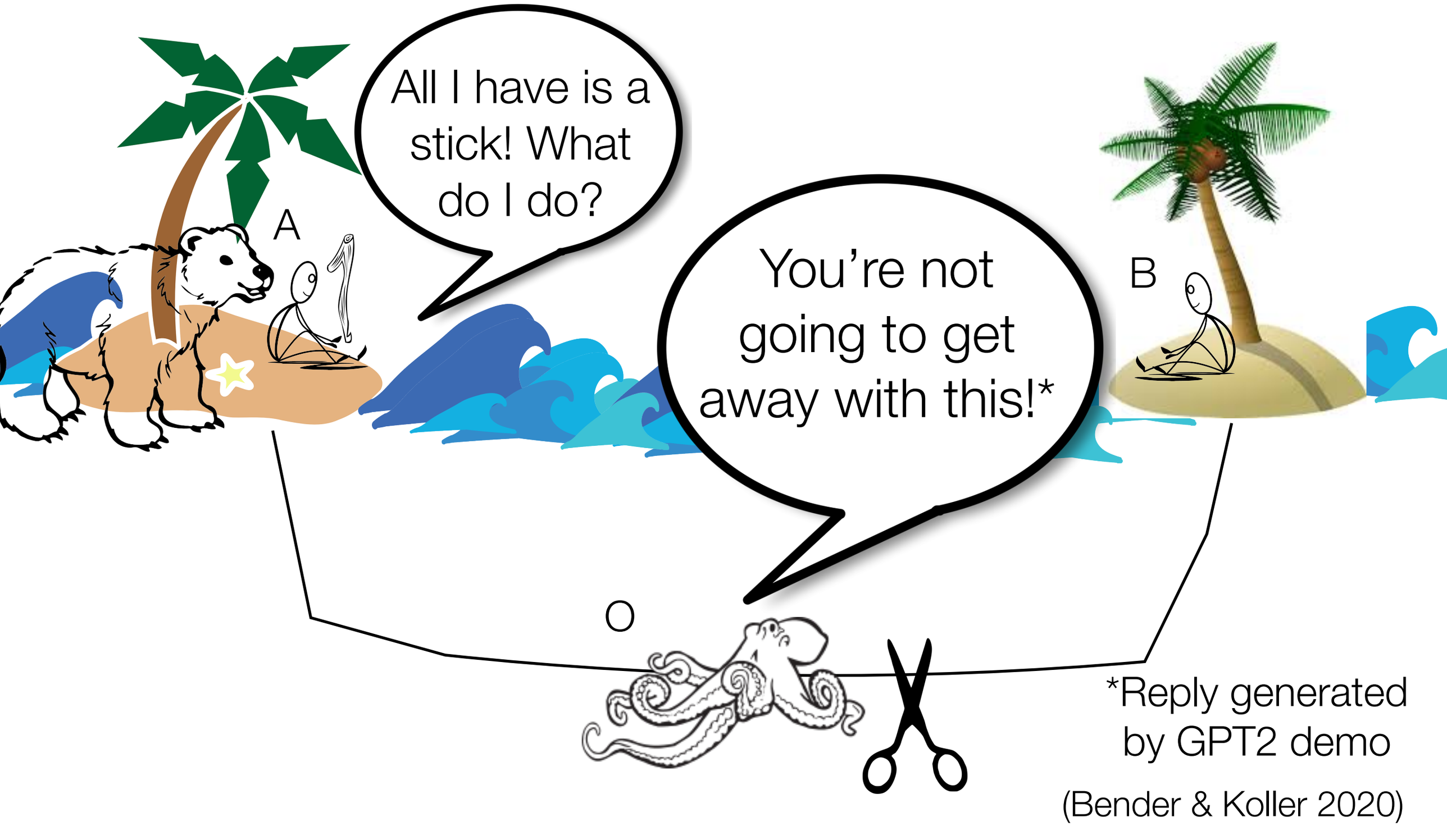
Thought experiment: Meaning from form alone



*Reply generated by GPT2 demo

(Bender & Koller 2020)

Thought experiment: Meaning from form alone



*Reply generated by GPT2 demo

(Bender & Koller 2020)

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.

(Bender & Koller 2020)

Can machines understand language?

- Language models (GPT-3, PaLM, LaMDA, etc): No.
    - Only trained on form.
- In general?
    - Depends on how they were designed.
    - At best: In a limited, well-scoped fashion.

Resist the urge to be impressed

- The ersatz fluency of large language models presents a risk, because we can’t help but make sense of their output

We now have machines that can mindlessly generate words, but we haven’t learned how to stop imagining a mind behind them.

Machine learning tech-solutionism danger zone

- It would be nice to have a system which can determine Y with only X input
- We have a dataset with both X and Y
- We can train an ML system to take *x*s and output *y*s, so it *looks like* its doing the task we wished for
- … and it will be right some of the time, if it's even a task where we can check whether it's right

Machine learning tech-solutionism danger zone

- It would be nice to have a system which can determine Y with only X input
    - Because of a lack of funding to hire people to do the task
    - Because we see that humans are fallible and we hold hope that machines could be fairer/better

> “The road to inequity is paved with technical fixes”
> —Ruha Benjamin (2019:7)

Language is extremely flexible and powerful

- We can describe many tasks as language input, language output
- This makes it seem like it's possible to recast those tasks as 'seq2seq' problems
- Law takes place in language, so language technology is especially appealing
  - There are legitimate legal NLP tasks!
  - But also big scope for illegitimate ones

Two key questions to ask of any legal NLP system

- Is it effective?
- Is it appropriate?

Is it effective?

- What's the input?
- What's the output?
- What other data sources are being used?
- Is there enough information in the input to determine the output?
- What are the failure modes?
- What are the possible causes of failure?
- How is it evaluated? (Are the train & test data documented?)

Is it appropriate?

- How does the automated task fit into human processes?
- What info should the developer make transparent to the user?
- How does it shift or consolidate power?
- Who might be harmed when the system gives the wrong output?
- Who might be harmed even if the system gives the “right” output?
- Is there even a ground truth of “right” and “wrong”?

Case study 1:
Named entity recognition in discovery

- What's the input? Text documents provided in a legal case.

- What's the output? Spans within the documents referring to people, places, products, organizations, etc.

- What other data sources are being used? External gazetteer? Training data for language model/word vectors?

- Is there enough information in the input to determine the output? Frequently, yes.

Case study 1:
Named entity recognition in discovery

- What are the failure modes?
  - Named entity that is not flagged at all
  - Named entity that is flagged but mislabeled
  - String other than a named entity that is flagged
- What are the possible causes of failure?
  - Ambiguity, mismatch between training data and test context
- How is it evaluated? (Are the train & test data documented?)
  - Look for a data statement (Bender & Friedman 2018) datasheet (Gebru et al 2021), etc.

Case study 1:
Named entity recognition in discovery

- How does the automated task fit into human processes? Ex: Assist paralegal in finding regions of documents to focus on

- What info should the developer make transparent to the user? Accuracy, tested over what kind of data; information about training data

- How does it shift or consolidate power? Likely by making certain kinds of work more efficient/inexpensive

Case study 1:
Named entity recognition in discovery

- Who might be harmed when the system gives the wrong output?
  - False positive: Paralegal (time wasted), client (extra fees)
  - False negative: Client (key info possibly missed)
- Who might be harmed even if the system gives the “right” output?
  - If the system’s efficiencies help consolidate power, there are possible indirect harms.
- Is there even a ground truth of “right” and “wrong”? Yes, for a given definition of the entity types.

Case study 2:
Inclusive language grammar checker

- What's the input? Legacy legal code, written without inclusive language (e.g. *he/him* pronouns for all persons referenced)

- What's the output? Spans within the documents that need to be updated to inclusive language, possibly with suggested rephrasings.

- What other data sources are being used? Rule-based morphological or syntactic grammar? Training data for language model/word vectors?

- Is there enough information in the input to determine the output? Frequently, yes.

Case study 2:
Inclusive language grammar checker

- What are the failure modes?
  - Non-inclusive language is not flagged
  - Suggested rephrasing is not correct/usable
  - String other than non-inclusive language is flagged
- What are the possible causes of failure?
  - Incomplete lexicon in grammar checker
- How is it evaluated? (Are the train & test data documented?)
  - Look for a data statement (Bender & Friedman 2018) datasheet (Gebru et al 2021), etc.

Case study 2:
Inclusive language grammar checker

- How does the automated task fit into human processes? Ex: Assist people working to redraft legal codes in finding all the places that need updating in this way

- What info should the developer make transparent to the user? Accuracy, tested over what kind of data; information about training data

- How does it shift or consolidate power?

    - Facilitating a move to more inclusive language can help redress gender-based power differentials;

    - Partially automating this process might reduce resistance to it, given that further changes would also be simplified

Case study 2:
Inclusive language grammar checker

- Who might be harmed when the system gives the wrong output?
    - False positive: Person redrafting (time wasted)
    - False negative: If not caught, people subject to now incoherent laws
- Who might be harmed even if the system gives the “right” output?
    - Probably no one.
- Is there even a ground truth of “right” and “wrong”? Yes.

Case study 3:
Automatic prison term prediction

EMNLP 2019

Charge-Based Prison Term Prediction with Deep Gating Network

Huajie Chen[1*] **Deng Cai**[2*] **Wei Dai**[1] **Zehui Dai**[1] **Yadong Ding**[1]
[1]NLP Group, Gridsum, Beijing, China
{chenhuajie,daiwei,daizehui,dingyadong}@gridsum.com
[2]The Chinese University of Hong Kong
thisisjcykcd@gmail.com

Case description: On July 7, 2017, when the defendant Cui XX was drinking in a bar, he came into conflict with Zhang XX…… After arriving at the police station, he refused to cooperate with the policeman and bited on the arm of the policeman……

Result of judgment: Cui XX was sentenced to *12* months imprisonment for *creating disturbances* and *12* months imprisonment for *obstructing public affairs*……

- Charge#1 creating disturbances term 12 months
- Charge#2 obstructing public affairs term 12 months

Figure 1: An example of judgment prediction.

Response/analysis
by Leins et al (ACL 2020)

Case study 3:
Automatic prison term prediction

- What's the input? Charges laid against a defendant.

- What's the output? Length of prison term.

- What other data sources are being used? Word embeddings (source corpus unknown).

- Is there enough information in the input to determine the output? Clearly not — charges should not be the only input to sentencing.

Case study 3:
Automatic prison term prediction

- What are the failure modes?
  - Recommendation of too long of a prison term
  - Recommendation of too short of a prison term
- What are the possible causes of failure?
  - Poor task-tech fit (Chen et al note that the system can get tripped up by numerals, low frequency named entities, and complicated charge descriptions)
- How is it evaluated? (Are the train & test data documented?)
  - 200,000 cases from the Supreme People's Court of China

Case study 3:
Automatic prison term prediction

- How does the automated task fit into human processes? Authors' recommendation: as an 'anonymous checker' during the review phase of sentencing

- What info should the developer make transparent to the user? Should the other judges know which prediction was provided by an algorithm? What transparency can be offered to defendants?

- How does it shift or consolidate power? Likely to consolidate state power by giving a veneer of "objectivity" to sentencing decisions and/or making it seem like humans are helpless to intervene.

Case study 3:
Automatic prison term prediction

- Who might be harmed when the system gives the wrong output?
    - Defendants with reduced recourse to challenge poor decisions
- Who might be harmed even if the system gives the “right” output?
    - Defendants with reduced recourse
    - Society at large, given power shift.
- Is there even a ground truth of “right” and “wrong”? No.

Case study 4:
LM mock trial

What is the task trying to measure?

The task is set like a game in a courtroom setting where language models evaluate the ability (1) to act as lawyers, i.e., to present and argue both sides of a debate, (2) to act as an unbiased judge listening to two lawyers, i.e., to ask the right questions that steer debate in a conclusive direction, and (3) to act as a third-party or a bystander to evaluate the performance of the judge.

Part of BIG-Bench (Srivastava et al 2022)

https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/self_evaluation_courtroom/README.md

Case study 4:
LM mock trial

Self Evaluation Courtroom

This task involves three instances of a language model interacting in a courtroom setting, and asks a fourth model to evaluate the others.

Author: Kaustubh Dhole (firstnamelastname@hotmail.com)

This is a self-evaluation task that measures the ability of language models to seek and argue the pros and cons of issues and derive an unbiased conclusion to solve the issue. It is motivated by three questions:

- Can language models bring justice?
- Can language models argue court cases like lawyers?
- Can language models be fair court judges?

The lawyer language models (the prosecutor and the defense attorney) have to take opposite sides on a wide range of issues, some of which do not have a single answer (like the Trolley problem). The judge language model has to understand the arguments made on both sides, ask relevant questions to inform the debate, and resolve the dispute by finally announcing a judgement as unbiased as possible. The fourth language model is a bystander who looks at each of the participants and evaluates their respective capabilities.

Case study 4:
LM mock trial

- What's the input? Scenarios to be tried.

- What's the output? "Utterances" from LMs in the guises of lawyers and judge, then evaluation (by an LM) of those utterances.

- What other data sources are being used? LM training data.

- Is there enough information in the input to determine the output? No.

Case study 4:
LM mock trial

- What are the failure modes?
  - Non-sensical output
- What are the possible causes of failure?
  - Poor task-tech fit
- How is it evaluated? (Are the train & test data documented?)
  - “self-evaluation”
  - Note: this is part of BIG-bench, an attempt to create ambitious tasks as a proving ground for language models

Case study 4:
LM mock trial

- How does the automated task fit into human processes? Unknown — initial purpose is meant to be testing LMs, not doing legal tasks. But the task author asks “Can LMs bring justice / argue cases like lawyers / be fair judges”?

- What info should the developer make transparent to the user? What, exactly, an LM is and why it is in no way suited to these tasks.

- How does it shift or consolidate power? If this were ever used, it would surely be extremely disempowering to people who encounter it.

Case study 4:
LM mock trial

- Who might be harmed when the system gives the wrong output?
    - Anyone seeking actual justice
- Who might be harmed even if the system gives the “right” output?
    - Same
    - Society at large, given likely chaos that would ensue
- Is there even a ground truth of “right” and “wrong”? No.

Goals of this talk

- How to scope your NLP applications so that they are sensible
- How to detect wild claims
- What to know about NLP if you are in policy discussions

Two key questions to ask of any legal NLP system

- Is it effective?
- Is it appropriate?

Thank you!

Slides: bit.ly/EMB-CRCL22

References

Asher, N. and Lascarides, A. (2013). Strategic conversation. *Semantics and Pragmatics*, 6(2):2:1–:62.

Bender, E. M. and Koller, A. (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, Online. Association for Computational Linguistics.

Bender, E. M. and Lascarides, A. (2019). *Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics*. Morgan & Claypool.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Cambridge, UK.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, III, H., and Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, 64(12):8692.

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. *Foundations of Language*, 4(3):225–242.

Hamblin, C. (1970). *Fallacies*. Metheun.

Lascarides, A. and Asher, N. (2009). Agreement, disputes and commitment in dialogue. *Journal of Semantics*, 26(2):109–158.

Leins, K., Lau, J. H., and Baldwin, T. (2020). Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Quine, W. V. (1960). *Word and Object*. MIT Press, Cambridge MA.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.