

Natural Language Processing with Language in Focus

Emily M. Bender
University of Washington
@emilymbender

Conversazioni Linguistiche
Università di Trento
16 April 2021



UNIVERSITÀ
DI TRENTO

Dipartimento di
Lettere e Filosofia

CeASUm

Centro di Alti Studi Umanistici

Laboratorio
Lingue e Linguaggio
LaLL



Conversazioni Linguistiche

Upcoming event:

Natural Language Processing with Language in Focus

Emily M. Bender

University of Washington, Department of Linguistics

Slides: <http://bit.ly/BenderTrento>

Talk outline

- Situating this talk; a bit about my trajectory
- Why build language technology?
- What linguistics has to contribute to language technology
 - Typology
 - Semantics & pragmatics
 - Child language acquisition
 - Sociolinguistics
 - Descriptive and documentary linguistics

Not exhaustive!

My journey into computational linguistics

- Discovered linguistics freshman year of university; AB (UC Berkeley), MA, PhD (Stanford) all in Linguistics
- First programming language: Logo (4th grade)
- First programming class: CS 60A @ Cal, in Scheme
 - Concurrently: Morphology with Prof. Sharon Hargus & TA David A. Peterson
 - First compiling project: Luganda morphological analyzer in Scheme



My journey into computational linguistics

- Grad school: Introduction to computational linguistics (Martin Kay), phenomenology (Terry Winograd)
 - RAship in grammar engineering, with Ivan Sag and Dan Flickinger
 - Dissertation (2001): *Syntactic Variation and Linguistic Competence: The Case of AAVE Copula Absence*
- No luck on the job market as syntactician or sociolinguist
- Short stint in industry (YY Technologies) as a grammar engineer for Japanese

My journey into computational linguistics



- While at YY, started the Grammar Matrix, in connection with Project Deep Thought
- After a couple more years of temporary positions, hired by UW Linguistics to start the CLMS program
- At the time: strong language group in EE working on MT & ASR (Mari Ostendorf, Jeff Bilmes, Katrin Kirchhoff)
- CSE had AI/IE folks, who worked with language data

Language *per se* vs.
Information encoded in language

Why build language technology?

- Learn something about language
- Build something of direct practical use
- As part of a broader program of AI

Learn something about language

- Annotation tools for corpora, to support linguistic research (Davies 2009, Meurer et al 2013, Kouylekov and Oepen 2014, Bender et al 2012)
- Computer-assisted transcription (language documentation, sociolinguistics) (e.g. Wassink et al 2018)
- Precision models of grammar
 - Linguistic hypothesis testing (Bierwisch 1963, Friedman et al 1971, Müller 1999, Butt et al 2002, Bender 2008, Fokkens 2014, Müller 2015, Zamaraeva 2021)
 - Language documentation (Bender et al 2013, Howell 2020)

Build something of direct practical use

- Automatic speech recognition
- Text-to-speech
- Machine translation
- Virtual assistants
- Search engines
- Spelling & grammar checkers
- Writing assistants
- Computer-assisted language learning

<https://www.aclweb.org/anthology/venues/ws/>

As part of a broader program of AI

- If machines can “do” language, does that prove intelligence? (Turing 1950)
- If machines can “do” language, can they learn lots of “world knowledge” and “common sense” and otherwise do self-teaching through interaction and/or machine reading?
- If a given machine learning algorithm can “solve language”, does that prove that it's a general purpose learning algorithm?



Problem: language-as-a-proving-ground-for-AI papers, and some work on practical language technology that takes an end-to-end approach, looks right through the language.

Solution: Keep the language in focus, and apply the lessons of linguistics.

Towards more multilingual NLP

- Bender 2009 “Linguistically naïve != language independent”
- Bender 2011 Dos & don'ts for language independent NLP, including:

Do state the name of the language that is being studied, even if it's English. Acknowledging that we are working on a particular language foregrounds the possibility that the techniques may in fact be language-specific. Conversely, neglecting to state that the particular data used were in, say, English, gives false veneer of language-independence to the work.



The #BenderRule

- “Always state the name of the language you are working on, even if it is English”
- Coined by (at least) Nathan Schneider, Yuval Pinter, Rob Munro & Andrew Caines



Emily M. Bender

@emilymbender



Dear Computer Scientists,

"Natural Language" is **not** a synonym
for "English".

That is all.
-Emily

9:32 AM - 26 Nov 2018

255 Retweets 1,132 Likes



The #BenderRule

- Why does this matter, if we always know it's English unless otherwise specified?
- Status quo: Work on non-English is “language specific”, work on English is “NLP”
- But English is just one language, like any other and not representative of all!
 - A window with its own specific pattern of raindrops



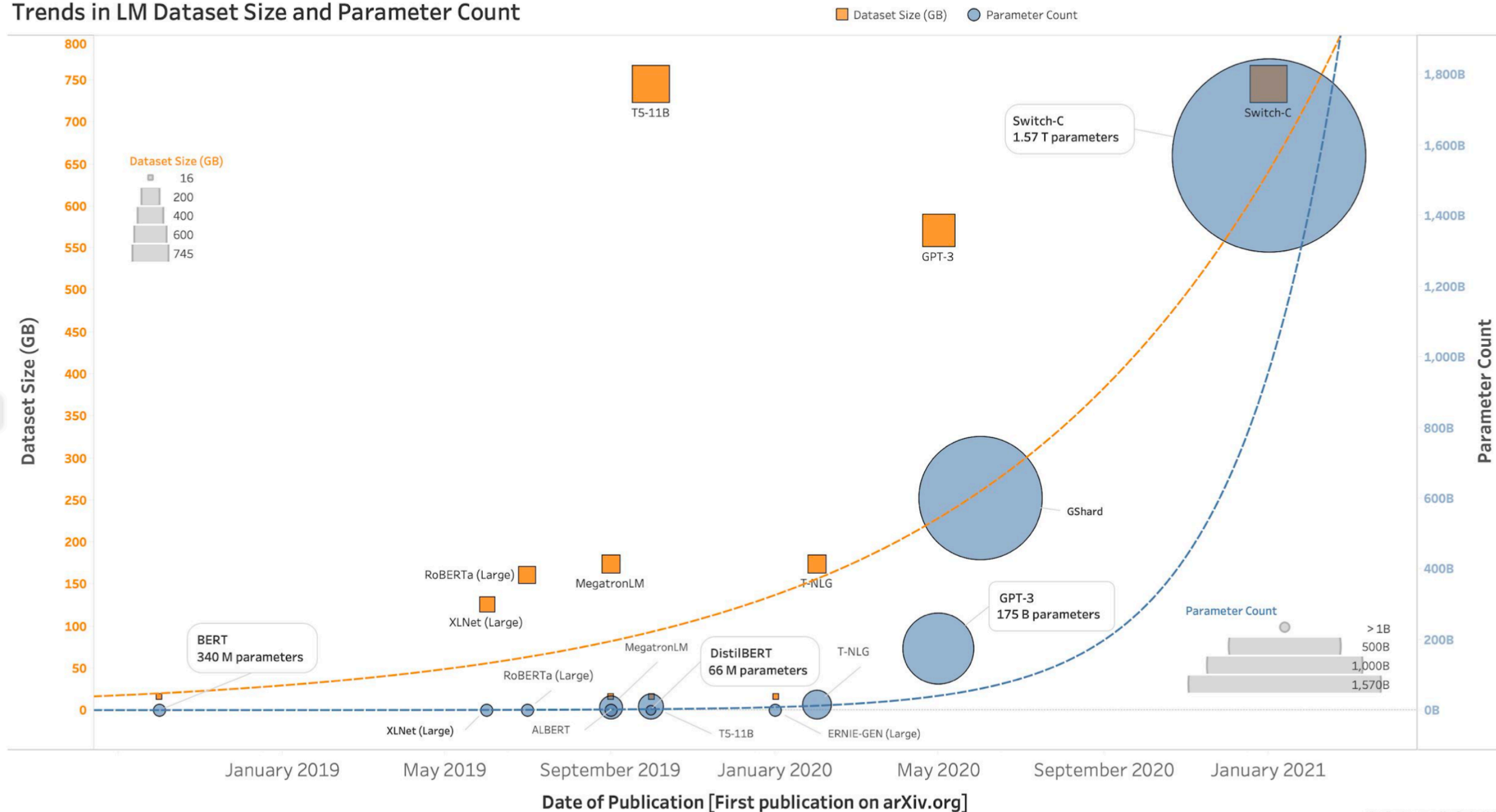
(Bender 2019)

How is English non-representative?

- It's a spoken language, not a signed language
- It has a well-established, long-used, roughly phone-based orthographic system
- ... with white space between words
- ... using (mostly) only lower-ascii characters
- It has relatively little morphology and thus fewer forms of each word
- It has relatively fixed word order
- English forms might 'accidentally' match database field names, ontology entries, etc.
- It has massive amounts of training data available (like the 3.3B tokens used to train BERT (Devlin et al 2019) or 499B tokens used to train GPT-3 (Brown et al 2020))

Large Language Models: Trends in dataset and parameter size (from Bender, Gebru et al 2021)

Trends in LM Dataset Size and Parameter Count



Thanks to Denise Mak for graph design

Linguistic Typology

- In what ways do languages vary, and what are the bounds on that variation? (Plank 2007)
- Linguistic interest:
 - What's where why? (Bickel 2007)
 - What is a possible human language and what does that tell us about human brains and human language development? (Slobin and Bowerman 2007)

Linguistic Typology

- In what ways do languages vary, and what are the bounds on that variation? (Plank 2007)
- Language technology interest:
 - Design NLP systems which are more portable across languages
 - Test NLP systems more thoroughly
- Promising developments:
 - ACL SIGTYP <https://sigtyp.github.io/>
 - <https://universaldependencies.org/> (Nivre et al 2020)

Understanding the relationship between form & meaning

- Form: text, speech, sign (+ paralinguistic information like gesture or tone)
- Conventional/standing meaning: logical form (or equivalent) that the linguistic system pairs with that form
- Communicative intent of the speaker: what they are publicly committed to by uttering that form (+ additional plausibly deniable inferences)
- Relationship between communicative intent & the world, e.g.:
 - True assertion, mistaken assertion, lie, accidentally true assertion, social act related to construction of social world, question about the interlocutor's beliefs, ...

Form/meaning/intent/world get flattened in NLP

- “Bag-of-words” approaches to NLP
- End-to-end approaches to meaning sensitive tasks: Mapping speech/text directly to machine actions
- Mistaking language modeling for understanding



Photo credit: NASA/NOAA

Language modeling

- Predicting linguistic form based on other linguistic form
 - Next word, given preceding sequence
 - Missing word, given surrounding context (“masked language models”)
 - Next sentence/sentence pair classification
- Can capture detailed information about word distribution and possibly also syntax
 - Super useful in many tasks, but not actually understanding

BERT fancub

- “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 fancub

- “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)

GLUE & SuperGLUE (Wang et al 2019a, b)

- Designed as tests for ‘natural language’ (actually English) understanding
- Key idea: A system that is really leveraging the linguistic system to understand should be able to apply that knowledge to different tasks
- Suites of multiple tasks
- ... including a ‘diagnostic’ task in GLUE designed to check for specific phenomena



GLUE/SuperGLUE sample tasks

Multi-sentence Reading Comprehension

(Kashabhi et al 2018)

MultiRC

Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*

Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)*

(Wang et al 2019b)



GLUE/SuperGLUE sample tasks

Commitment Bank

(de Marneffe et al 2019)

- (6) **John:** Tess was our star in the marathon this year. She's always trained with all her heart and soul. After all that training, she was happy to cross the finish line.
Prompt: Tell us how certain John is that Tess crossed the finish line.
- (7) **A:** Did you hear anything about Olivia's chemistry test?
B: Well, she studied really hard. But even after putting in all that time and energy, she didn't manage to pass the test.
Prompt: Tell us how certain speaker B is that Olivia passed the test.



GLUE as proving ground for language models

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	89.0	91.0	96.9	93.9	94.8	92.5
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	76.1	93.8	62.3	88.2	92.5	93.3
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

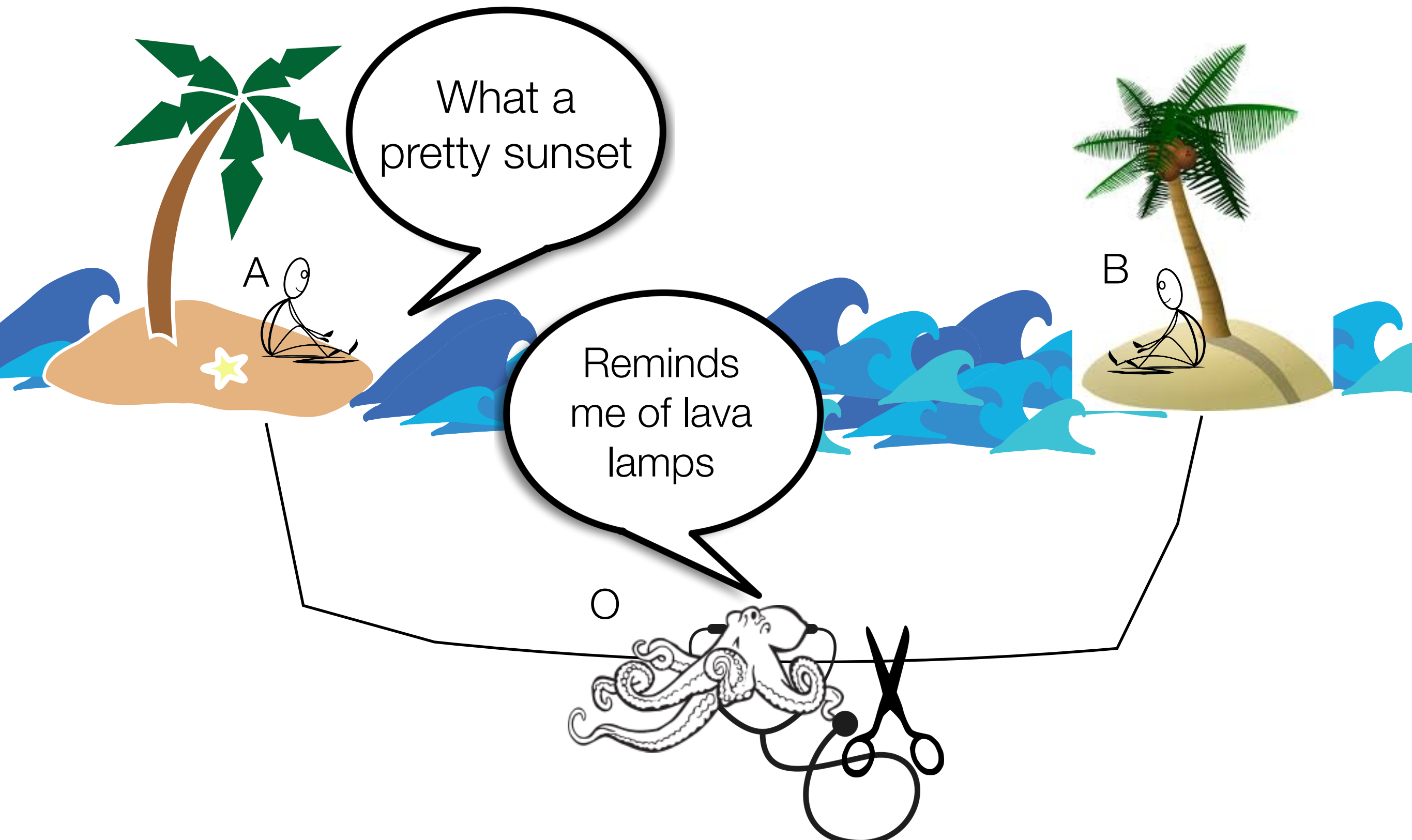
Table 3.8: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

(Brown et al 2020)



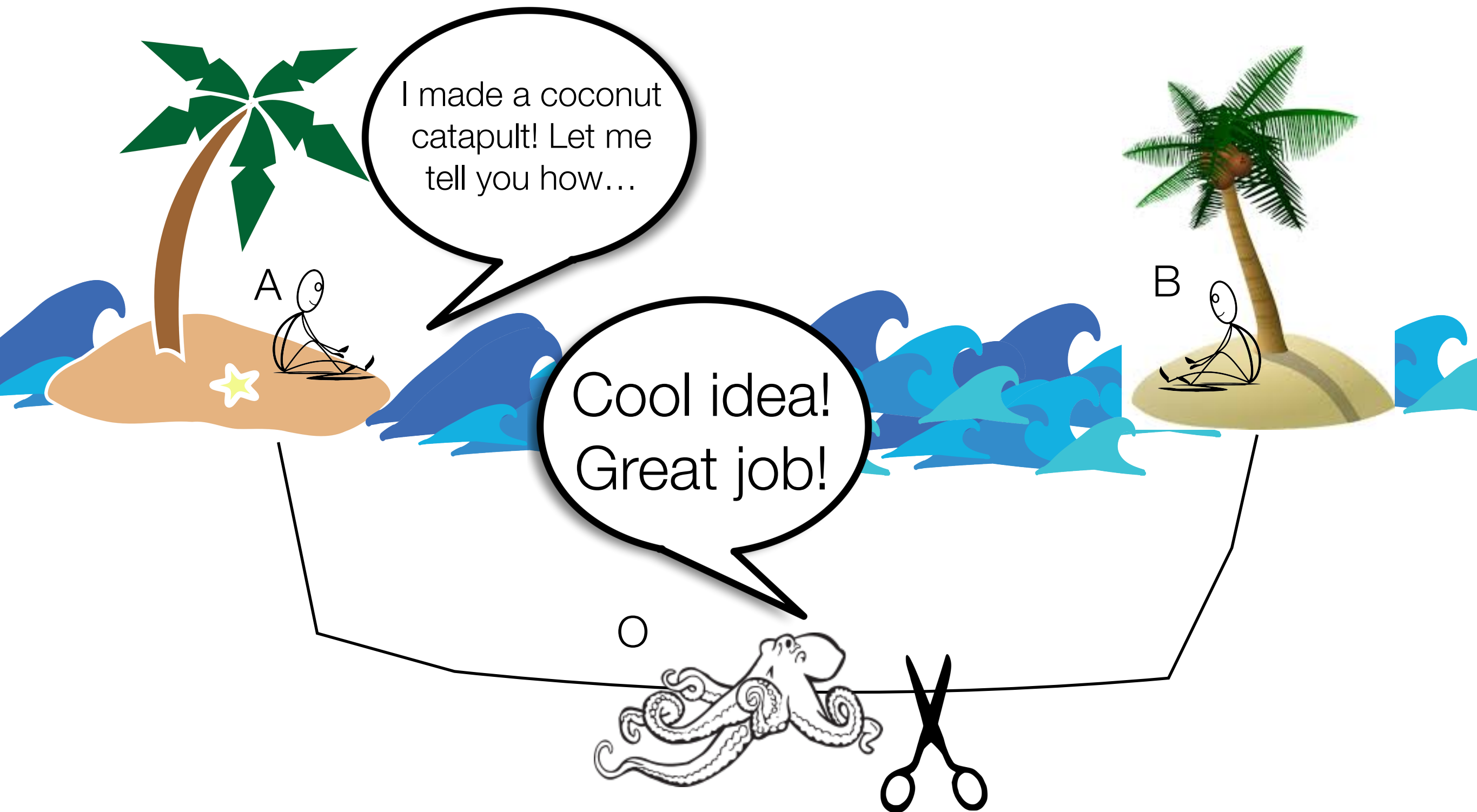
Thought experiment: Meaning from form alone

(Bender & Koller 2020)



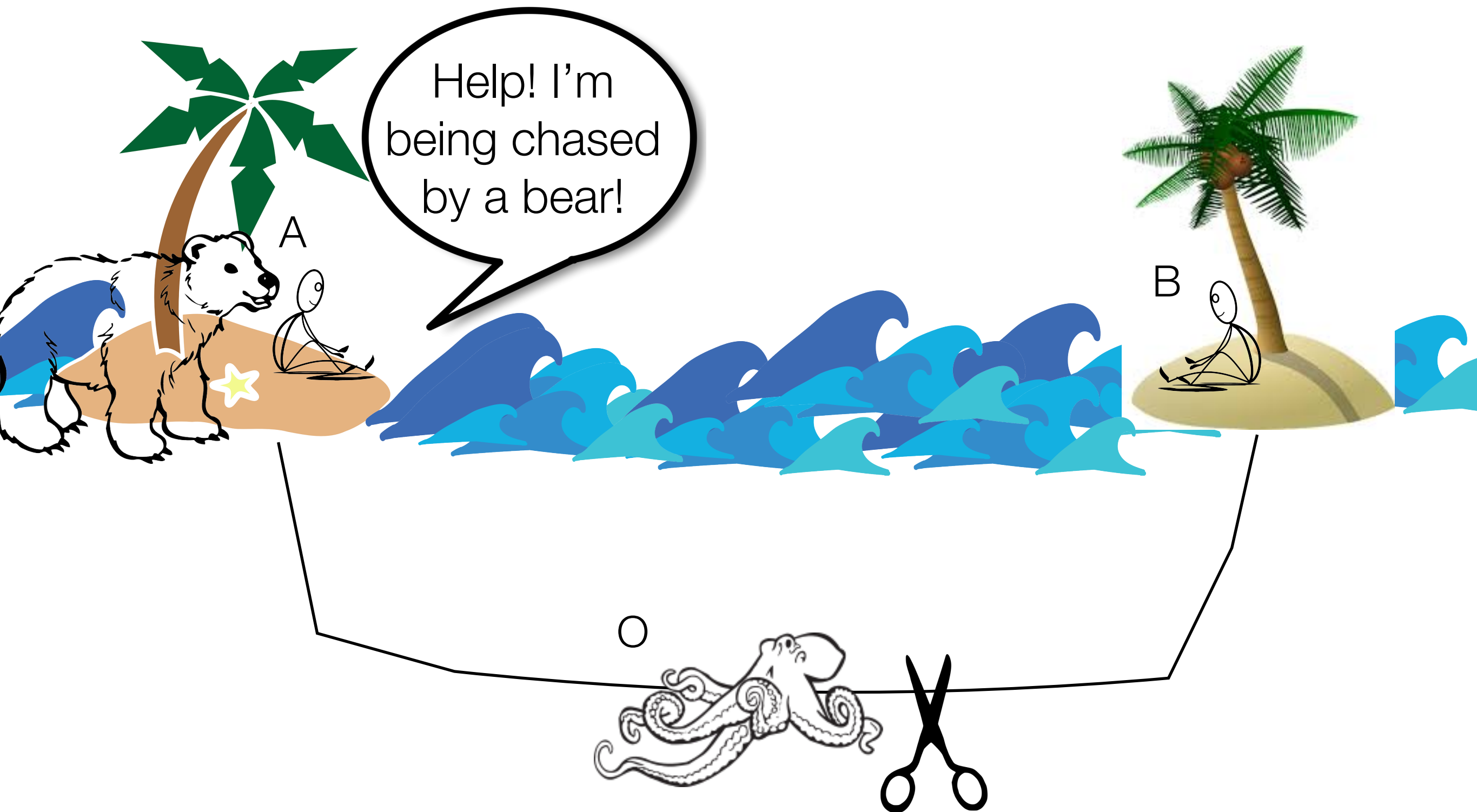
Thought experiment: Meaning from form alone

(Bender & Koller 2020)



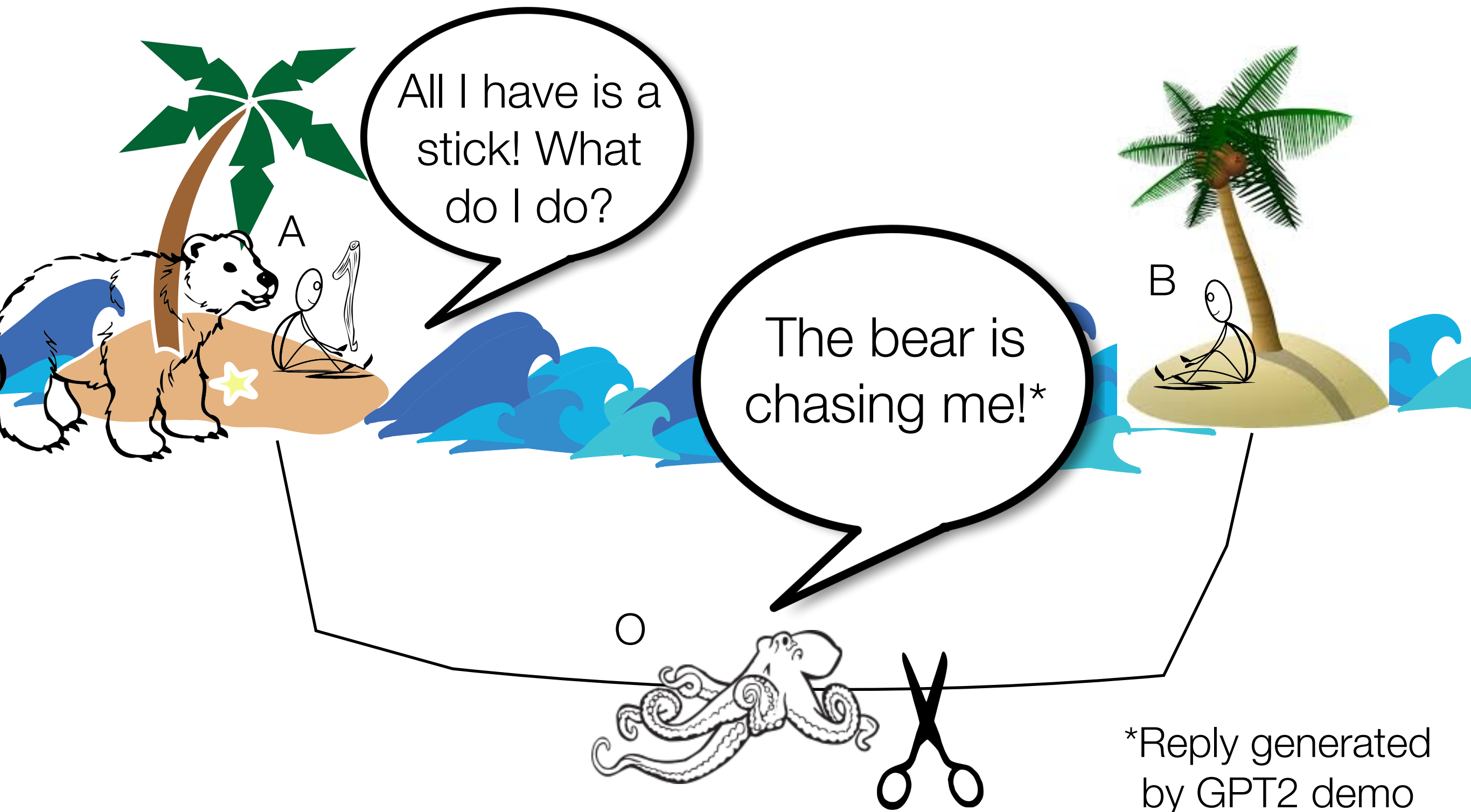
Thought experiment: Meaning from form alone

(Bender & Koller 2020)



Thought experiment: Meaning from form alone

(Bender & Koller 2020)



*Reply generated by GPT2 demo

Thought experiment: Meaning from form alone

(Bender & Koller 2020)



*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.

Understanding the relationship between form & meaning

- Form: text, speech, sign (+ paralinguistic information like gesture or tone)
- Conventional/standing meaning: logical form (or equivalent) that the linguistic system pairs with that form
- Communicative intent of the speaker: what they are publicly committed to by uttering that form (+ additional plausibly deniable inferences)
- Relationship between communicative intent & the world, e.g.:
 - True assertion, mistaken assertion, lie, accidentally true assertion, social act related to construction of social world, question about the interlocutor's beliefs, ...

So what are (large) language models learning?

- Fine-grained representations of word similarity both syntactic (Lin et al 2015) and semantic (Rubenstein and Goodenough 1965, Mikolov et al 2013)
- Structural phenomena like English subject-verb agreement (Goldberg 2019, Jawahar et al 2019)
- Constituent types, dependency labels, named entities, (core) semantic role labels (all in English; Tenney et al 2019)
- Something like unlabeled dependency structures (Hewitt and Manning 2019)
- ... but not any kind of sophisticated composition (Yu and Ettinger 2020)
- ... and lots of ‘short-cuts’ to getting the answer right (e.g. Niven & Kao 2019)

Language learning

- Frequently seen in discussions of machine learning: appeals to learning “without supervision” like babies do (e.g. Manning quoted in Andrews 2020)
 - Surely babies aren’t presented with corpora annotated with syntactic structure or word sense labels!
- But what do we know about how babies actually learn language?



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.



Language learning

- Frequently seen in discussions of machine learning: appeals to learning “without supervision” like babies do (e.g. Manning quoted in Andrews 2020)
 - Surely babies aren’t presented with corpora annotated with syntactic structure or word sense labels!
- But what do we know about how babies actually learn language?
- Machines don’t have to learn the same way, but knowledge of how language acquisition works in humans can inject realism into task design



Sociolinguistics

(e.g. Labov 1966, Eckert & Rickford 2001)

- Variation is the natural state of language
 - Variation in pronunciation, word choice, grammatical structures
- Status as ‘standard’ language is a question of power, not anything inherent to the language variety itself
 - Language varieties & features associated with marginalized groups tend to be stigmatized
- Meaning, including social meaning, is negotiated in language use
- Our social world is largely constructed through linguistic behavior


Sociolinguistics is critical to building equitable language technology

- *I choose to use this voice assistant, dictation software, machine translation system...*
 - ... but it doesn't work for my language or language variety
 - Suggests that my language/language variety is inadequate
 - Makes the product unusable for me

Sociolinguistics is critical to building equitable language technology

- *My screening interview was conducted by a virtual agent*
- *I can only access my account information via a virtual agent*
- *Access to a emergency response system requires interaction with a virtual agent first*
 - ... but it doesn't work or doesn't work well for my language variety
 - I scored poorly on the interview, even though the content of my answers was good
 - I can't access my account information or emergency services

Language use encodes stereotypes; ML can pick these up (to our peril)

- McConnell-Ginet (1984, 2020): Divergent paths of lexical semantic change of once parallel pairs like *buddy/sissy*, *master/mistress*, due to contexts of use and conversational dynamics
- Speer (2017): Tried building sentiment analysis system for English language restaurant reviews
 - Input: review text; Output: number of stars
 - System component: Word vectors from general web garbage
 - Problem: Underestimating stars assigned to Mexican restaurants
- Bender, Gebru et al (2021)  : Working at scales where we can document and account for whose language & whose world view is ingested is key

Language documentation, revitalization, learning and digital support

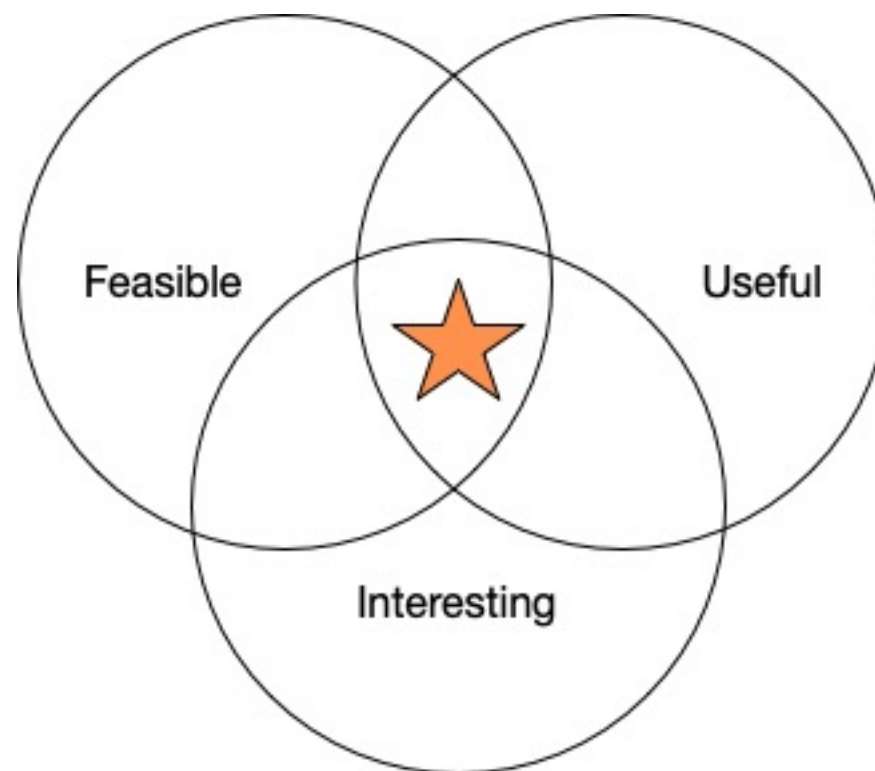
- Some excitement in NLP around low-resource languages as a proving ground for certain kinds of learning techniques
- Endangered languages are frequently among the most low-resource
- But what is actually needed, wanted, helpful?

What do communities need?

- Look to community-led projects
 - The ASL app (theaslapp.com) v. perennial “sign language gloves” (Erard 2017, Hill 2020)
 - firstvoices.com
 - lakotabears.com
 - yugtun.com
- ICLDC conference: a great meeting place for community members and others working on language documentation and conservation

What helps with language documentation and description?

- EL-STECC 2016 (NSF #1500157); STREAMLInED (NSF #1760475)
- Working meeting with field linguists and language technologists



What helps with language documentation and description?

- EL-STEAC 2016 (NSF #1500157); STREAMLINE (NSF #1760475)
- Working meeting with field linguists and language technologists
- Planned tasks (Levow et al 2017, 2021):
 - “Grandma’s hatbox”: speaker diarization, speaker ID, [genre ID, language ID, metadata extraction, transcription alignment]
 - Orthographic regularization
 - Auto-glossing of interlinear glossed text

Rule-based approaches: Case study of morphology

- Finite-state technology is up to the task of modeling natural language morphology (morphophonology + morphotactics) (Karttunen and Beesley 2005)
- Linguist-friendly tools exist for designing finite-state transducers (Beesley and Karttunen 2003, Hulden 2009)
 - Map surface forms to underlying strings of morphs or lemmas + tags
 - Bidirectional
 - Very efficient (time and memory)

Rule-based approaches: Case study of morphology

- Use cases:
 - spell checkers
 - dictionaries for morphologically complex (especially prefixing) languages
 - preprocessing of corpora for further analysis
 - generation of training data for ML systems (Schwartz et al 2019)
- Doable for most languages in 1-2 years, with linguistic expertise (Butt 2020)
 - Recent example: Strunk 2020 (on community impact, see <http://bit.ly/Strunk-KYUK>)

Talk outline

- Situating this talk; a bit about my trajectory
- Why build language technology?
- What linguistics has to contribute to language technology
 - Typology
 - Semantics & Pragmatics
 - Child language acquisition
 - Sociolinguistics
 - Descriptive and documentary linguistics

Not exhaustive!

From linguistics generally: Focus on data

- Understanding languages as fundamentally social entities, spoken by particular communities in particular times & places
- Viewing words and sentences as objects with internal structure and as members of larger systems
- Even when working ‘at scale’, the particulars of the data that make up a dataset matter
- Dataset documentation is valuable (Bender & Friedman 2018, Mitchell et al 2019, Gebru et al 2020)
 - See also: <https://sites.google.com/uw.edu/data-statements-for-nlp/>

Summary

- Linguistics can (and should) inform:
 - The design of language technology
 - The evaluation of language technology
 - The prudent and liberatory deployment of language technology



Keep in touch: @emilymbender
Slides: <http://bit.ly/BenderTrento>

Thank you!

References

- Andrews, E. L. (2020). How AI systems use mad libs to teach themselves grammar. *Stanford University HAI Blog*. <https://hai.stanford.edu/blog/how-ai-systems-use-mad-libs-teach-themselves-grammar>.
- Baldwin, D. A. (1995). Understanding the link between joint attention and language. In Moore, C. and Dunham, P. J., editors, *Joint Attention: Its Origins and Role in Development*, pages 131–158. Psychology Press.
- Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications, Stanford CA.
- Bender, E. M. (2008). Grammar engineering for linguistic hypothesis testing. In Gaylord, N., Palmer, A., and Ponvert, E., editors, *Proceedings of the Texas Linguistics Society X Conference: Computational Linguistics for Less-Studied Languages*, pages 16–36, Stanford. CSLI Publications.
- Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.
- Bender, E. M. (2011). On achieving and evaluating language independence in NLP. *Linguistic Issues in Language Technology*, 6:1–26.
- Bender, E. M. (2019). The #BenderRule: On naming the languages we study and why it matters. *The Gradient*. <https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/>.
- Bender, E. M., Crowgey, J., Goodman, M. W., and Xia, F. (2014). Learning grammar specifications from IGT: A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of FAccT 2021*.
- Bender, E. M., Ghodke, S., Baldwin, T., and Dridan, R. (2012). From database to treebank: Enhancing hypertext grammars with grammar engineering and treebank search. In Nordhoff, S. and Poggeman, K.-L. G., editors, *Electronic Grammaticography*, pages 179–206. University of Hawaii Press, Honolulu.
- 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.
- Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.
- Bierwisch, M. (1963). *Grammatik des deutschen Verbs*, volume II of *Studia Grammatica*. Akademie Verlag.
- Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Butt, M. (2020). Building resources: Language comparison and analysis. Keynote presentation at SIG-TYP2020, <https://slideslive.com/38939785>.
- Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The parallel grammar project. In Carroll, J., Oostdijk, N., and Sutcliffe, R., editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 1–7.
- Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.
- De Marneffe, M.-C., Simons, M., and Tonhauser, J. (2019). The CommitmentBank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

- de Saussure, F. (1959). *Course in General Linguistics*. The Philosophical Society, New York. Translated by Wade Baskin.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Eckert, P. and Rickford, J. R., editors (2001). *Style and Sociolinguistic Variation*. Cambridge University Press, Cambridge.
- Erard, M. (2017). Why sign-language gloves don’t help Deaf people. *The Atlantic*. <https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/>.
- Fedus, W., Zoph, B., and Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.
- Fokkens, A. S. (2014). *Enhancing Empirical Research for Linguistically Motivated Precision Grammars*. PhD thesis, Department of Computational Linguistics, Universität des Saarlandes.
- Friedman, J., Bredt, T. H., Doran, R. W., Pollack, B. W., and Martner, T. S. (1971). *A Computer Model of Transformational Grammar*. Elsevier, New York.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. (2020). Datasheets for datasets.
- Goldberg, Y. (2019). Assessing BERT’s syntactic abilities. *CoRR*, abs/1901.05287. <http://arxiv.org/abs/1901.05287>.
- Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hill, J. (2020). Do deaf communities actually want sign language gloves? *Nature Electronics*, 3(9):512–513.
- Howell, K. (2020). *Inferring Grammars from Interlinear Glossed Text: Extracting Typological and Lexical Properties for the Automatic Generation of HPSG Grammars*. PhD thesis, University of Washington.
- Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. Association for Computational Linguistics.
- Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Karttunen, L. and Beesley, K. R. (2005). Twenty-five years of finite-state morphology. *Inquiries Into Words, a Festschrift for Kimmo Koskeniemi on his 60th Birthday*, pages 71–83.
- Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., and Roth, D. (2018). Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
- Kouylekov, M. and Oepen, S. (2014). Semantic technologies for querying linguistic annotations. An experiment focusing on graph-structured data. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 4331–4336, Reykjavik, Iceland.
- Kuhl, P. K. (2007). Is speech learning ‘gated’ by the social brain? *Developmental Science*, 10(1):110–120.
- Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC.
- Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. (2020). Gshard: Scaling giant models with conditional computation and automatic sharding.
- Levow, G.-A., Ahn, E., and Bender, E. M. (2021). Developing a shared task for speech processing on endangered languages. In *Proceedings of the 4th Workshop on the Use of Computational Methods in*

- the Study of Endangered Languages*, pages 96–106.
- Levow, G.-A., Bender, E. M., Littell, P., Howell, K., Chelliah, S., Crowgey, J., Garrette, D., Good, J., Hargus, S., Inman, D., Maxwell, M., Tjalve, M., and Xia, F. (2017). STREAMLInED challenges: Aligning research interests with shared tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47, Honolulu. Association for Computational Linguistics.
- Lin, C.-C., Ammar, W., Dyer, C., and Levin, L. (2015). Unsupervised POS induction with word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1311–1316, Denver, Colorado. Association for Computational Linguistics.
- McConnell-Ginet, S. (1984). The origins of sexist language in discourse. In White, S. J. and Teller, V., editors, *Discourses in Reading and Linguistics*, pages 123–135. New York Academy of Sciences, New York.
- McConnell-Ginet, S. (2020). *Words Matter: Meaning and Power*. Cambridge University Press, Cambridge.
- Meurer, P., Dyvik, H., Rosén, V., De Smedt, K., Lyse, G. I., Smørdal Losnegaard, G., and Thunes, M. (2013). The INESS treebanking infrastructure. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 453–458, Oslo, Norway. Linköping University Electronic Press, Sweden.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.
- Müller, S. (1999). *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche*. Max Niemeyer Verlag, Tübingen.
- Müller, S. (2015). The CoreGram project: Theoretical linguistics, theory development and verification. *Journal of Language Modelling*, 3(1):21–86.
- Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Ohsugi, Y., Saito, I., Nishida, K., Asano, H., and Tomita, J. (2019). A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. *CoRR*, abs/1905.12848.
- Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Plank, F. (2007). Extent and limits of linguistic diversity as the remit of typology—But through constraints on WHAT is diversity limited? *Linguistic Typology*, 11(1):43–68.
- Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In Ortony, A., editor, *Metaphor and Thought*, pages 284–310. Cambridge University Press.
- Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.
- Schwartz, L., Chen, E., Hunt, B., and Schreiner, S. L. (2019). Bootstrapping a neural morphological analyzer for St. Lawrence Island Yupik from a finite-state transducer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 87–96, Honolulu. Association for Computational Linguistics.
- Slobin, D. I. and Bowerman, M. (2007). Interfaces between linguistic typology and child language research.

- Linguistic Typology*, 11(1):213–226.
- Snow, C. E., Arlman-Rupp, A., Hassing, Y., Jobse, J., Joosten, J., and Vorster, J. (1976). Mothers’ speech in three social classes. *Journal of Psycholinguistic Research*, 5(1):1–20.
- Speer, R. (2017). Conceptnet numberbatch 17.04: better, less-stereotyped word vectors. Blog post, <https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/>, accessed 6 July 2017.
- Strunk, L. (2020). A finite-state morphological analyzer for Central Alaskan Yup’ik. Master’s thesis, University of Washington.
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S., Das, D., and Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.
- Tomasello, M. and Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 57(6):1454–1463.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019b). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019a). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.
- Wassink, A., Squizzero, R., Fellin, C., and Nichols, D. (2018). Client libraries oxford (clox): Automated transcription for sociolinguistic interviews [computer software]. <https://clox.ling.washington.edu>.
- Yu, L. and Ettinger, A. (2020). Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.
- Zamaraeva, O. (2021). *Assembling Syntax: Modeling Constituent Questions in a Grammar Engineering Framework*. PhD thesis, University of Washington.