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

by
Emily M. Bender and Alexander Koller

Outline

- Introduction
- What is Meaning ?
- The Octopus Test
- On Climbing the Right Hills
- Conclusion

Introduction

Introduction

Recently we have a huge number of papers about language models like BERT or GPT-2 that claim that their model can “understand” natural language or captures “meaning”

Examples:

- In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task. (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)

Aim

Human-analogous natural language understanding is a grand challenge and a language model cannot learn "meaning" when it is trained only on form.

What is Meaning ?

What is Meaning ?

Form:

- It is any observable realization of language: marks on a page, pixels, or byte in a digital representation of text, or movements of the articulators

Meaning:

- It is the relation between the form and something external to language

What is Meaning ?

Communicative intent:

- It is something outside of a languages
- They are many types:
  - To convey some information to the other person
  - To ask a person to to do something
  - To socialize

Meaning:

- is the relation $M \subseteq E \times I$
- Which contains pairs *(e, i)* of natural language expressions *e* and the communicative intents *i*
- “Understand” refers to the process of retrieving *i* given *e*

What is Meaning ?

Conventional meaning:

- It is an abstract object that represents the communicative potential of a form

Linguistic system provides a relation:

- $C \subseteq E \times S$
- Which contains pairs *(e, s)* of expressions *e* and their conventional meanings *s*

Like the meaning relation *M, C* connects language to objects outside of the language

What is Meaning ?

Example:

- The photographer asked: “Please, can you do a Napoleon for the camera ?”

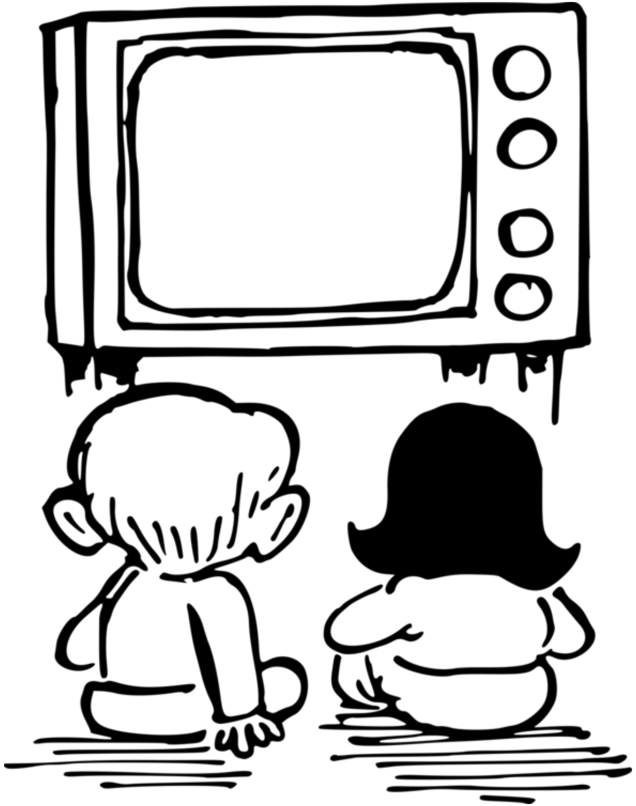
A natural language model that is trained purely on form will not learn meaning.

- There is not sufficient signal to learn:
  - The relation *M* between a form and the non-linguistic intent of human language users
  - The relation *C* between a form and the standing meaning the linguistic systems assigns to each form

To provide some signals to the model :

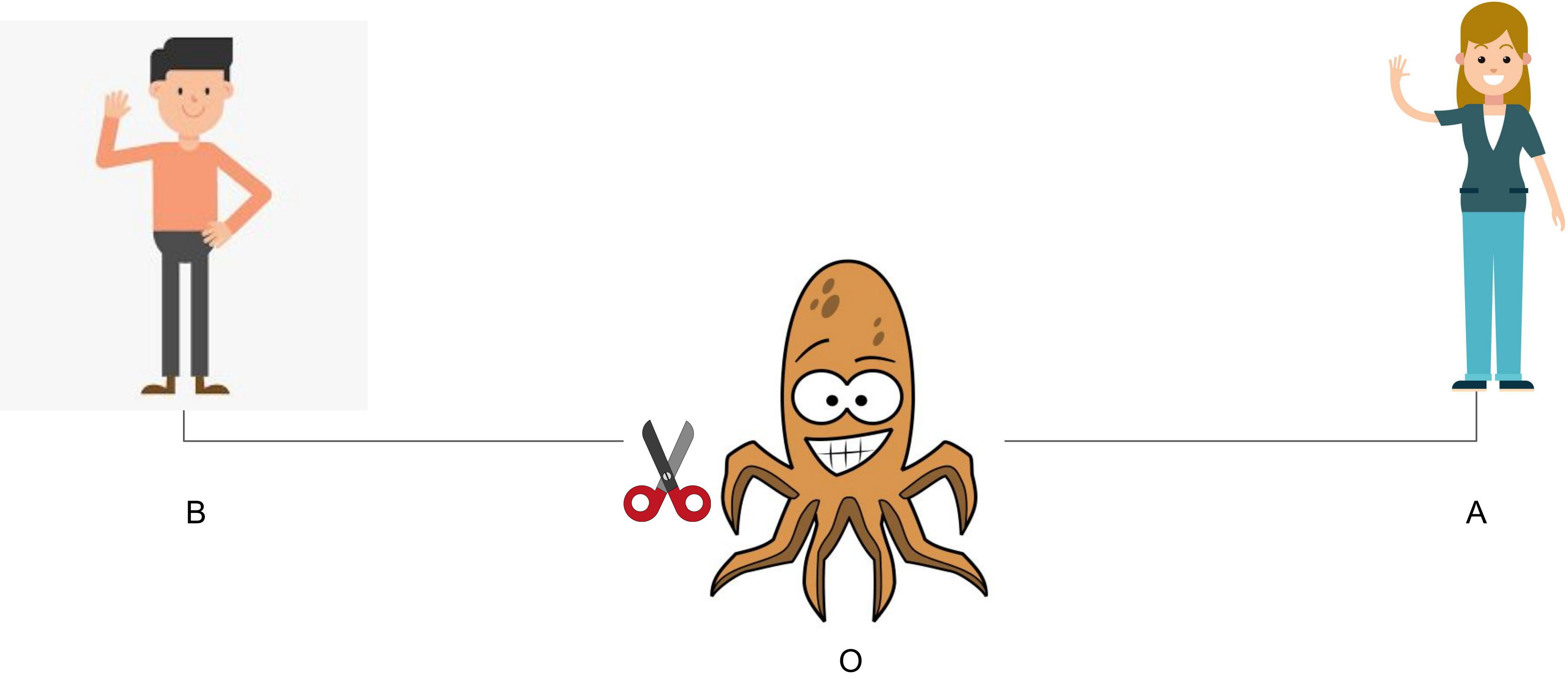
- We have to augment the data with perceptual data such as photos or other modalities
- Interaction data such as dialogue corpus

Human Language Acquisition



The Octopus Test

The Octopus Test



Experiments

Java Program:

- Imagine we trained a language model on all the well-formed Java code published on Github
- Input is only the code
- It is not paired with bytecode, nor a compiler, nor sample inputs and outputs for any specific program

We then ask the model to execute a sample program ???

We expect a correct program output

Right?

Wrong !!!

Experiments

GPT-2 on fighting a bear with sticks:

- **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** Go back to civilization. Take your gun...
- **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** #YouCanDoIt.” Police officers were quick to respond, finding a 3-foot female grizzly bear wandering off a dirt road near the shoreline...
- **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** Place these sticks over your head. Your heart will stop and you will not feel fear...

Experiments

GPT-2 on Arithmetic Operation:

- **Three plus five equals** seven," he replied. "Right, but I think that there should be more points added when our first family get in a power game, and even more when I start speaking like her mother!"...
- **Three plus five equals** ten" Bosh In this case, the conditions only need to be satisfied once, instead of...
- **Three plus five equals** five, right? Right? In the Catholic Church they say they don't look at church history, and no, I'm not going to say that I'm infallible either...

On Climbing the Right Hill

On Climbing the Right Hills

Bottom-up Perspective:

- A scientific results counts as a success if it solves a specific challenge

Top-down Perspective:

- The focus on the remote end goal of offering a complete, unified theory for the entire field

On Climbing the Right Hills

Hillclimbing Diagnostics:

- Ask top-down questions
- Be aware of the limitations of tasks
- Value and support the work of carefully creating new tasks
- Evaluate models of meaning across tasks
- Perform through analysis of both errors and successes

Conclusion

Conclusion

In this paper:

- Argued that model cannot learn the meaning from form alone
- It is call to use the proper language when talking about the success of language models
- Encouraged researchers to have a top-down perspective on NLP

Thank you !!!

Additional Slides

Some Possible Counterarguments

Counterarguments:

- “But” ‘meaning ‘ doesn’t mean what you say it means”
- “But meaning could be learned from …”
- “But there is so much form out there -- surely that is enough”
- “But aren’t neural representations meaning too?”
- “But BERT improves performance on meaning-related tasks, so it must have learned something about meaning”

References

References

1. Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained anal- ysis of sentence embeddings using auxiliary predic- tion tasks. In *Proceedings of ICLR*.
2. Dare A. Baldwin. 1995. Understanding the link between joint attention and language. In Chris Moore and Philip J. Dunham, editors, *Joint Attention: Its Origins and Role in Development*, pages 131–158. Psychology Press.
3. Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. ObjectNet: A large- scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche´ Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9453– 9463. Curran Associates, Inc.
4. Marco Baroni, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9(6):5–110.
5. Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. ArXiv preprint.
6. Ned Block. 1981. Psychologism and behaviorism. *The Philosophical Review*, 90(1):5–43.
7. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
8. Rechele Brooks and Andrew N. Meltzoff. 2005. The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543.
9. Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

References

10. Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantic Theory*, second edition, pages 493–522. Wiley-Blackwell.
11. Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg. Springer.
12. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.
13. Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.
14. Guy Emerson. 2020. What are the goals of distributional semantics? In*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.
15. Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics & Pragmatics*, 9(17):1–63.
16. Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
17. Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. ArXiv preprint.

References

18. H. Paul Grice. 1968. Utterer's meaning, sentence- meaning, and word-meaning. *Foundations of Language*, 4(3):225–242.
19. Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.
20. William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
21. Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.
22. Junxian He, Graham Neubig, and Taylor Berg- Kirkpatrick. 2018. Unsupervised learning of syntactic structure with invertible neural projections. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Brussels, Belgium. Association for Computational Linguistics.
23. Benjamin Heinzerling. 2019. NLP's Clever Hans moment has arrived. Blog post, accessed 12/4/2019.
24. Aurelie Herbelot. 2013. What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*, pages 321–327, Potsdam, Germany. Association for Computational Linguistics.
25. Aurélie Herbelot, Eva von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54, Avignon, France. Association for Computational Linguistics.

References

26. John Hewitt and Christopher D. Manning. 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.
27. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6):118:1–118:36.
28. Ganesh Jawahar, Benoˆıt Sagot, and Djame´ Seddah. 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.
29. Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. ArXiv preprint.
30. Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231– 236, Beijing, China. Association for Computational Linguistics.
31. Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal. Association for Computational Linguistics.
32. Alexander Koller, Konstantina Garoufi, Maria Staudte, and Matthew Crocker. 2012. Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 30–39, Seoul, South Korea. Association for Computational Linguistics.

References

33. Patricia K. Kuhl. 2007. by the social brain? Is speech learning 'gated' *Developmental Science*, 10(1):110–120.
34. Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.
35. Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. 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.
36. Sally McConnell-Ginet. 1984. The origins of sexist language in discourse. *Annals of the New York Academy of Sciences*, 433(1):123–135.
37. Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019.
38. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
39. Daniel McDuff and Ashish Kapoor. 2019. Visceral machines: Reinforcement learning with intrinsic physiological rewards. In *International Conference on Learning Representations*.
40. Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
41. Timothy Niven and Hung-Yu Kao. 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.

References

42. Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova´, Dan Flickinger, Jan Hajicˇ, and Zdenˇka Uresˇova´. 2015. SemEval 2015 Task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.
43. Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A simple but effective method to incorporate multi-turn con- text with BERT for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17, Florence, Italy. Association for Computational Linguistics.
44. Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.
45. Fabio Petroni, Tim Rockta¨schel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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.
46. W. V. O. Quine. 1960. *Word and Object*. MIT Press.
47. Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Open AI Blog, accessed 12/4/2019.
48. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

References

49. Michael J. Reddy. 1979. The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony, editor, *Metaphor and Thought*, pages 284–310. Cambridge University Press.
50. Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.
51. John Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457.
52. Catherine E Snow, Anjo Arlman-Rupp, Yvonne Hassing, Jan Jobse, Jan Joosten, and Jan Vorster. 1976. Mothers' speech in three social classes. *Journal of Psycholinguistic Research*, 5(1):1–20.
53. an Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.
54. Michael Tomasello and Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child Development*, 57(6):1454–1463.
55. Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.
56. Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche´ Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.
57. Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Ha- gen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investi- gating BERT's knowledge of language: Five anal- ysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

References

58. Joseph Weizenbaum. 1966. ELIZA—A computer program for the study of natural language communication between men and machines. *Communications of the ACM*, 9:36–45.
59. Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merrie ̈nboer, Armand Joulin, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *Proceedings of ICLR*.
60. Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
61. Ludwig Wittgenstein. 1953. *Philosophical Investigations*. MacMillan, New York.
62. Thomas Wolf. 2018. Learning meaning in natural language processing — The semantics mega-thread. Blog post, accessed 4/15/2020.
63. Dani Yogatama, Cyprien de Masson d’Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Ling- peng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. ArXiv preprint.