Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data (Bender and Koller, 2020)

Primary argument: Because LMs are only trained on *form*, they cannot in principle acquire *meaning*

- *Language model*: any system trained **only on the task of string prediction**
- *(Linguistic) meaning*: relationship between **linguistic form** and **communicative intent**

Aim of the paper: Advocate between alignment of claims and methodology

- "Climbing the right hill"
- We need to maintain clarity on notions such as *meaning* and *understanding* in task design and reporting of experimental results

1. **Terminology used to describe LMs can be problematic and leads to AI hype**
    - The use of words like "understanding" and "comprehension" wrt. LMs are either:
        - Overclaims, if meant to describe human-analogous understanding, or
        - Intended as technical terms, in which case should be explicitly defined
    - A consequence of this is the feeding of AI hype in popular press
    - So while we (as researchers) may understand that we use a different sense of "understanding" in research literature, this doesn't translate well in public communication
    - Part of the reason for this imprecise language is that we don't exactly know what about language LMs implicitly represent
        - "BERTology", LMs as psycholinguistic subjects
        - Overview of the literature highlights the extent to which LMs can learn aspects of formal linguistic structure (e.g. agreement, dependency structure), and their apparent ability to "reason" is built upon artifacts (of linguistic form) in the training data

2. **Now onto the spicy stuff: What is meaning?**
    - *Form*: any observational relation of language
    - *Meaning*: relation between form (natural language expressions) and something external to language—communicative intent
        - Humans use language in order to achieve some communicative intent
        - *Understanding*: process of retrieving intent given a natural language expression
    - Communicative intent is grounded in the real world
        - "You can't learn language from the radio"
        - It is distinct from *conventional (standing)* meaning, or the communicative potential of a form
    - A speaker has some intent and chooses an expression with some conventional meaning to express it in some communicative situation; a listener reconstructs this intent using their own knowledge about the situation and their hypotheses about the speaker's state of mind
    - Humans are very willing to attribute communicative intent to a linguistic signal, even when the origin of the signal is not an entity that could have communicative intent
        - Especially artificial agents, like ELIZA

- **Form alone is not a sufficient signal to learn the relation between that form and the non-linguistic intent, nor between the form and conventional meaning**
    - (Personal note: I think I agree with the first part of this statement, but not necessarily the second. Curious to know other's opinions on this.)
- The *symbol grounding problem* (+Chinese Room)

3. **The Octopus Test Saga: Sparknotes edition**
    - A and B are people on an island communicating in English through an underwater cable; O is a hyper-intelligent octopus who taps into their cable and intercepts their messages
    - O knows nothing about English initially, but is very good at picking up statistical patterns in their messages
    - O starts feeling lonely and cuts the cable, pretends to be B. Can O successfully pose as B without making A suspicious?
    - The extent to which O can do this is highly dependent on the task:
        - A and B talk frequently about trivial things in their daily lives; it seems possible that O could do this successfully "because the utterances in such conversations have a primarily social function, and do not need to be grounded in…the physical situation"
        - But what if A invents a coconut catapult and starts referring to objects like *ropes* and *coconuts*? How could O possibly understand what A is referring to (other than making comparisons to their distributions to other words like *mangos* and *nails*)?
            - Even without understanding, O could respond by just saying "Cool idea, great job!", and A would accept this because they have done all the work in attributing meaning to O's response
            - Q: Humans also talk about a great deal of things that are not grounded at our level of perception (particle physics, abstract ideas, ESP), as well as hypothetical objects and events. How is this any different from what O experiences?
        - Let's take it one step further: What if A has an emergency, like A is being chased by an angry bear? Can O give A suitable advice? Unfortunately no, so A is eaten.
    - Main point: "Without access to a means of hypothesizing and testing the underlying communicative intents, reconstructing them from the forms alone is hopeless"
        - O only fools A because A is such an active listener: "it is not that O's utterances make sense, but rather, that A can make sense of them"

More constrained thought experiments

- Bring up code since the form-meaning mappings are far less ambiguous than natural language
- "A system trained only on the form of Java or English has no way to learn their respective meaning relations"
- I wonder what the author's takes on things like Code LMs are, given recent progress on that front

4. **So what do babies do?**

- Not only requires grounding in the physical world, but interaction with people as well
- Cites a bunch of work in language acquisition literature:
  - They don't pick up passive exposure to a language, such as by listening to TV or radio (Snow et al. 2017; Kuhl 2007)
  - Joint attention (by the child and other communicator, like a caregiver) to the same object seems to play a crucial role (Baldwin 1995; Tomasello and Farrar 1986; Brooks and Meltzoff 2005)
- "Human children don't learn meaning from form alone and we should not expect machines to do so either"

5. **A side of distributional semantics** ("You shall know a word by the company it keeps")
   - Grounding has been an issue in distributional semantics for. A while.
     - Lexical similarity between words doesn't actually connect their concepts to the real world, and distributions in text don't necessarily match distributions in the world
   - So how can we provide grounding?
     - Augment with perceptual data
       - Q: Don't we have the same issue with images? *Ceci n'est pas une pipe*
     - Interaction data
       - Cues for emotional stress, eye gaze

6. **Best practices for less error-prone NLP/comp-lingy mountaineering**
   - NLP and comp-ling have had many a research fashion cycle (ever increasing in pace). How can we develop some foresight for whether we are making progress in the right direction?
   - They argue for a top-down perspective
   1. Develop a humility towards language
   2. Be aware of task limitations
   3. Value and support the work of carefully creating new tasks
   4. Evaluate models of meaning across tasks

7. **Possible counterarguments**
   - "But 'meaning' doesn't mean what you say it means."
   - "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."

Can Machine Learning Provide Understanding?

How Cosmologists Use Machine Learning to Understand Observations of the Universe

- This article discusses the how the use ML models as part of the scientific investigation process affects the scientific understanding we get from it
- It is based on one particular scientific discipline that has started to use ML: **cosmology**

- It starts by discussion the concept of **black-box**
    - Distinction between the two senses for the term black-box
        - **Black-box systems:** systems whose internals are unobservable, and where we can only study behavior
            - *"In our daily lives we are confronted at every turn with systems whose internal mechanisms are not fully open to inspection, and which must be treated by the methods appropriate to the Black Box"* (Ashby 1956, 86).
        - **Black-boxing as a methodology (The Method of Ignoration)**: ignoring mechanisms that are irrelevant to the object of study
            - Using *abstractions* to ignore lower-level details
                - Eg: treating cells a units rather than as processes of proteins
            - Delegating to other disciplines
- Introduces some background on cosmology:
    - <u>Problem:</u> How do cosmological clusters/structures arise?
        - Can they be predicted from gravitational theory?
        - Simple model:
            - Assume (dark) matter distribution is homogeneous fluid
            - Add perturbations to density across space
            - Study evolution
        - However simple "linear" model cannot deal with large perturbations to density
            - Necessary for complex structure
            - What do they mean by linear?
    - More complex simulations are instead used
        - **N-body simulations**
            - "discretize" matter into *N* particles
            - run (theory-based) simulation, assuming values for cosmological constants
            - compare resulting structures with observed ones

                - *"these techniques are not intended to represent the motion of a discrete set of particles. The particle confguration is itself an approximation to a fuid"*
            - By comparing observed structures with the ones obtains from simulations, we can *infer* values for cosmological constants
                - MCMC with distribution over parameters
            - However, N-body simulations are <u>expensive</u>
                - **Simulation** is expensive
                - **Posterior Inference** is expensive
        - Some discussion on what cosmologists learn from N-body simulations
            - Simulations assume simplified physical models
                - Only model gravity
                - Lower-level details are intentionally ignored
                    - "minimalist idealization"
                    - The method of ignoration
            - What is explanatory power of these simulations
                - Type I & II "*why*" questions
                    - <u>Type I</u>: why a phenomenon occurred in some particular circumstance
                        - "Why does our universe have the particular statistical distribution of matter that it does?"
                    - <u>Type II:</u> why questions ask why phenomena of this general type occur across a variety of circumstances
                        - "Why does the universe exhibit structures across a variety of cosmological parameters?"
                - Authors argue that N-body simulations answer both questions
                    - Inferred cosmological parameters answer type I
                    - Emergence of structures across wide range of parameters show density perturbations+gravitation answers type II

- The authors then move to how ML has been used in Cosmology
    - Background history
    - How ML has helped with the **simulation** cost: the use of **emulators**
        - ML models fitted to the "simulator function"
            - Extract training set by running similar many times across wide range of parameters
    - Two case studies
        - `The Cosmic Emulator`
            - Gaussian Process model
            - Has five """parameters"""

            - Not really parameters (GPs are non-parametric), more like inputs
        - Achieved 1% error with only 37 inputs
    - `PkANN`
        - Neural Network model
        - Better accuracy, but less sample-efficient
- Discussion on what cosmologist can learn when using **emulators**
    - How can we learn anything if the **emulator** is not modeling the relevant physics?
    - Authors argue that we can still use emulators by *the principle of ignoration*
        - We don't need to know all the details of how computer work to understand an algorithm
    - Machine Learning algorithms cannot answer Type II questions, but can answer Type I
        - To know why gravitation leads to clusters, one needs to model gravity and see clusters arising across simulations
        - However, we can still use emulators to infer the specific configuration of parameters that lead to the structure of the current universe
            - No difference between *emulators* and *simulators* for type I questions
    - Analogy to the three types of transparency of code:
        - **Functional:** *"knowledge of the algorithmic functioning of the whole"*
        - **Structural:** *"how a particular algorithm is instantiated in the code"*
        - **Run:** *"knowledge of how the program was run in a particular instance"*
            - Hardware, training data, etc...
    - `The Cosmic Emulator` & `PkANN` have all three kinds of transparency:
        - Easy to show that both have *function* and *structural* transparency
            - Functional: Gradient Descent
            - Structural: PyTorch/Autograd
        - *Run* transparency is a trickier issue since it depends on the training data
            - Are the samples used to train sufficient for a good model?
            - Adversarial cases, OOD generalization, etc...
        - However, given good evaluation results of the emulators, evidence is needed for why they aren't run transparent

- Author concludes arguing that, while MLs doesn't necessarily provide understanding in all contexts, there conditions in which ML can provide explanatory understanding
    - These conditions somewhat misspecified?
    - *"in contexts where a ML algorithm or an ANN is answering causal questions and there are no underlying structural simulations to appeal to"*

Physics Informed Machine Learning

- Modelling and forecasting the dynamics of multiphysics and multiscale systems is very difficult
    - Simulations are expensive
    - Inolve solving PDEs
    - Data isn't good enough
- Potential for application of ML is high
    - Can automatically extract features
    - Can model high dimensional nonlinear correlations
- Just naively throwing data at a large model has not worked
    - Physically impossible solutions
    - Bad generalization
- Physics Informed Machine Learning is needed
    - Leveraging prior knowledge to make ML systems better
- Most interesting setting
    - Some physics laws are known
    - Several scattered observations

Three ways of learning:

1. Observational bias: Learn from the data
2. Inductive bias: hard code invariances and properties into the architecture
3. Learning bias: soft constraints in the loss

Theoretical connections to well understood methods

- Kernel methods: Can be seen a kernel regression on the initial layer(s)
- Numerical methods: there are equivalences and mappings between simple architectural units and techniques to methods in physics

Benefits of using PNNs

- Incomplete physics and data is okay
- Low data is ok
- High dimensionality deal with well
- Uncertainty Quantification?? B-PNN gave error bars that scaled with the error,

Limitations

- Multiscale and multiphysics: e.g. frequency bias in NNs hurts when component frequencies vary
- Training methods suffer with multiobjective targets, no clear way to model select
- Data, Benchmarks, Optimization
- The need for theoretical results: e.g: A fundamental question is can a network find solutions to PDE via gradient-based optimization?

Future ideas:

- Fusion of models
- Interpretibility
- Useful representations