▼ *15주차 예습과제*

1. Extremely large models and GPT3

   about GPT3

   -Better than other models at language modeling and related tasks such as story completion

   -GPT-3 demonstrates some level of fast adaptation to completely new tasks.

   -The language model training(outer loop) is learning how to learn from the context

   limitations and open questions about GPT-3

   -Seems to do poorly on more structured problems that involve decomposing into atomic / primitive skills

   -Performing permanent knowledge updates interactively is not well studied

   -Doesn't seem to exhibit human like generalization (systematicity)

   -Language is situated and GPT-3 is merely learning from text without being exposed to other modalities.

2. Compositional Representations and Systematic Generalization

   Systematicity: The ability to produce/understand some sentences is intrinsically connected to ability to produce/understand certain others. This means there is a "definite and predictable pattern among the sentences we understand"

   E.g. any speaker that understands the sentence "John love Mary" should be able to understand "Mary loves John"

   -Compositionality of representations is a helpful prior that could lead to systematicity in behavior.

-Do neural networks generalize systematically on challenging benchmarks involving realistic language?

- Basic Machinery for producing compositionally challenging splits

  Let $\mathcal{F}_A$(data) $\equiv$ normalized frequency distribution of atoms

  Let $\mathcal{F}_C$(data) $\equiv$ normalized frequency distribution of compounds

  Define atom and compound divergence as:

  $$\mathcal{D}_A(\text{train}\,||\,\text{test}) = 1 - C_{0.5}(\mathcal{F}_A(\text{train})\,||\,\mathcal{F}_A(\text{test}))$$
  $$\mathcal{D}_C(\text{train}\,||\,\text{test}) = 1 - C_{0.1}(\mathcal{F}_C(\text{train})\,||\,\mathcal{F}_C(\text{test}))$$

  where,

  $$C_\alpha(P\,||\,Q) = \sum_k p_k^\alpha q_k^{1-\alpha}$$

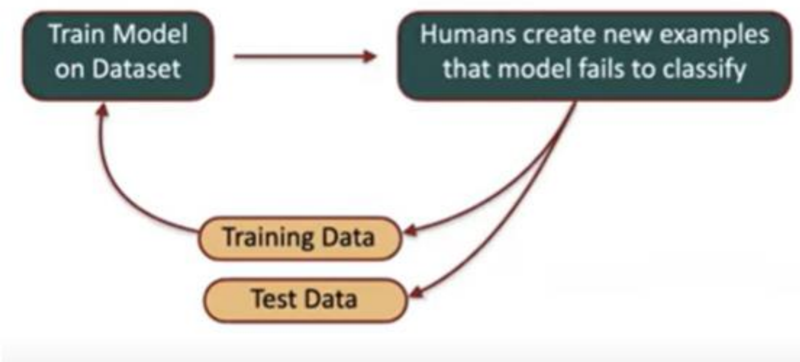
  is the chernoff coefficient between two categorical distributions that measures similarity.

- goal: split data into train/test such that compound divergence is maximized and atom divergence is minimized

3. Improving how we evaluate models in NLP

   While we are making progress in terms of perfromance on benchmarks, it's unclear if the gains are coming from spurious correlations or real task understanding

   -Instead of testing models on static benchmarks, evaluated on an ever changing dynamic benchmark.

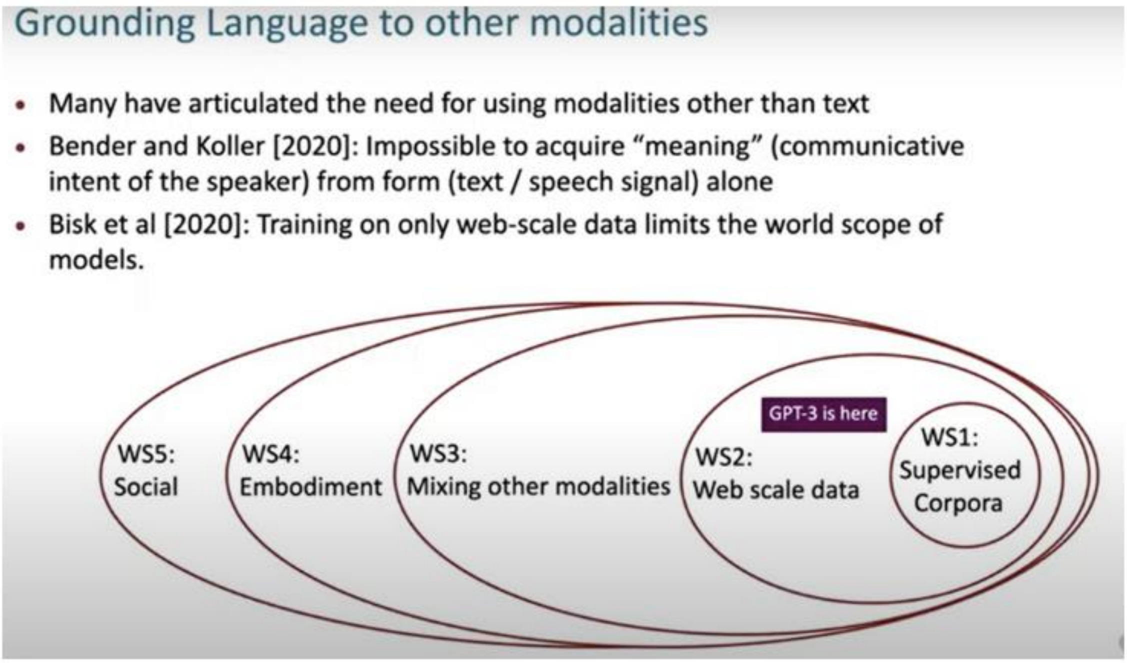


Overview of dynamic benchmarks

   -main challenges: ensuring that humans are able to come up with hard examples and we are not limited by creativity.

   -current approaches use examples from other datasets for the same task as prompts

4. Grounding language to other modalities



[serious progress in the last decade thanks to data+hardware+neural networks]

we now have amazing technologies such as GPT-3 that can do truly exciting things

in the short term:

- Scaling helps, so perhaps even larger models?
- Scaling requires very non-trivial enginerring efforts so a lot of interesting systems work to be done here

in the long term:

- making progress towards systematicity, fast adaptation, generalization
- Improved evaluation so we can trust benchmarks
- Figuring out how to move beyond text in a tractable way