Experience Grounds Language

Yiqi Tang

World scopes(a roadmap of Language Interface)

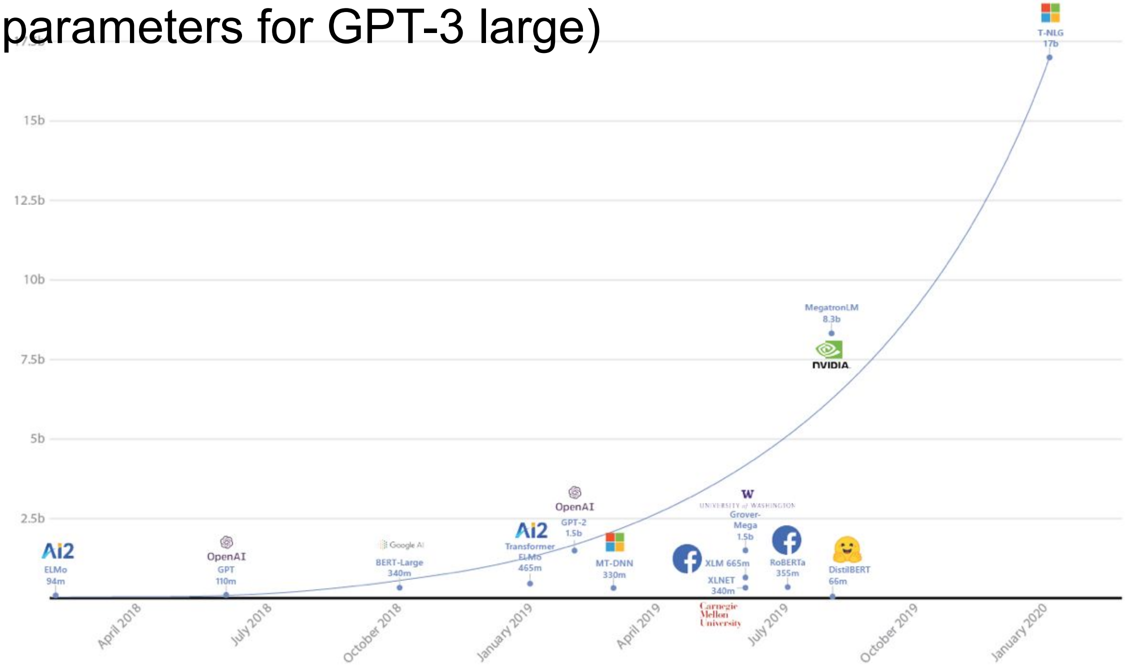
- WS1. Corpus *(our past)*
- WS2. Internet *(our present)*
- WS3. Perception
- WS4. Embodiment
- WS5. Social

WS1. Corpus *(our past)*

- Subset of naturally generated language, processed and annotated for the purpose of studying representations
- Dataset example：the Penn Treebank
- Works completely rely on corpora
- Methods from Baum-Welch to LSI (define words as distribution over clusters)

WS2. The Written World *(our present)*

- unstructured, unlabeled, multi-domain, and multilingual data
- transfer learning enabled by representations in deep models
- require scale (175B parameters for GPT-3 large)



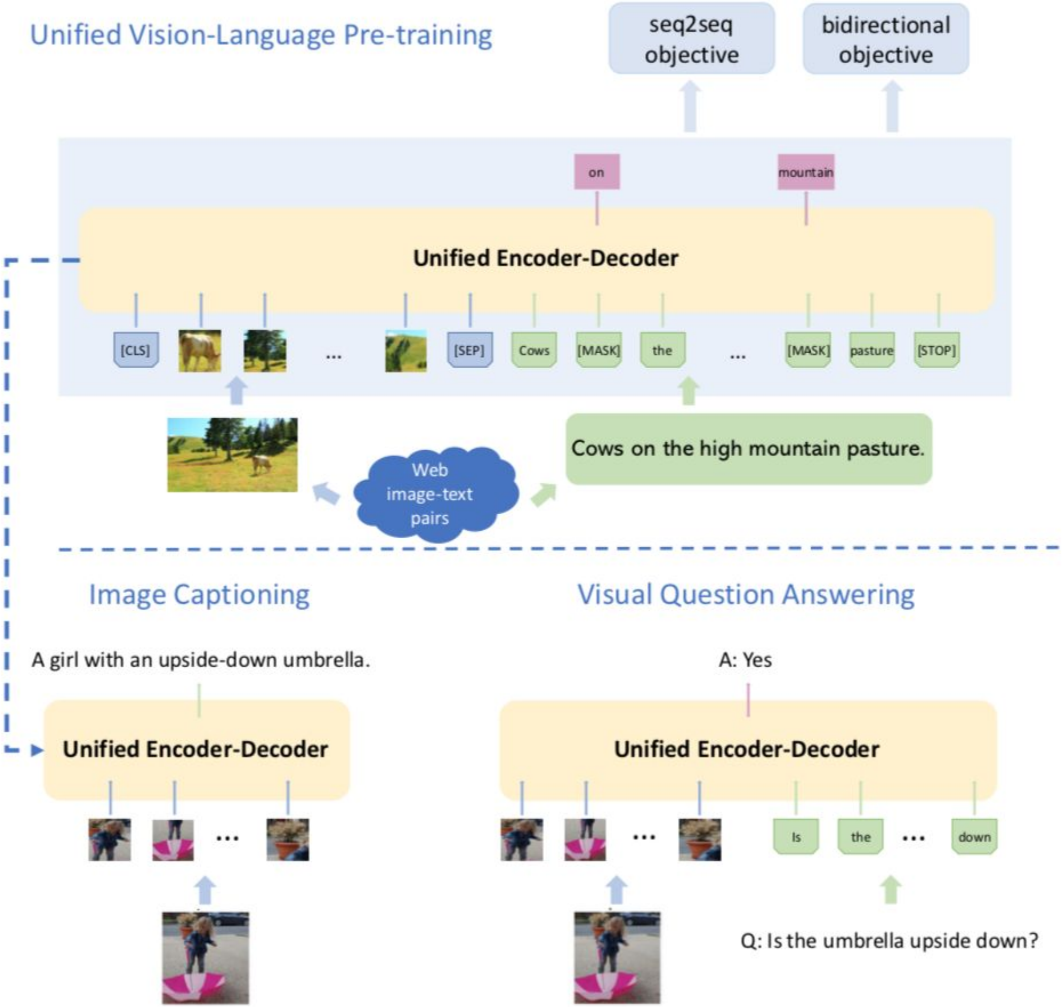
WS2. The Written World *(our present)*

- Two observations
    - Larger models see diminishing returns
        - For task LAMBADA, TuringNLG will achieve 67.8 in accuracy with 17B parameters
        - GPT-3 will improve 8% with 175B parameters
        - the slope of the increase is quickly decreasing!
    - The extent to which they capture deeper notions of contextual meaning remains an open question
        - pretrained word and sentence representations fail to capture many grounded features of words
- How can we improve?

WS3. The World of Sights and Sounds

- Language learning needs perception
  - Perception includes auditory, tactile, and visual input.
  - Cognitive science shows that children require grounded sensory perception, not just speech, to learn language
- Advances in CV
  - image captioning
  - visual question answering
  - visual reasoning/common sense
  - multilingual captioning/translation via video
- Combining text and vision
  - train large-scale, multimodal transformers even include audio

Unified Vision-Language Pre-training



seq2seq objective

bidirectional objective

on

mountain

Unified Encoder-Decoder

[CLS] ... [SEP] Cows [MASK] the ... [MASK] pasture [STOP]

Web image-text pairs

Cows on the high mountain pasture.

Image Captioning

A girl with an upside-down umbrella.

Unified Encoder-Decoder

Visual Question Answering

A: Yes

Unified Encoder-Decoder

Is the ... down

Q: Is the umbrella upside down?

WS3. The World of Sights and Sounds

- Benefits
  - An ideal WS3 agent will exhibit better long-tail generalization and understanding than any language-only system could.
  - most prominent in a test of zero-shot cir- cumstances
  - “Will *this* car fit through *that* tunnel?,”
  - Cognitive science shows that children require grounded sensory perception, not just speech, to learn language
- Limits
  - The agent has not tried to throw various objects and understand how their velocity and shape interact with the atmosphere to create lift.
  - Cannot test their actions in the physical world

WS4. Embodiment and Action

- *interactive* multimodal sensory experience forms the basis of action-oriented categories (robots: situated action taking)
- In a virtual world
    - 2D maze
    - a grid world
    - simulated house(ALFRED task)
        - *contain both high-level goals like “Rinse off a mug and place it in the coffee maker.” and low-level language instructions like “Walk to the coffee maker on the right.”*

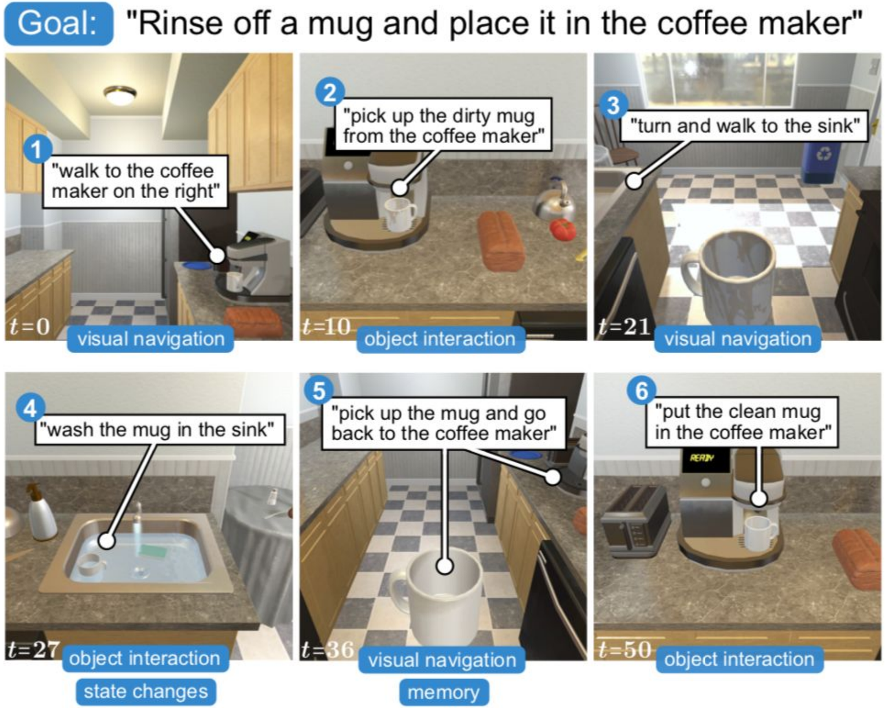

Goal: "Rinse off a mug and place it in the coffee maker"
1 "walk to the coffee maker on the right"
t=0
visual navigation
2 "pick up the dirty mug from the coffee maker"
t=10
object interaction
3 "turn and walk to the sink"
t=21
visual navigation
4 "wash the mug in the sink"
t=27
object interaction
state changes
5 "pick up the mug and go back to the coffee maker"
t=36
visual navigation
memory
6 "put the clean mug in the coffee maker"
t=50
object interaction

Figure 1: ALFRED consists of 25k language directives corresponding to expert demonstrations of household tasks. We highlight several frames corresponding to portions of the accompanying language instruction. ALFRED involves interactions with objects, keeping track of state changes, and references to previous instructions.

WS4. Embodiment and Action

- Benefits
  - The nuance of the physical wolrd
    - the orange and baseball afford similar manipulation *because* they have similar texture and weight, while the orange and banana both contain peels, deform, and are edible
  - Demonstration is unlimited
  - Allows the agent to construct rich pre-linguistic representations
    - much of the knowledge humans hold about the world is intuitive
      - possibly incommunicable by language, but still required to understand language
      - metaphur like “a distant concern”
- Limits
  - Robotics and embodiment are not available in the same off-the-shelf manner as computer vision models

WS5. The Social World

- Interpersonal communication
- History
    - Turing (1950)’s Imitation Game
    - a naïve tester could easily be tricked

Why do we need WS5

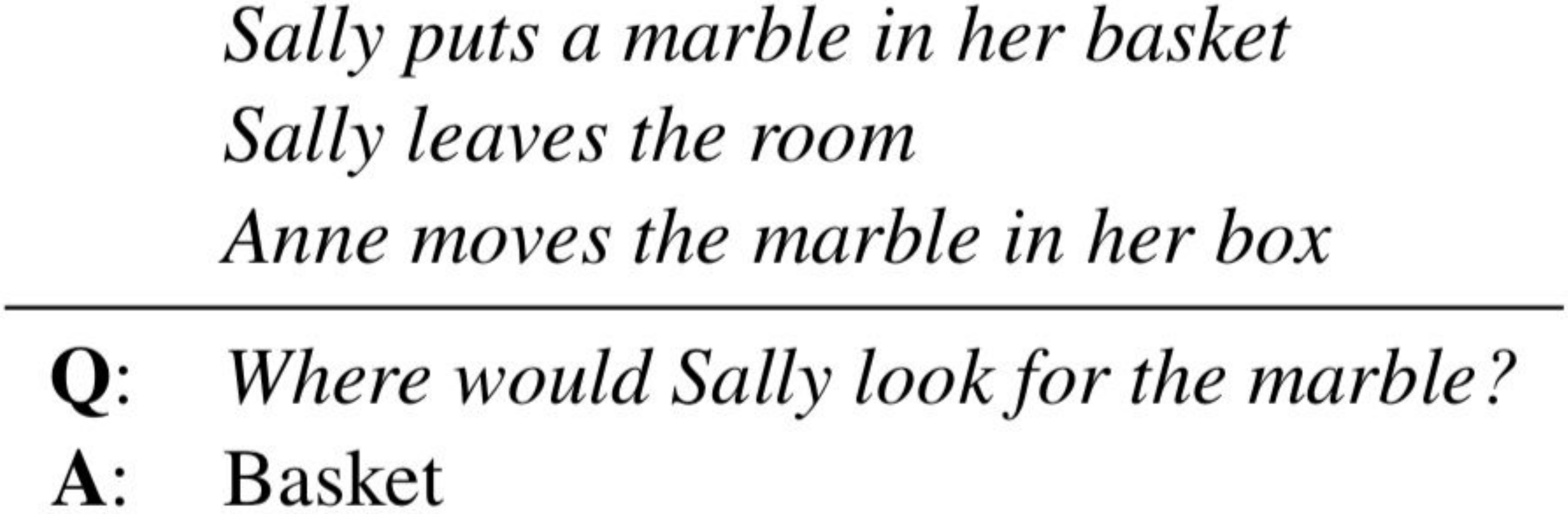
Function

- “Function is the source of meaning”
  - Evidence from cognitive science and philosophy
  - **Active** experimentation to learning that effect(feedback) from the world
    - Get extra demonstration of distribution of generated response(My understanding)

Theory of mind

- Reason about the mental state of the other agents
- Sally-Anne test(psychological version)
    - Ability to reason about other agents’ false beliefs, two agents(Sally, Anne)
    - puts an object into a container
    - Anne moves the object without Sally observing this action.
    - questions about reality and the agents’ beliefs
        - *First-Order Belief* : Where will Sally look for the marble?
        - *Reality*: Where is the marble really?
        - *Memory*: Where was the marble in the beginning?
        - *Second-Order Belief* :Where does Anne think Sally will look for the marble?

bAbi dataset



Sally puts a marble in her basket

Sally leaves the room

Anne moves the marble in her box

Q: *Where would Sally look for the marble?*

A: Basket

limitation for WS 1-4 to capture Theory of mind

- With only WS 2, the model cannot capture the Theory of mind
  - Is solvable without using theory of mind
  - After adding distractor phrases, locations, characters, the state-of-the-art failed
- Other limitation for WS 1-4
  - there is a lack of inductive bias
  - current cross entropy training losses actively discourage learning the tail of the distribution properly

Language in a social context

- Information like status, role, intention
- Collecting data about rich natural sit- uations is often impossible.
  - To address this gap, learning by participation, where users can freely interact with an agent, is a necessary step to the ultimately *social* venture of communication.

Summary

- You can’t learn language ...
  - ... from the radio (Internet). WS2 $\subset$ WS3
    - *A task learner cannot be said to be in* WS3 *if it can succeed without perception (e.g., visual, auditory).*
  - ... from a television. WS3 $\subset$ WS4
    - *A task learner cannot be said to be in* WS4 *if the space of its world actions and consequences can be enumerated.*
  - ... by yourself. WS4 $\subset$ WS5
    - *A task learner cannot be said to be in* WS5 *unless achieving its goals requires cooperating with a human in the loop.*