Integrating Cognitive Architecture and Neural Generative Language Models (LMs)

Bob Wray, John Laird

Jun 2021

CIC @ IQMRI

Motivation: Humans know how to find “missing” knowledge

Lexical:

- Dr. Stevens inserted the **broach** and successfully removed the pulp.
  Stevens is the only machinist who we trust with the 25-ton vertical pull-down **broach**.
  After foolishly **broaching** the boat in the rough seas, Stevens was soon capsized.

Semantic:

- *In 1987, the Prime Minister was briefed on the history of punk acts, including acts like the Clash and the Sex Pistols. “You may not like this,” the memo warned.*

Action-informing:

- How do I make soft-boiled eggs?

Note that we would have ideas for finding the missing knowledge before “Googling” was an option.

Various methods an agent might use to find missing knowledge?

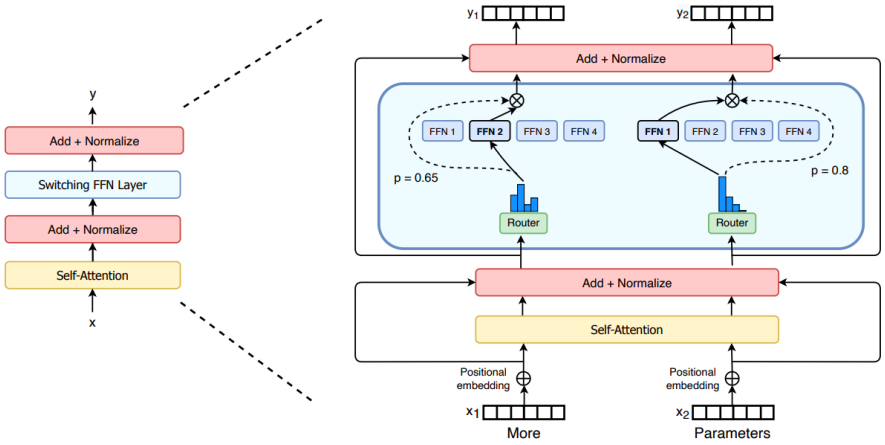
- Ask someone else
  - Examples from ITL
  - Responder may need to work around agent knowledge and language limitations
  - Limited by availability of someone who can provide an answer

- Mimic human capabilities / Investigate human-accessible resources (“Google it”)
  - Provides significant power and flexibility
  - (Probably) requires sophisticated language understanding

- Ask/access machine-interpretable, curated resources:
  - Example: Knowledge graphs, semantic web, etc.
  - Comparably easy to access, but generally a lot less knowledge to draw from

Seeking knowledge from neural generative language models

- Transformer-based neural generative language models (LMs)
  - ANN models trained to complete sentences
  - Trained on massive corpora
  - Set new benchmarks in sentence completion, language understanding, question answering, etc.
  - Huge investments in these models
    - Google (Switch-*/BERT), Facebook (RoBERTa), Microsoft (Turing-NLG), Open AI (GPT*), …



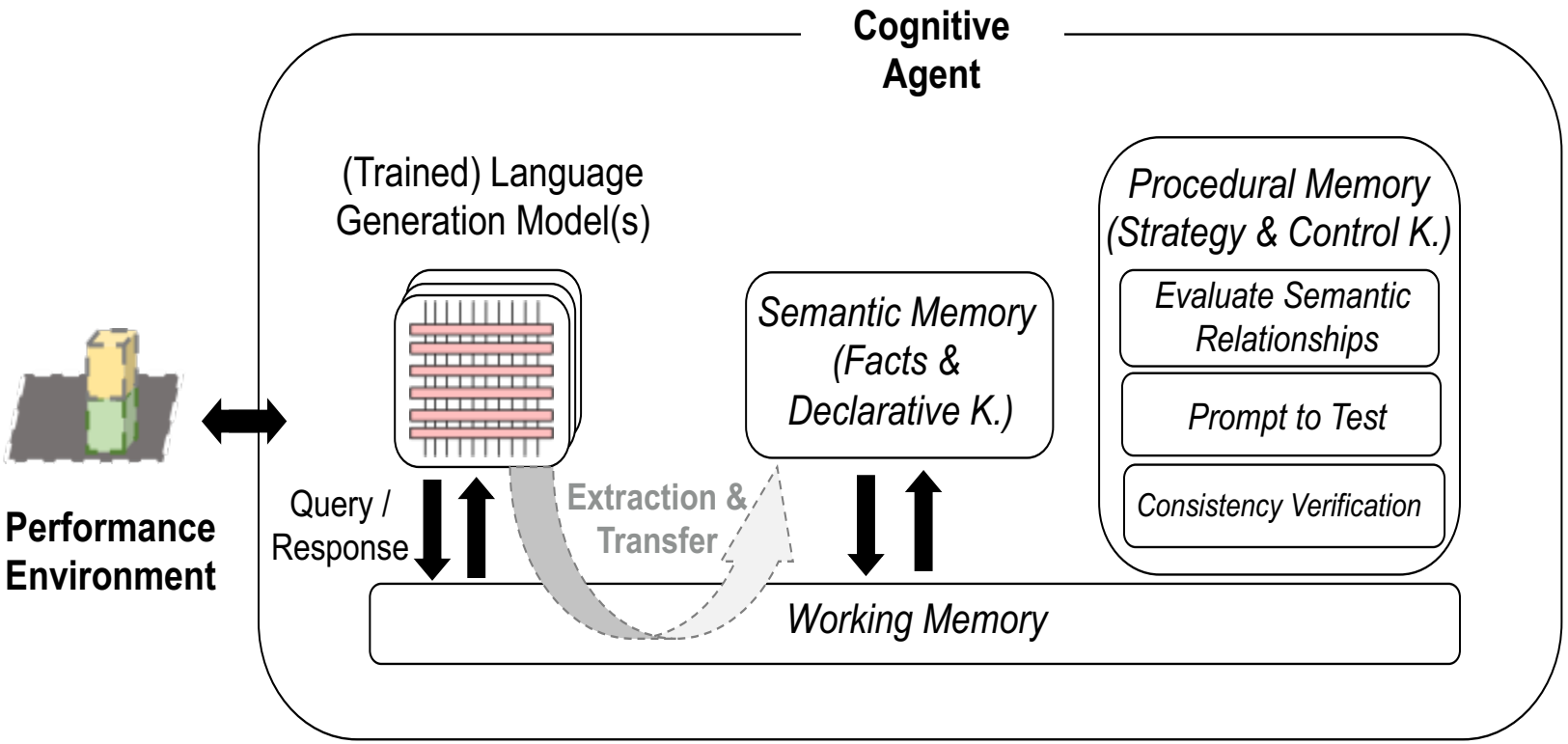
Switch Transformer (from Google)

- New class of information system?

- Research Questions:
  - How can we extract “missing” information from these models?
  - What kinds of knowledge can be extracted from these models?

CIC @ IQMRI

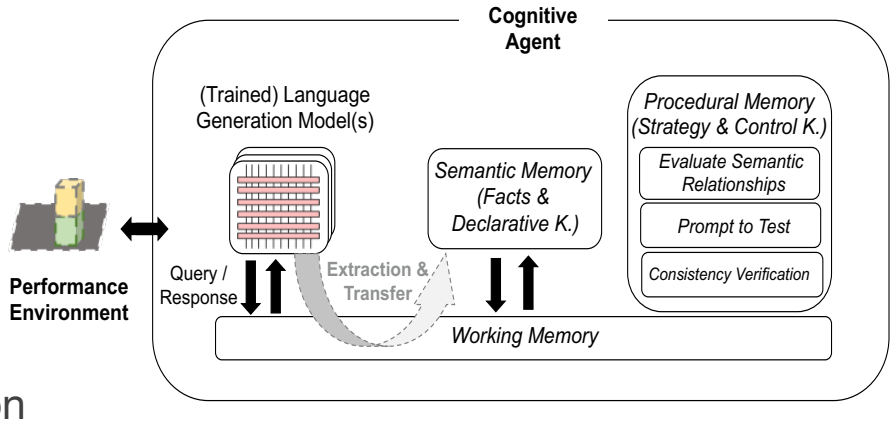
Initial Concept: Language Model as a Type of Memory



How can we extract knowledge for agent use?

Primary challenges:

- Functional integration
  - Explore patterns of integration (e.g., deliberate vs. spontaneous retrieval)
- Strategies for knowledge access
  - Deciding when/how to access
  - Creating cues for access
  - Interpreting and extracting relevant information
- Testing and verifying extracted information
  - *In 1987, the Prime Minister was briefed on the history of punk acts, including acts like the Clash and the Sex Pistols.*



Responses from GPT-2:

Talking to the writer Tim Butcher, Tony Blair said he kept the record (that day) in his top drawer. "I'm sure it's going to be played many times by the successors of this band," he said.

David Cameron is the latest of many current and former leaders to be asked about their thoughts on the Sex Pistols – with Cameron's response raising a few eyebrows.

What's misleading/wrong wrt these responses?

CIC @ IQMRI

What kinds of knowledge can be extracted?

- Numerous researchers now looking at this question
  - Two related, high-level conceptual questions:
    - What kinds of knowledge are present but latent (and what kinds are not present)?
    - How does the language-model encoding align (and not) with human language use?
      - Example: LM results largely derive from lexical encodings, not compositionality (Ettinger)

- Plan: Use the language model to development increasingly complex capabilities
  - Disambiguation: Resolve an intended meaning when the task context is insufficient (“bank”)
  - Relational discovery: Discover additional details about an imprecisely-defined concept (“tellers” are employed-by “banks”)
  - Word learning: Understand and use* a new word/concept
    - * Performative use as well as communicative use
  - …

Nuggets

- Language models (LMs) offer potential as a source of significant knowledge for agents.
- Lots of investment and interest in LMs (work we can draw from and build on)
- Cognitive architectures have potential to provide larger framework/context for improved utility of LMs
- Patterns of query, interpretation, test may be applicable to other knowledge sources (e.g., knowledge graph)

Coal

- Non-trivial time/computation cost to explore alternatives
- More direct/explicit language understanding models likely preferable in the long term

Nuggets & Coal