Implicit Representations of Meaning in Neural Language Models

Belinda Z. Li Maxwell Nye Jacob Andreas

Massachusetts Institute of Technology

Presented By: Arya Honraopatil

Agenda

1. Summary
2. Methodology
3. Results
4. Strengths
5. Weaknesses
6. How it relates to Interactive Fiction/Story Generation?
7. Questions

Summary

- explores how neural language models (NLMs), such as BART and T5 learn to represent meaning
- investigates if NLMs implicitly encode representations of entities and their dynamic states
- uses two datasets, Alchemy and TextWorld

Key Findings:

a. NLMs develop structured, queryable, and manipulable semantic models
b. These models update entity properties and relations as a discourse evolves
c. NLMs trained only on text can capture meaningful world states

Methodology

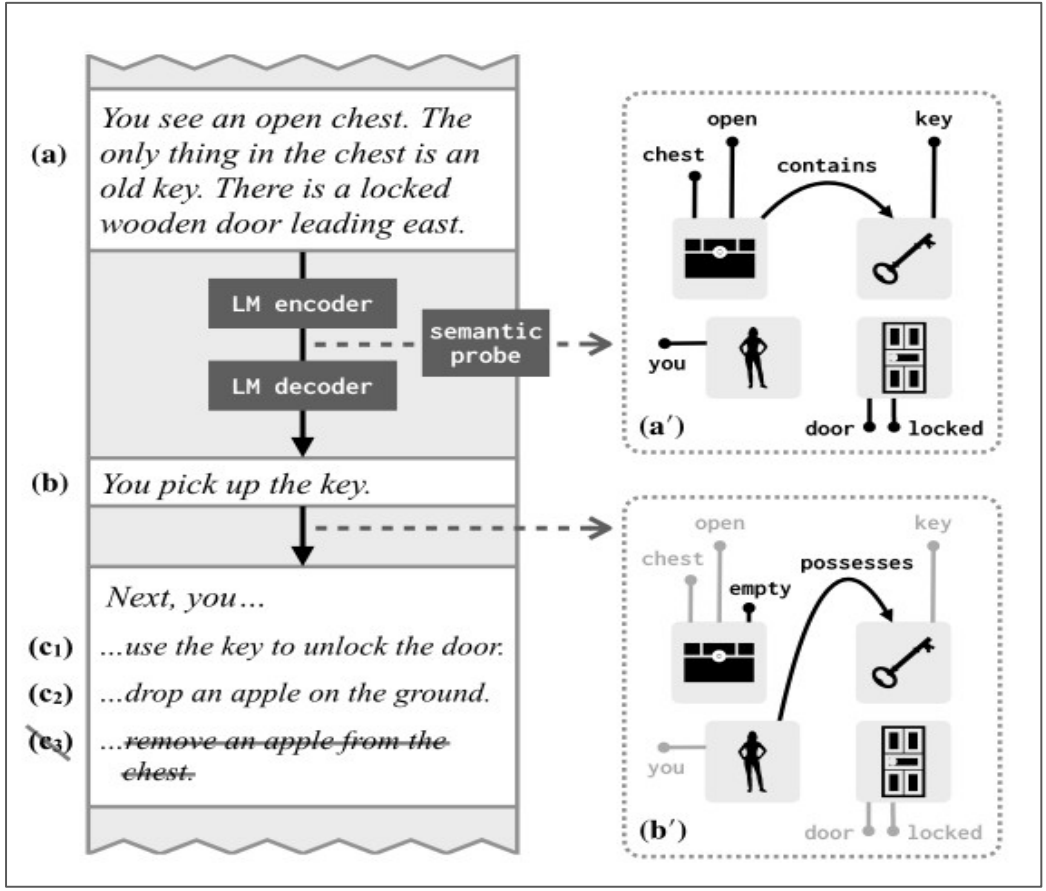


Figure 1. Neural language models trained on text alone (a–c) produce semantic representations that encode properties and relations of entities mentioned in a discourse (a′). Representations are updated when the discourse describes changes to entities' state (b′).

Observations:

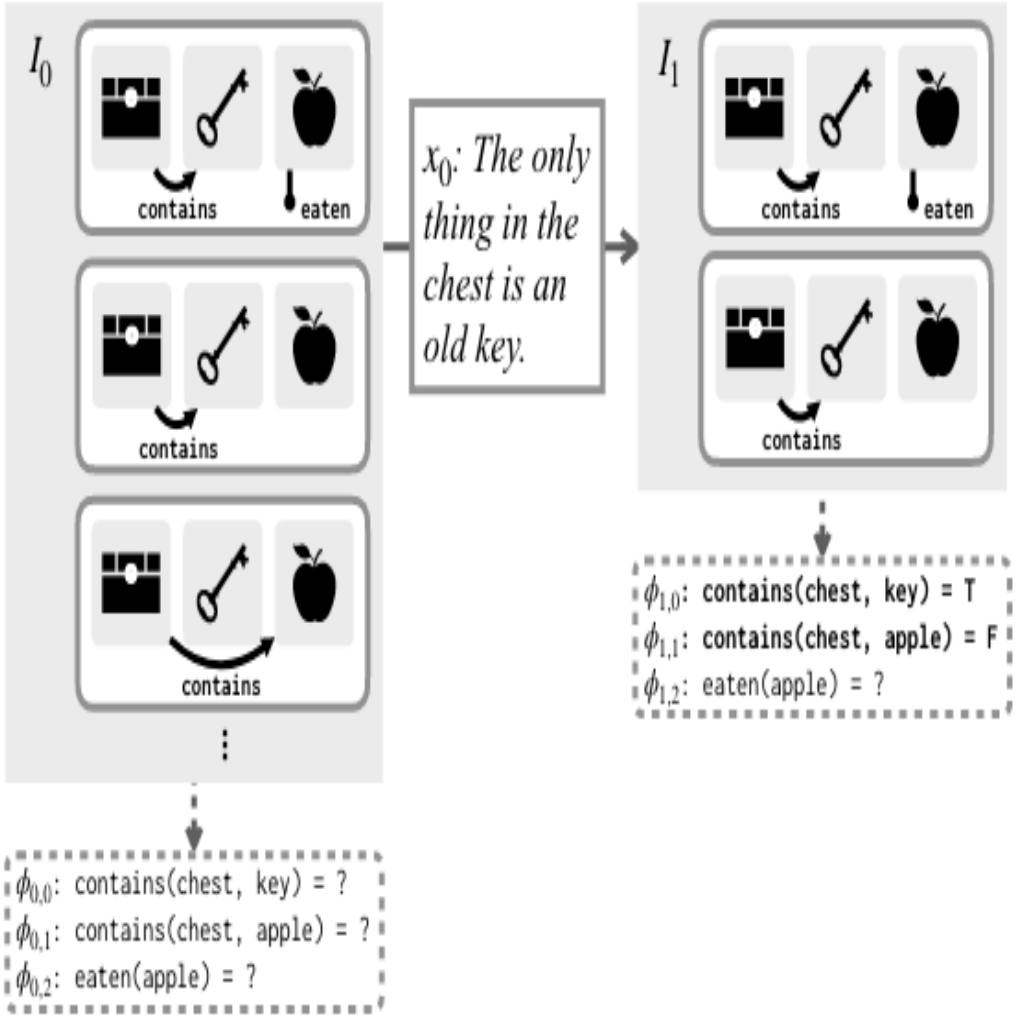
- NLMs face the problem of learning to generate coherent text
- LMs cannot represent meaning at all

Arguments from the paper:

- NLMs learn implicit models of meaning that are translatable into formal state representations like (a′)–(b′) (from fig. 1)
- they follow a semantically necessary consequence (represent set of entities & update facts)
- NLM training can produce models of meaning

Figure 2:

- collection of possible situations is an information state (I0)
- information states assign values to propositions φi,j according to whether they are true, false, or undetermined
- appending a new sentence discourse causes the information state to be updated (I1).
- In this case, the sentence, “The only thing in the chest is an old key” causes contains(chest, key) to become true, contains(chest, apple) to become false, and leaves eaten(apple) undetermined.



Information State:

- the set of possible states of the world consistent with a discourse (I0 and I1 in Fig. 2)
- each new sentence in a discourse provides an update
- are represented logically
- are decoded via the truth values that they assign to logical propositions

Approach:

- probe for the truth values of logical propositions about entities mentioned in the text For example, in Fig. 1, we test whether a representation of sentences (a)–(b) encodes the fact that empty(chest) is true and contains(chest, key) is false
- and then train probing models to test whether NLMs represent the information states specified by the input text

Probing:

- is a process of analyzing internal representations learned by a model to understand what linguistic or semantic information is encoded at different layers or positions within the model
- uses a technique called *probe tasks*, where a simple classifier (probe) is trained on top of the representations (embeddings or hidden states) from a specific layer of the language model

Types of Probing:

Syntactic Probes: checks if the model encodes syntactic structures like parse trees or part-of-speech tags

Semantic Probes: checks if the model encodes word meanings, relationships between words, or sentence entailment

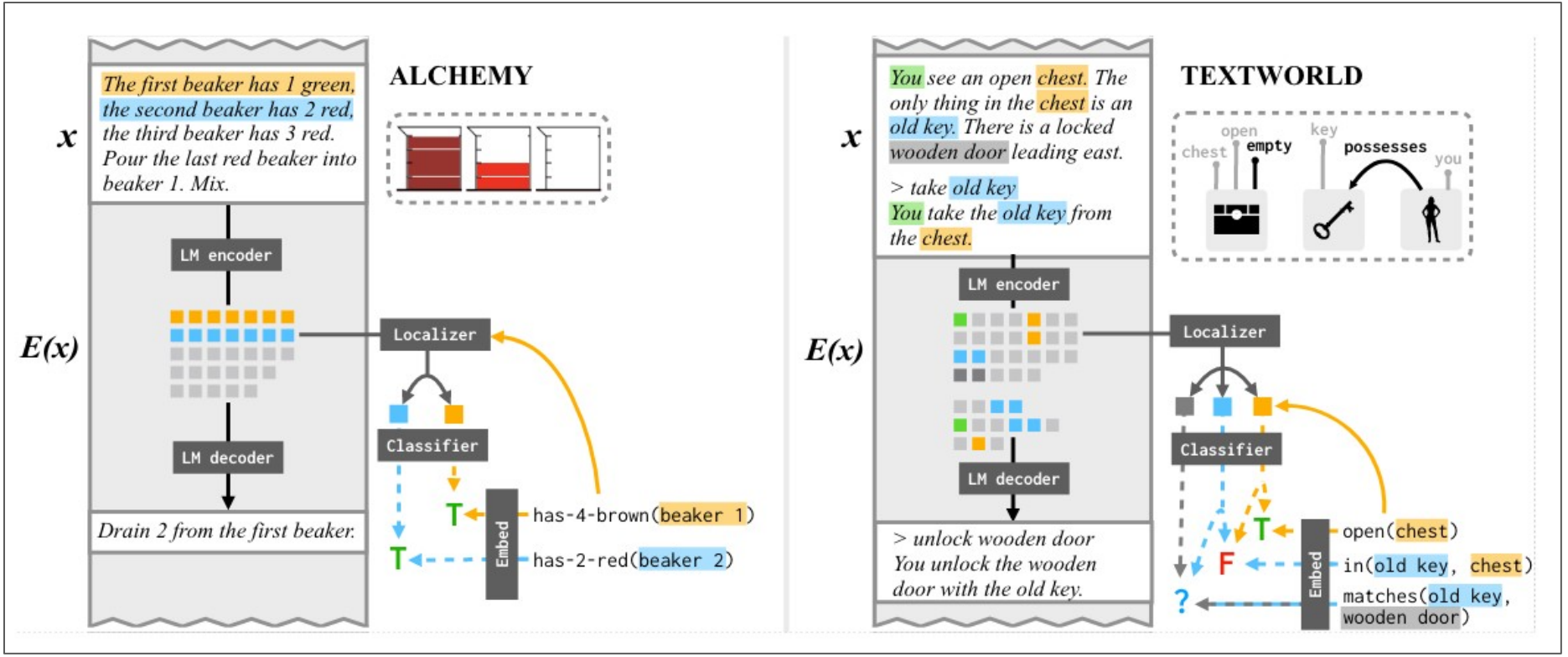


Figure 3. Overview of the probe model. Left: Alchemy. Right: Textworld. The LM is first trained to generate the next instruction from prior context (left side, both figures). Next, the LM encoder is frozen and a probe is trained to recover (the truthfulness of) propositions about the current state from specific tokens of encoder outputs.

The probe has 3 components:

Embed - converts propositions about entities and their states into dense vector representations

Localizer - extracts the relevant token embeddings from the language model, corresponding to the entities mentioned in the proposition

Classifier - compares the proposition's vector with the context embeddings to determine the truth of the proposition

Results

		Alchemy				TextWorld			
		State EM		Entity EM		State EM		Entity EM	
		BART	T5	BART	T5	BART	T5	BART	T5
main probe (§4.2)		7.6	14.3	75.0	75.5	48.7	53.8	95.2	96.9
baselines & model ablations (§4.2)	+pretrain, -fine-tune	1.1	4.3	69.3	74.1	23.2	38.9	91.1	94.3
	-pretrain, +fine-tune	1.5		62.8		14.4		81.2	
	random init.	0.4		64.9		11.3		74.5	
	no change	0.0		62.7		9.73		74.1	
	no LM	0.0		32.4		1.77		81.8	

Strengths

- introduces a new method to probe NLMs for representations of meaning
- provide concrete evidence of how NLMs encode dynamic situations and entities
- the work has broad implications for improving coherence and factuality in NLMs
- shows implicit models of meaning could enhance performance in various NLP tasks

Weaknesses

- the datasets used feature relatively simple situations with few objects and relations
- LM output and implicit state representations are not perfect (even in the best case, complete information states can only be recovered 53.8% of the time)
- the semantic representations do not have the expressiveness needed to support human-like generation
- whether the errors in language model prediction are attributable to errors in the underlying state representation is not known

How it relates to Interactive Fiction/Story Generation

In story generation:

- to manage dynamic plot developments (for e.g. as characters move through a narrative, the model can track their relationships, items they possess, or changes in their environment)
- generated stories can maintain coherence across multiple scenes or interactions

How it relates to Interactive Fiction/Story Generation

In interactive fiction:

- to improve the responsiveness and realism of interactive fiction (for e.g. when a player's action changes the state of the world, the model also updates it's entity states)
- can be used for personalized storytelling (for e.g. a player's decisions can lead to unique story paths or endings)

QUESTIONS?

THANK YOU!