Towards Explainable and Ontologically Grounded Language Models*

Walid S. Saba[1]

[1] Institute for Experiential AI, Northeastern University, 100 Fore St, Portland, ME 04101 USA

Abstract
We argue that the relative success of large language models (LLMs) is not a reflection on the symbolic vs. subsymbolic debate but a reflection on employing an appropriate bottom-up strategy of a reverse engineering of language at scale. However, and due to their subsymbolic nature whatever knowledge these systems acquire about language will always be buried in millions of weights none of which is meaningful on its own, rendering such systems utterly unexplainable. Furthermore, and due to their stochastic nature, LLMs will often fail in making the correct inferences in various linguistic contexts that require reasoning in intensional, temporal, or modal contexts. To remedy these shortcomings we suggest employing the successful bottom-up strategy employed in LLMs but in a symbolic setting, resulting in explainable, language-agnostic, and ontologically grounded language models.

Keywords
Large language models, ontology, bottom-up reverse engineering

1. Introduction

To arrive at a scientific explanation there are generally two approaches we can adopt, a top-down approach or a bottom-up approach (Salmon, 1989). However, for a top-down approach to work, there must be a set of established general principles that one can start with, which is clearly not the case when it comes to language and how our minds externalize our thoughts in language. In retrospect, therefore, it is not surprising that decades of top-down work in natural language processing (NLP) failed to produce satisfactory results since most of this work was inspired by theories that made questionable assumptions where, for example, an innate universal grammar was assumed (Chomsky, 1957), or that we metaphorically build our linguistic competence based on a set of idealized cognitive models (Lakoff, 1987), or that natural language could be formally described using the tools of formal logic (Montague, 1973). In a similar vein, it is perhaps for the same reason that decades of top-down work in ontology and knowledge representation (Lenat and Guha, 1990 and Sowa, 1995) also faltered since most of this work amounted to pushing, in a top-down manner, metaphysical theories of how the world is supposedly structured and represented in our minds, and again without any agreed upon general principles to start with. On the other hand, unprecedented progress has been made in only a few years of NLP work that employed a data-driven bottom-up strategy, as exemplified by recent advances in large language models (LLMs) that are essentially a massive experiment of a bottom-up reverse engineering of language at scale (e.g., ChatGPT and GPT-4)[2].

1.1. Issues with LLMs

Despite their relative success, LLMs do not tell us anything about how language works since these models are not really models of language but are statistical models of regularities found in language[3]. In

KiL'24: Workshop on Knowledge-infused Learning co-located with 30th ACM KDD Conference, August 26, 2024, Barcelona, Spain.
* Corresponding author.
† These authors contributed equally.
w.saba@northeastern.edu (W. Saba)

[2] GPT stands for 'Generative Pre-trained Transformer', an architecture that OpenAI built on top of the transformer architecture (Vaswani, A. et. al., 2017).
[3] In looking inside the neural network (NN) of an LLM one does not find concepts, meanings, linguistic structures, etc. but weights associated with neural connections, which is exactly what one will find in an object recognition or any other NN.

fact, and due to their subsymbolic nature, whatever 'knowledge' these models acquire about language will always be buried in millions of weights (microfeatures) none of which is meaningful on its own, rendering these models utterly unexplainable (Guizzardia and Guarino, 2024). Besides unexplainability, LLMs are also oblivious to truth (Borji, 2023), since for LLMs all text (factual or non-factual), is treated equally. Finally, and while LLMs have been shown to do poorly in a number of tasks that require high-level reasoning such as planning (Valmeekam et. al., 2023), analogies (Lewis and Mitchell, 2024) and formal reasoning (Arkoudas, 2023) what concerns here is the failure of LLMs in making the right inferences in various linguistic contexts. As an illustration of the kinds of failures in deep language understanding we consider here three linguistic contexts involving copredication, intension and prepositional attitudes.

Example 1. *Show the entities and the relations that are implicit in the following text*: "I threw away the newspaper I was reading because they fired my favorite columnist".

Example 2. *Since Madrid is the capital of Spain, can I replace one for the other in the following:* "Maria thinks Madrid was not always the capital of Spain"?

Example 3. *Suppose Devon knows that if someone is a client, then s/he is a student, and suppose that Olga is a client.* Then what does Devon know?

The first example involves a phenomenon called *copredication* (see Asher and Pustejovsky, 2005) which occurs when the same entity is used in the same context to refer to more than one semantic (ontological) type. All LLMs tested[4] failed in recognizing that 'newspaper' in the text is used to simultaneously refer to three entities: (i) the physical object I *threw away*; (ii) the content of the newspaper *I was reading*; and (iii) the 'editorial board' of the newspaper that *did the firing of the columnist*. Note that the failure of the LLMs was more acute when the LLMs were asked to draw a graph showing all entities and relations implied by the text since to show all the relations in the text all the different types of entities must be extracted. Here all LLMs tested showed the same newspaper (*physical*) object doing the firing of the columnist.

In example 2 all LLMs we tested approved replacing 'the capital of Spain' by 'Madrid' resulting in 'Maria thinks that Madrid was not always Madrid'. It is worth noting that the LLMs tested were consistently oblivion to intension. For example, in 'Perhaps Socrates was not the tutor of Alexander the Great', 'Socrates' and 'the tutor of Alexander the Great' were also deemed replaceable (since they are extensionally equal) resulting in 'Perhaps Socrates was not Socrates'. These results were expected since neural networks (deep or otherwise), that are the computing architecture behind all LLMs, are purely extensional models and are based on the 'empiricist theory of abstraction' where their similarity semantics has no notion of 'object identity' (Lopes, 2023).

Finally, example 3 illustrates failures of LLMs in making the correct inferences in modal (belief) contexts: the response of the LLMs tested was that 'Devon knows that *Olga* is a student' which is clearly the wrong inference since inferring $\mathbf{K}(Devon, \text{student}(Olga))$ from $\mathbf{K}(Devon, \text{client}(Olga) \supset \text{student}(Olga))$ requires $\mathbf{K}(Devon, \text{client}(Olga))$, i.e., it requires Devon knowing that *Olga* is a client. We have collected many other tests that we make available elsewhere for the sake of saving space.[5]

1.2. LLMs: A Glass Half Empty, Half Full

So where do we stand now? On one hand, LLMs have clearly proven that one can get a handle on syntax and quite a bit of semantics in a bottom-up reverse engineering of language at scale; yet on the other hand what we have are unexplainable models that do not shed any light on how language actually works. Moreover, it would seem that due to their purely extensional and statistical nature, LLMs will always fail in making the correct inferences in many linguistic contexts. Since we believe the relative success of LLMs is not a reflection on the symbolic vs. subsymbolic debate but is a reflection on a successful bottom-up reverse engineering strategy, we think that combining the advantages of symbolic and ontologically grounded representations with a bottom-up reverse engineering strategy is a worthwhile effort. In fact, the idea that word meaning can be extracted from how words are actually used in language is not exclusive to linguistic work in the empirical tradition, but in fact it can be traced back to Frege.

In the rest of the paper we will (i) first argue that current word embeddings that are the genesis of modern-day large language models can be constructed in a symbolic setting instead of being the result of statistical cooccurrences; (ii) we will show that symbolic vectors perform better than current embeddings on a well-known word similarity benchmark; (iii) we will discuss how our symbolic

[4] Our experiments were conducted on GPT-4o (chat.openai.com).

[5] https://shorturl.at/ejmH8

vectors can be used to *discover* the ontological structure that is implicit in our ordinary language.

2. Concerning 'the Company a Word Keeps'

The genesis of modern LLMs is the *distributional semantics hypothesis* which states that the more semantically similar words are, the more they tend to occur in similar contexts – or, similarity in meaning is similarity in linguistic distribution (Harris, 1954). This is usually summarized by a saying that is attributed to the British linguist John R. Firth that "you shall know a word by the company it keeps". When processing a large corpus, this idea can be used by analyzing co-occurrences and contexts of use to approximate word meanings by word embeddings (vectors or tensors), that are essentially points in multidimensional space. Thus, at the root of LLMs is a bottom-up reverse engineering of language strategy where, unlike top-down approaches, "reverse engineers the process and induces semantic representations from contexts of use" (Boleda, 2020). But nothing precludes this idea from being carried out in a symbolic setting. In other words, the 'company a word keeps' can be measured in several ways, other than the correlational and statistical measures that underlie modern word embeddings.

2.1. Symbolic Dimensions of Meaning

In discussing possible models of the world that can be employed in computational linguistics Hobbs (1985) once suggested that there are two alternatives: (i) on one extreme we could attempt building a "correct" theory that would entail a full description of the world, something that would involve quantum physics and all the sciences; (ii) on the other hand, we could have a promiscuous *model of the world that is isomorphic to the way we talk it about in natural language* (emphasis is ours). Since the first option is a project that is most likely impossible to complete, what Hobbs is clearly suggesting here is a reverse engineering of language to discover how we actually use language to talk about the world we live in. This is also not much different from Frege's Context Principal that suggests "never ask for the meaning of words in isolation" (Dummett, 1981) but that a word gets its meanings from analyzing all the contexts in which the word can appear (Milne, 1986). Again, what this suggests is that the meaning of words is embedded (to use a modern terminology) in all the ways we use these words in how we talk about the world. While Hobbs' and Frege's observations might be a bit vague, the proposal put forth by Fred Sommers (1963) was very specific. Again, Sommers suggests that "to know the meaning of a word is to know how to formulate some sentences containing the word" and this would lead, like in Frege's case, to the conclusion that a complete knowledge of some word *w* would be all the ways *w* can be used. For Sommers, the process of understanding the meaning of some word *w* starts by analyzing all the properties *P* that can **sensibly** be said of *w*. Thus, for example, [*delicious Thursday*] is not sensible while [*delicious apple*] is, regardless of the truth or falsity of the predication. Moreover, and since [*delicious cake*] is also sensible, then there must be a common type (perhaps food?) that subsumes both *apple* and *cake*. This idea is similar to the idea of type checking in strongly typed polymorphic programming languages. For example, the types in an expression such as '$x + 3$' will only unify (or the expression will only 'make sense') if/when x is an object of type number (as opposed to a tuple, for example). As it was suggested in (Saba, 2007), this type of analysis can thus be used to 'discover' the ontology that seems to be implicit in the language, as will be discussed below. First, however, we describe how a bottom-up reverse engineering of language can be done in a symbolic setting.

2.2. Symbolic Reverse Engineering of Language

The procedure we have in mind assumes a Platonic universe where all concepts, physical or abstract, including states, activities, properties (tropes) (Moltmann, 2013), processes, events, etc. are considered entities that can be defined by a number of language-agnostic primitives (Smith, 2005) that we call the 'dimensions of meaning'. We consider here the following dimensions: AGENTOF, OBJECTOF, HASPROP, INSTATE, PARTOF, INSTATE, INPROCESS, and OFTYPE. For every word *w* in the language, and for every dimension **D**, a reverse-engineering process is conducted to compute a set $w^{\mathbf{D}} = \{(x, t) \mid \mathbf{D}(w, x)\}$ where t is a weight in [0,1]. Here are example sets computed for 'book' along four dimensions of meaning along with the masking prompt that queries what an LLM has 'learned' about how we talk about books:

book . HASPROP
Everyone likes to read a [**MASK**] *book.*
=> {(popular, 0.9), (educational, 0.8), (famous, 0.8), ... }

book . OBJECTOF
Everyone I know enjoyed [**MASK**] *'The Prince'.*
=> {(reading, 0.9), (writing, 0.8), (editing, 0.8), ... }

book . AGENTOF
Das Kapital has [MASK] *many people over the years.*
=> {(influenced, 0.9), (inspired, 0.8), (changed, 0.8), ... }

book . PARTOF
Hamlet should be part of every [**MASK**].
=> {(collection, 0.9), (archive, 0.8), (library, 0.8), ... }

book . INSTATE
I was told that my book is now in [**MASK**].
=> {(print, 0.9), (circulation, 0.8), (review, 0.8), ... }

What the above says is the following (i) in ordinary spoken language we speak of a 'book' that is *popular*, *educational*, *famous*, etc.; (ii) we speak of *reading*, *writing*, *editing*, etc. a 'book'; (iii) we speak of 'book' that may *change*, *influence*, *inspire*, etc.; and (iv) we speak of a b 'book' that is part of a *collection*, an *archive*, or a *library*; and (v) a book can be in *review*, in *print*, in *circulation*, etc. The nominalization process can be conducted using the copular 'is' as shown in table 1. For example, 'John is famous' can be restated as 'John *has the property* of fame'; 'Jim is sad' as 'Jim is *in a state* of sadness'; etc. (see [Smith, 2005] for more on the relationship between the copular and abstract entities and [Moltmann, 2013] for more on abstract objects.) What should be noted here is that even with the simple conceptual structure discovered thus far one can *generate* plausible text, such as the following:

(1) *enjoyed the interesting reading of the new book*
(2) *completed a boring reading of a controversial book*

Table 1: From propositions to relations and entities

Copular form	Nominalized form
John is famous	John has the PROPERTY fame
Jim is sad	Jim in a STATE of sadness
Maria is recognized	Maria as OBJECT of recognition
Olga is dancing	Olga as AGENT of a dancing
Sara is maturing	Sara in PROCESS of maturation

The sensible (and meaningful) fragment in (1) can be generated because a *book* can be 'read' and described by 'new', and *readings* can be 'interesting' and the object of *enjoyment*; and similarly for (2) where a *reading* of a *controversial* book can be *boring* and the object of a *completion*, etc. Note, however, that text generation in this case is not a function of 'predicting' the most likely continuation, but a function of plausible filling in of subjects, objects, agents, descriptions, etc. to any propositional structure.

2.3. Symbolic Embeddings

The process we described thus far results in symbolic word embeddings as the one shown in figure 2 below. In figure 2(a) we show the symbolic embedding for 'boy' and 'lad' along the HASPROP dimension. Thus, in ordinary spoken language it is sensible to speak of a 'handsome boy' and a 'funny boy' as well as a 'clever lad' and a 'talented lad'. We note here that in this process generic descriptions are removed using a function that computes the information content of some adjectives, where the information content of an adjective *adj* is inversely proportional to the set of types of *adj* can sensibly be applied to. For example, 'beautiful' will have a low information content score since 'beautiful' can sensibly be said of many concepts, both physical and abstract (e.g., *car*, *movie*, *poem*, *night*, *girl*, ...) while 'tasty' can sensibly be said of 'food' and just a few others. The symbolic embeddings in figure 2(b) are those of 'automobile' and 'car' along the OBJECTOF dimension. Note now that word similarity along these symbolic dimensions can be computed using cosine similarity as well as weighted Jaccard similarity where max and min can be used in fuzzy union and fuzzy intersection. We are currently experimenting with the optimal number of dimensions using a number of word similarity benchmarks, including the WordSim353 dataset (Finkelstein, Lev. et al., 2001)[6].

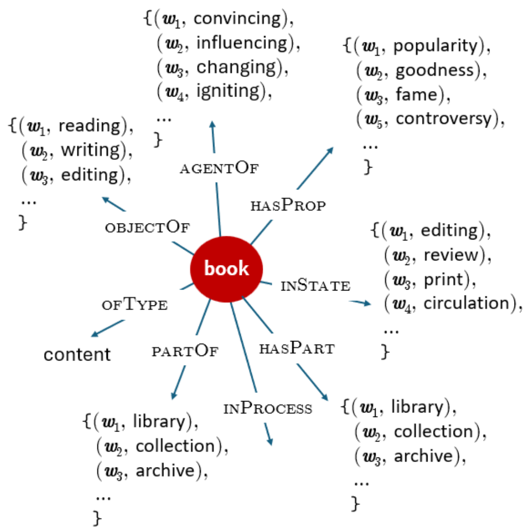

Figure 1: We speak of a 'book' (i) that influence, change, convince; (ii) that is edited, read, written; (iii) that can be popular, controversial, famous; (iv) that is part of a library, an archive, etc.

3. The Ontology of the Language of Thought?

The reverse engineering process we have described above would result in symbolic embeddings along

[6] https://kaggle.com/datasets/julianschelb/wordsim353-crowd

various dimensions, as the ones shown in figure 2. As a result of this, however, we could then analyze the subset relations between these embeddings to discover the ontological structure that seems to be implicit in our ordinary language. To illustrate, consider the following:

(3) car . objectOf
= {(driving, 0.9), (repairing, 0.8), (buying, 0.8), ... }
(4) book . objectOf
= {(reading, 0.9), (writing, 0.8), (buying, 0.8), ... }
(5) person . AGENTOF
= {(reading, 0.9),(writing, 0.8), (driving, 0.8), ... }
(6) person. HASPROP
= {(popularity, 0.9), (fame, 0.8), (beautiful, 0.8), .. }
(7) car. hasProp
= {(popularity, 0.9), (fame, 0.8), (beautiful, 0.8), .. }
(8) book . HASPROP
= {(popularity, 0.9), (fame, 0.8), (beautiful, 0.8), .. }

Note that car can be the object of 'buying' and so can be a book and this means that car and book must, at some level of abstraction, share the same parent (perhaps 'artifact'?) Note also that a car as well as a book and a person can be popular. An analysis along these lines would result in the following:

(9) read(person, book)
(10) write(person, book)
(11) buy(person, T_1 = car ∪ book ...)
(12) drive(person, car)
(13) beautiful(T_2 = person ∪ car ∪ book ...)

What the above says is the following: in ordinary spoken language we speak of people reading and writing books (9 and 10); we speak of people buying cars and books, and thus of buying objects that are of some type that subsumes both cars and books (11); we speak of people driving cars (12); and we speak of beautiful people, cars, and books (and thus beautiful seems to be a property that can sensibly be said of concepts that are at very high level of generality). As suggested by Sommers (1963) this type of analysis that can be fully automated with the help of LLMs can help us discover what he called 'the Tree of Language' – which is essentially the ontology that seems to be underneath our ordinary language. This might also be what Hobbs (1985) was seeking when he suggested building a model of the world that isomorphic to the we talk about it in natural language.

4. Concluding Remarks

Large language models (LLMs) have shown impressive capabilities that pioneers in artificial intelligence and natural language processing would marvel at.

a very [MASK] **boy**	a very [MASK] **lad**
('handsome', 0.93)	('handsome', 0.91)
('cute', 0.90)	('clever', 0.84)
('naughty', 0.83)	('nice', 0.82)
('nice', 0.80)	('talented', 0.81)
('clever', 0.78)	('young', 0.77)
('lucky', 0.77)	('cute', 0.76)
('pretty', 0.74)	('lucky', 0.75)
('funny', 0.73)	('brave', 0.75)
('brave', 0.73)	('funny', 0.74)
('obedient', 0.72)	('naughty', 0.72)
('talented', 0.72)	('pretty', 0.71)
('young', 0.71)	('adorable', 0.70)
('polite', 0.68)	('polite', 0.69)
('pleasant', 0.67)	('annoying', 0.69)

(a)

they were [MASK] an **automobile**	they were [MASK] a **car**
('driving', 0.9253)	('driving', 0.9253)
('buying', 0.8541)	('buying', 0.8233)
('repairing', 0.7624)	('chasing', 0.7492)
('chasing', 0.7603)	('repairing', 0.7193)
('designing', 0.7596)	('expecting', 0.7171)
('creating', 0.7446)	('owning', 0.7147)
('constructing', 0.7272)	('needing', 0.7143)
('owning', 0.7263)	('looking', 0.7133)
('purchasing', 0.7224)	('rocking', 0.712)
('wearing', 0.7205)	('destroying', 0.7084)
('looking', 0.7185)	('restoring', 0.7053)
('racing', 0.716)	('checking', 0.703)
('leasing', 0.7146)	('leasing', 0.703)
('destroying', 0.7136)	('inspecting', 0.7027)
('possessing', 0.7119)	('charging', 0.7026)

(b)

Figure 2: (a) the symbolic embeddings of 'boy' and 'lad' along the HASPROP dimension (with a weighted Jaccard similarity of 0.876) and (b) those of 'automobile' and 'car' along the OBJECTOF dimension (the weighted Jaccard similarity is 0.91)

However, we believe that LLMs are not the answer to the language understanding problem nor to reasoning in general and in particular commonsense reasoning. Due to their paradigmatic unexplainability LLMs will also not shed any light on how language works and how we externalize our thoughts in language. Since, in our opinion, the relative success of LLMs is not due to their subsymbolic nature but due to applying a successful bottom-up reverse engineering strategy, we suggested here applying the same strategy but in a symbolic setting, something that has been argued for by logicians dating back to Frege. By combining the successful bottom-up strategy and symbolic and ontological methods we arrive at explainable and ontologically grounded language models that can be used in problems requiring commonsense reasoning.

We are still in the early stage of this work, but we currently have the tools to realize the dream of Frege and Sommers and perhaps shed some light on the 'language of thought' Fodor (1998) – the internal language that we use to construct and process our thoughts.

References

[1] Nicholas Asher and James Pustejovsky. 2005. Word Meaning and Commonsense Metaphysics. In: Course Materials for Type Selection and the Semantics of Local Context, ESSLLI 2005.

[2] Gemma Boleda. 2020. Distributional Semantics and Linguistic Theory, Annual Review of Linguistics, 6, pp. 213–34.

[3] Borji, A. 2023. A Categorical Archive of ChatGPT Failures, Available online at https://arxiv.org/abs/2302.03494

[4] Noam Chomsky. 1957. Syntactic Structures, Mouton de Gruyter, NY.

[5] Michael Dummett. 1981. Frege: Philosophy of Language. Harvard University Press.

[6] Finkelstein, Lev, et al. "Placing search in context: The concept revisited." Proceedings of the 10th international conference on the World Wide Web. ACM, 2001.

[7] Jerry Fodor, 1988. Concepts: Where Cognitive Science Went Wrong, Oxford University Press.

[8] Giancarlo Guizzardia and Nicola Guarino, 2024. Semantics, Ontology, and Explanation, Data & Knowledge Engineering (to appear).

[9] Zellig S. Harris. 1954. Distributional Structure. Word 10, pp. 146–62

[10] Jerry Hobbs. 1985. Ontological promiscuity. In Proceedings. of the 23rd Annual Meeting of the Association for Computational Linguistics, Chicago, Illinois, 1985, pp. 61–69.

[11] George Lakoff. 1987. Women, Fire, and Dangerous Things: What Categories Reveal About the Mind, University of Chicago Press.

[12] Doug Lenat and Guha, R. V. 1990. Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project. Addison-Wesley.

[13] Martha Lewis and Melanie Mitchell. 2024. Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in LLMs, https://arxiv.org/abs/2402.08955

[14] Jesse Lopes. 2023. Can Deep CNNs Avoid Infinite Regress/Circularity in Content Constitution? Minds and Machines, vol. 33, pp. 507-524.

[15] Friederike Moltmann, 2013. Abstract Objects and the Semantics of Natural Language, Oxford University Press.

[16] Peter Milne. 1986. Frege's Context Principle, Mind, Vol. 95, No. 380, pp. 491-495.

[17] Richard Montague. 1973. The Proper Treatment of Quantification in Ordinary English. In: Kulas, J., Fetzer, J.H., Rankin, T.L. (eds) Philosophy, Language, and Artificial Intelligence. Studies in Cognitive Systems, vol 2.

[18] Walid Saba. 2020. Language, Knowledge, and Ontology: Where Formal Semantics Went Wrong, and How to Go Forward, Again, Journal of Knowledge Structures and Systems, 1 (1): 40-62.

[19] Walid Saba. 2007. Language, logic and ontology: Uncovering the structure of commonsense knowledge, International Journal of Human Computer Studies, 65(7): 610-623.

[20] Wesley Salmon. 1989. Four decades of scientific explanation, in: P. KITCHER & W. SALMON (Eds) Minnesota Studies in the Philosophy of Science, Vol. XIII (Minnesota, University of Minnesota Press), pp. 3-21.

[21] Barry Smith, 2005. Against Fantology, In J. Marek and E. M. Reicher (eds.), Experience and Analysis, pp. 153–170

[22] Fred Sommers. 1963. Types and ontology. Philosophical Review, 72 (3), pp. 327-363.

[23] John Sowa. 1995. Knowledge Representation: Logical, Philosophical and Computational Foundations, PWS Publishing Company, Boston.

[24] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, Subbarao Kambhampati. 2023. On the Planning Abilities of Large Language Models - A Critical Investigation, In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

[25] Vaswani, A., Shazeer, N., et. al. 2017. Attention is All You Need, Available online at https://arxiv.org/abs/1706.03762.