Using what LLMs Have Learned in Discovering the Ontology of Natural Language

Anonymous ACL submission

Abstract

Large language models (LLMs) have undeniably reached impressive human-like performance on a number of tasks. However, and independent of scale and the underlying neural architecture, LLMs (i) are susceptible to the toxicity and bias in their training data; (ii) are hopelessly unexplainable since their knowledge will always be buried in billions of weights (microfeatures); and (iii) have no access to the ground truth since for these models all text (factual and nonfactual) was created equal. To get at an explainable and more predictable theory of how language works what we suggest here is to systematically reverse engineer what LLMs have learned about language, thereby uncovering the conceptual structure and the ontology that must be implicit in our natural languages.

1 Introduction

There are two competing philosophical viewpoints as to the proper scientific method: a rationalist (top-down) method and an empirical (bottom-up) method (Salmon, 1989). In the early days of natural language processing (NLP) the rational top-down approach dominated the field using theories like generative linguistics (Chomsky, 1957), cognitive linguistics (Lakoff, 1987), and model-theoretic semantics (Montague, 1974), but these theories lacked solid foundations since they did not start from a set of generally agreed upon principals, which inevitably led to their limited success. In contrast, and particularly after the introduction of a transformer with an attention mechanism model (Vaswani et. al., 2017), empirical bottom-up methods have made remarkable progress as exemplified by large language models (LLMs) that are essentially the result of a massive experiment in a data-driven, bottom-up reverse engineering of language at scale. The undeniably impressive capabilities of these LLMs is arguably challenging some previously held beliefs in linguistics and cognitive science including some philosophical skepticism about the possibility of machine understanding of language. For instance, criticisms involving Dreyfus's relevancy argument (Dreyfus, 1972) and John Searle's Chinese Room Argument (Searle, 1980) appear less convincing in the face of LLMs' proficiency in syntax, semantics, and even their apparent capturing of some commonsense knowledge. However, and notwithstanding their impressive linguistic fluency, LLMs have a number of limitations. For one thing, LLMs are not *really* models of language, but statistical models of regularities found in our linguistic communication. Due to their subsymbolic nature, all 'knowledge' acquired by LLMs will always be buried in billions of microfeatures (weights), none of which is meaningful on its own and thus one cannot hope to have an explainable theory of how language actually works. Furthermore, (i) LLMs will always be susceptible to the bias and toxicity found in their training data (Bender et. al., 2021); and (ii) can never be fully trusted concerning the "truthfulness" of the content they generate (Borji, 2023). For other shortcomings see (Riedl, 2023) and (Saba, 2023).

Given the current state of affairs we believe an important question that one can now ask is this: to remedy their paradigmatic limitations, can we use what LLMs 'learned' about language to discover the conceptual structure (along with the ontology) of natural language? We believe there is a positive answer to this question. In particular, we believe that the symbolic dimensions of meaning and the ontological structure implicit in natural language can be reverse engineered by querying the massive syntactically valid and semantically plausible patterns that LLMs have learned (memorized?). In the remainder of this short paper we discuss how this could be done.

2 Concerning 'the Company a Word Keeps'

The genesis of modern LLMs lies in distributional semantics (Lenci & Sahlgren, 2023, Ruder, 2016) which claims that semantic equivalence and similarity amount to the similarity of contexts in which words occur; or stated in other words, similarity in meaning is similarity in linguistic distribution (Harris, 1954). This idea is well captured by the now quite famous slogan "you shall know a word by the company it keeps" (Firth, 1957). While early models of word embeddings did well in approximating word meanings (by approximating 'lexical semantics'), it was not until the transformer model with attention mechanisms (Vaswani et al, 2017) that neural language models started the encoding of syntax and quite a bit of semantics. This has been achieved by creating embeddings for valid sequences of words rather than just single words. Stephen Wolfram (2023) commented on the scale of such models, noting that their size is often comparable to the size of the training data, suggesting a limitation in their explanatory power and hinting at a possible memorization of all word combinations in context.

Regardless of the details of how we got here, at the root of the 'embedding' approach is the idea that word meanings can be inferred by the words' context, i.e., via their usage (Boleda, 2020). While this is profoundly different from top-down approaches and the use of symbols, nothing precludes us from carrying out this bottom-up and data-driven approach in a symbolic setting and these ideas have in fact been discussed going back to Frege. We discuss this idea below.

3 Symbolic Reverse Engineering of Language at Scale

In discussing possible models (or theories) of the world to be used in computational linguistics Hobbs (1985) considered two possibilities: either we build a "complete and correct" theory, comprising a full description of the world, involving quantum physics and all the sciences, or we can build a promiscuous model of the world that is isomorphic to the way we talk about it in natural language. What Hobbs is suggesting here is clearly a reverse engineering approach to reveal how we use language to talk about the world we live in. In essence, this is not really different from Frege's Context Principal that suggests to avoid asking "for the meaning of words in isolation" (Dummett, 1981). That is, to capture the possible meanings of words we need to analyze the various contexts in which they may occur (Milne, 1986).

Again, what this suggests is that the actual meaning of words is conveyed or embedded (to use a modern terminology) in their usage, i.e., the ways they are used by humans when talking about the world. While Hobbs' and Frege's observations seem a bit vague, Sommer's (1963) proposal was very specific. He 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 same conclusion: the complete knowledge of some word *w* entails all its usages in every possible sentence. 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*. For example, while [*delicious apple*] is quite sensible, [*delicious Thursday*] is not. Moreover, since [*delicious cake*] is also sensible, there must be a common type (perhaps **food**?) subsuming both 'apple' and 'cake'. This idea seems similar to the idea of type checking in 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 suggested elsewhere (Saba, 2007), this type of reasoning can be used not only to discover the dimensions of word meanings, but also for the 'discovery' of the ontology underlying natural languages (Arp et al. 2015; Studer & Staab, 2004).

But how do determine all what can 'sensibly' be said of some concept, *C*? We will turn to this question below, but first let us assume that *C* can be any concept, physical or abstract, including abstract concepts such as properties, events, activities, states, etc. For example, an adjective such as 'wise' corresponds ontologically to the abstract entity (a property, in the case), called *wisdom* (see Moltmann, 2013). Thus, saying *Sara is wise* is essentially the same as saying *Sara has the wisdom property*. Similarly, saying *David is ill* is exactly the same as saying *David is in an illness state*. Note, further, that *Sara has the wisdom property* is a statement that expresses a relationship between the two entities *Sara* and *wisdom* (which we call this relationship **hasProp**)

while *David is ill* expresses a relationship between *David* and *illness* (a relationship that we call **inState**). What we now have are triples such as <*Sara*, **hasProp**, *wisdom*> and <*David*, **inState**, *illness*>. We consider relations such as **hasProp** and **inState** to be part of the dimensions of meaning, since we believe that they determine how we talk about the world we live in. Other dimensions are those that describes what an entity can do and what can be done to/with an entity. For example, in *Das Capital changed many opinions on capitalism*, we have a *book* that is the **agentOf** change; while *Dave read The Prince* we have a book that is the **objectOf** a reading event. The dimensions of meaning that we consider are the following (see):

- **agentOf**
- **objectOf**
- **hasProp**
- **inState**
- **inProcess**
- **hasPart**
- **partOf**
- **subtypeOf**

Note that in saying x-**agentOf**-y we are not making a statement truth or falsehood some fact, only that it is sensible to say that x-**agentOf**-y. For example, while *Das Capital* might have never inspired anyone, it is sensible (or it makes sense) to say *Das Capital inspired many people over the years*. Finally, given a concept e_1 the value along any some dimension **d** can be thought of as a set of pairs (e_2, w) where e_1 is an entity (abstract or physical) and where w is a weight that represents how frequent (or common) is the relation **d** between e_1 and e_2. For example, and assuming that $w_1 > w_2$ then the following

($influence$, w_1) $\in$ *book* . **hasProp**
($profoundness$, w_2) $\in$ *book* . **hasProp**

would correspond to the fact that the property *influence* is ascribed to books (in ordinary language use) a lot more than *profoundness*.

Now how do we compute a set such as *book* . **hasProp** (i.e., the set of weighted properties that can sensibly be ascribed to a book); or the set *book* . **agentOf** (i.e., the set of weighted activities that a book can be the agent of)? The answer is by consulting the massive machines we call large language models since they have presumably averaged out all the 'sensible' ways we talk about things in ordinary spoken language.

4 UnMASKing LLMs

Virtually all publicly available LLMs provide a public function that can be queried to suggest replacement tokens to a [MASK] (in some LLMs we must use <mask>). To populate the set of activities that some *C* can sensibly be the object of we can execute a call such as

```
objectOf
= predict("Everyone enjoys [MASK] a good " + C + ".")
```

For the word 'book' (and using one LLM) we get the following:

```
<reading, 0.76>
<writing, 0.64>
<reviewing, 0.61>
<editing, 0.58>
etc.
```

To populate the set of properties that can sensibly be said of some *C* we can execute a call such as

```
hasProp
= predict("This is a very [MASK] + C + ".")
```

One specific LLM would produce the following for the word 'book'

```
<interesting, 0.81>
<famous, 0.70>
<entertaining, 0.63>
<provoking, 0.59>
etc.
```

Thus a book can be (mostly) interesting, famous, entertaining, provoking, etc. Finally, we should note that the result of this experiment will also yield the ontological structure that is implicit in how we talk about the world, since a subset analysis of these sets would translate into a hierarchical structure. Moreover, entirely different sets for the same lexeme would be an indication that we are using the same label to refer to another meaning or to a metaphor, while very similar sets for different tokens would correspond to synonymity. In all cases, analyzing how we talk about things, tells us what things there are, as some of the most penetrating minds in linguistics, cognitive science and philosophy have suggested a couple of centuries ago. Now that our collective ingenuity

has resulted in large language models that have encoded how we talk about the world in ordinary language, we can systematically reverse engineer these models to 'discover' the conceptual structure of language, as well as the ontological structure of the world we talk about in language.

5 Concluding Remark

Admittedly, the proposal discussed here is in the very early stages. Although quite a bit of experiments have already been conducted (on a number of LLMs), there is a lot to do making sure we are conducting the 'right experiment', for example that we are executing the right query on every dimension, and that we are subsequently analyzing the results properly as far as metaphor, lexical ambiguity, and the ontological structure that we will discover to be implicit in our findings.

There is much to be done, but we believe that this 2-pass approach is very promising: in the first pass we managed to construct LLMs that saved (in the weights of a massive neural network) all the linguistic patterns we seem to use to talk about the world; in the second phase we systematically query what the LLM has learned to discover the conceptual structure underneath our language(s).

References

Bender, E. M., Gebru, T. et. al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623

Boleda, G. (2020). Distributional Semantics and Linguistic Theory, *Annual Review of Linguistics*, 6:213–34.

Borji, A. (2023). A Categorical Archive of ChatGPT Failures, Online at https://arxiv.org/abs/2302.03494

Chomsky, N. (1957). *Syntactic Structures*, Mouton de Gruyter, New York.

Dreyfus, H. (1972), *What Computers Still Can't Do: A Critique of Artificial Reason*, Cambridge, MA, MIT Press.

Dummett, M. (1981). *Frege: Philosophy of Language*. Harvard University Press.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955, In *Studies in Linguistic Analysis* (Oxford: Philosophical Society, pp. 1-32

Harris, Z. S., (1954). Distributional Structure. *Word* 10:146–62

Hobbs, J. (1985). Ontological promiscuity. In *23rd Annual Meeting of the Assoc. for Computational Linguistics*, Chicago, Illinois, 1 985, pp. 61–69

Lakoff, G. (1987). Women, Fire, and Dangerous Things: What Categories Reveal About the Mind, University of Chicago Press.

Lenci, A., & Sahlgren, M. (2023). *Distributional Semantics*. Cambridge Uni versity Press.

Milne, P. (1986). Frege's Context Principle, *Mind*, Vol. 95, No. 380, pp. 4 91-495

Moltmann, F. (2013). *Abstract Objects and the Semantics of Natural Language*, Oxford University Press.

Montague, R. (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.

Riedl, M. (2023). A Very Gentle Introduction to Large Language Models without the Hype. https://mark-riedl.medium.com/a-very-gentle-introduc tion-to-large-language-models-without-the-hype-5f67941fa59e

Ruder, S. (2016). An overview of word embeddings and their connection to distributional semantic models. Blog, AYLIEN, https://aylien.com/blog/overview-word-embeddings-history-word2vec-cbow-glove

Saba, W. (2023), Stochastic LLMs do not Understand Language: Towards Symbolic, Explainable and Ontologically Based LLMs, In *42nd International Conference, ER 2023*, 6-9, LNCS 14320, Springer.

Saba, W. (2007). Language, logic and ontology: Uncovering the structure of commonsense knowledge, *Int. J. of Human Computer Studies*, 7(65): 610-623

Salmon, W. (1989). Four decades of scientific explanation, in: P. Kitcher & W. Salmon (Eds) *Minnesota Studies in the Philosophy of Science*, Vol. XIII (University of Minnesota Press), pp. 3-21

Searle, J. (1980). Minds, Brains, and Programs, *Behavioral and Brain Sciences*, Volume 3, pp. 417–457.

Sommers, F. (1963). Types and ontology. Philosophical Review, 72 (3), 32 7-363.

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

Wolfram, S. (2023). What is ChatGPT Doing, and Why Does it Work? https://shorturl.at/EKNP9