When classifying grammatical role, BERT doesn't care about word order...except when it matters

Anonymous ACL submission

Abstract

Because meaning can often be inferred from lexical semantics alone, word order is often a redundant cue in natural language. For example, the words *cut*, *chef*, and *onion* are more likely used to convey "The chef cut the onion," not "The onion cut the chef." Recent work has shown large language models to be surprisingly word order invariant, but crucially has largely considered natural *prototypical* inputs, where compositional meaning mostly matches lexical expectations. To overcome this confound, we probe grammatical role representation in BERT and GPT-2 on *non-prototypical* instances. Such instances are naturally occurring sentences with inanimate subjects or animate objects, or sentences where we systematically swap the arguments to make sentences like "The onion cut the chef". We find that, while early layer embeddings are largely lexical, word order is in fact crucial in defining the later-layer representations of words in semantically non-prototypical positions. Our experiments isolate the effect of word order on the contextualization process, and highlight how models use context in the uncommon, but critical, instances where it matters.

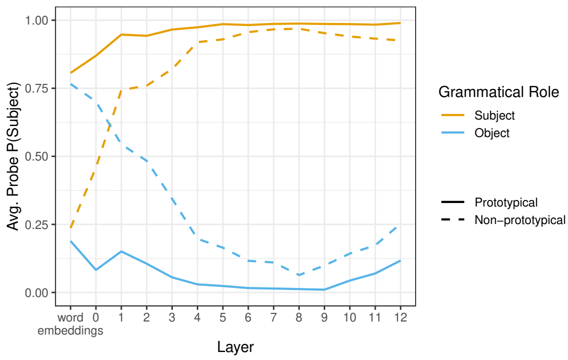

Figure 1: Probabilities of probes trained to differentiate subjects from objects in BERT embeddings. We separate our evaluation examples by prototypicality: whether the grammatical role is what we would expect given the word out of context. The majority of natural examples are prototypical (solid lines), and so if we average all cases we cannot see that grammatical information is gradually acquired in the first half of the network for cases where lexical information is non-prototypical. The equivalent figures for GPT-2 are in Appendix A.

1 Introduction and Prior Work

Large language models create contextual embeddings of the words in their input, starting with a static embedding of each word and progressively adding more contextual information in each layer (Devlin et al., 2019; Brown et al., 2020; Manning et al., 2020). While these contextual embedding models are often praised for capturing rich grammatical structure, a spate of recent work has shown that they are surprisingly invariant to scrambling word order (Sinha et al., 2021; Hessel and Schofield, 2021; Pham et al., 2019; Gupta et al., 2021; O'Connor and Andreas, 2021) and that grammatical knowledge like part of speech, often attributed to contextual embeddings, is actually also captured by fixed embeddings (Pimentel et al., 2020). These results point to a puzzle: how can syntactic contextual information be important for language understanding when the words themselves, not their order, are what matter?

We argue that this apparent paradox arises because of the redundant structure of language itself. Lexical distributional information alone captures a great deal of meaning (Erk, 2012; Mitchell and Lapata, 2010), and the local coherence of words is crucial for constructing meaning in both humans (Mollica et al., 2020) and machines (Clouatre et al., 2021). Viewing this redundancy from the perspective of **grammatical role** (whether a noun is the subject or the object of a clause), most clauses are **prototypical**: in a sentence like "the chef cut the onion", the grammatical roles of *chef* and *onion* are clear to humans from the words alone, without word order or context (Futrell et al., 2019, experiments in English and Russian). This means syntactic word order is redundant with lexical semantics. Whether hand-constructed or corpus-based, most studies probing contextual representations have used prototypical sentences as input, where syntactic context does not have much information to contribute to core meaning beyond the words themselves.

Yet human language can use syntax to deviate from the expectations generated by lexical items alone: we can also understand the absurd meaning of a rare **non-prototypical** sentence like "The onion cut the chef" (Gibson et al., 2013).

In this paper we evaluate BERT and GPT-2[1] on these rare non-prototypical examples, where the meaning of words in context is different from what we would expect from looking at the words alone. We train grammatical role probes on layer embedding spaces to examine the progression of grammatical representation through the layers. We focus on grammatical role since it is used to encode the basic compositional semantic structure of a sentence (Dixon, 1979; Comrie, 1989; Croft, 2001). While fixed lexical semantics contain information about grammatical role (animate nouns are likely to be subjects, etc), the grammatical role of a word in English is ultimately defined by syntactic word order. Probing grammatical role lets us examine the interplay between syntax and lexical semantics in forming compositional meaning.

Our experiments highlight two key findings. First, lexical semantics play a key role in organizing embedding space in early layer representations, and non-lexical compositional features are only expressed in later layers (Experiment 1, Figure 1). Second, if we control for distributional cooccurence factors by creating **argument swapped sentences** (like "The onion cut the chef", real sample in Appendix B), embeddings still represent meaning that is imparted *only* by syntactic word order, overriding lexical and distributional cues (Experiment 2, Figure 2). More generally, we highlight the importance of examining models using non-prototypical examples, both for understanding the strength of lexical influence in contextual embeddings, but also for accurately isolating syntactic processing where it is taking place.

2 Why non-prototypical probing?

As opposed to more general syntactic probing tasks (e.g., dependency parsing), grammatical role is a linguistically significant yet specific task that is both syntactic *and* semantic. As such, we can choose these linguistically-informed sets of non-prototypical examples where lexical semantics do not match the compositional meaning implied by the syntax.

Non-prototypical examples give us a unique perspective on how syntactic machinery like word order influences compositional meaning representation *independently* from lexical semantics. Studies in probing have controlled for lexical semantics by substituting content words for nonce words ("jabberwocky" sentences, as in Maudslay and Cotterell, 2021; Goodwin et al., 2020) or random real words ("colorless green idea" sentences, as in Gulordava et al., 2018). A tradeoff is that these methods lead to out-of-distribution sentences whose words are unlikely to ever co-occur. Rather than bleaching any effect of lexical semantics, our setup lets us examine the interplay between lexical semantics and syntactic representation in a controlled environment, isolating the effects of syntactic word order while using in-distribution examples.

Recent work on representation probing has focused on improving probing methodologies to make sure that extracted information is not spurious or not simply lexical (Hewitt and Liang, 2019; Belinkov, 2021; Voita and Titov, 2020; Hewitt et al., 2021; Pimentel et al., 2020). Our experiments are a complementary approach, where we use standard probing methods, but use linguistically-informed *data selection* to address the ambiguity of what classifiers are extracting.

3 Experiment 1: Grammatical Subjecthood Probes

In Experiment 1, we evaluate grammatical role probes on prototypical instances, where grammatical role lines up with lexical expectations, and non-prototypical instances, where it does not.

3.1 Methods

We train a 2-level perceptron classifier probe with 64 hidden units to distinguish the layer embeddings of nouns that are *transitive subjects* from nouns that are *transitive objects*, as in Papadimitriou et al. (2021). We train a separate classifier for each model layer, as well as training a classifier on the static word embedding space of the models without the position embeddings added (be-

[1]Results are similar for the two models, so we visualize BERT results here, and include GPT-2 figures in App. A.

fore layer 0). Our classifiers are binary, taking the layer embedding of a noun and predicting whether it is a transitive subject or a transitive object. Our probe training data comes from Universal Dependencies treebanks: we pass single sentences from the treebanks through the models, and use dependency annotations to label each layer embedding for whether it represents a transitive subject, a transitive object, or neither (not included in training). The training set is balanced to include an equal number of subjects and objects (1728 examples total). We use `bert-base-uncased` and `gpt2`. For our analysis, we call a noun a prototypical subject if the probe probability for its word embedding (pre-layer 0) is greater than 0.5, and a prototypical object if it is less [2].

3.2 Results

Prototypical and non-prototypical arguments differ in probing behavior across layers, as demonstrated in Figure 1. For prototypical instances (solid lines), syntactic information is conflated with type-level information and so probe accuracy is high starting from layer 0 (word embeddings + position embeddings), and stays consistent throughout the network. However, when we look at non-prototypical instances (dashed lines), we see that the embeddings from layer to layer have very different grammatical encodings, with type-level semantics dominating in the early layers and more general syntactic knowledge only becoming extractable by our probes in later layers.

Crucially, since prototypical examples dominate in frequency in any corpus, the average probe accuracy across all examples is high for all layers, and the grammatical encoding of subjecthood, which is accurate only after the middle layers of the model, would be hidden. Separating out non-prototypical examples illustrates how the syntax of a phrase can arise independently from type-level information through transformer layers, while also showcasing the importance of lexical semantics in forming embedding space geometry in the first half of the network.

4 Experiment 2: Controlling for Distributional Information by Swapping Subjects and Objects

In Experiment 1 we show that the contextualization process consists of gradual grammatical information gain for non-prototypical examples, even though this is largely obscured in the majority prototypical examples where lexical semantics also contains accurate syntactic information. In this experiment, we ask: does this contextualized information about grammatical role stem from word order and syntax, or from distributional (bag-of-words) effects when seeing all words in the sentence? We answer this question by creating example pairs where we control for distributional information by keeping all the words the same, but swapping the positions of the subject and the object. Such pairs of the type “The chef cut the onion” $\rightarrow$ “The onion cut the chef” have identical distributional information. To accurately classify grammatical role in both sentences, the model we're probing would have to be attuned to the ways in which small changes in word order globally affect meaning.

4.1 Methods

We use the same probing classifiers from Experiment 1, and evaluate on a special test set of pairs of sentences that have the subject and direct object of a clause swapped. To create the swapped sentences, we search for verbs that have lexical direct subjects and direct objects, check that the subject and object have the same number (singular or plural), and also check that neither of them are part of a compound word or a flat dependency word that would be separated. If a sentence contains a verb where its arguments fulfill all of these requirements, we swap the position of the subject and the object to create a second, swapped sentence, and add the sentence pair to our evaluation set. A random sample of our swapped sentences is in Appendix B.

4.2 Results

When testing our probes on pairs of normal and swapped sentences, we find that our probes from Experiment 1 correctly classify both the normal and the swapped sentences with high accuracy in higher layers. Since we test our probes on controlled pairs that have the same distributional information, we can isolate effect of syntactic word order in influencing meaning representation. This is demonstrated in Figure 2, where probe predictions for the same set of words in the same distributional context diverges significantly depending on whether the word is in subject or object position. Our results indicate that, separate from dis-

[2] We plan to release our code for reporoducibility

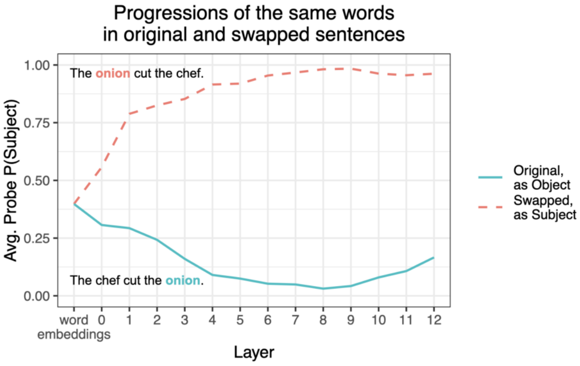

Figure 2: Probe probabilities for the same words when they are the object of an original treebank sentence (eg. “The chef cut the **onion**”, blue line) versus being the subject of that sentence after manual swapping (eg. “The **onion** cut the chef”, dashed red line). When probing the geometry of grammatical role, *the same words in the same distributional contexts* are clearly differentiated throughout contextualization in BERT layers, due to the impact of syntactic word order.

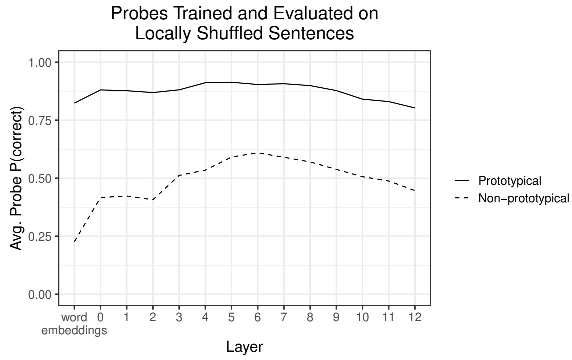

Figure 3: Probe accuracies for sentences where the words have been locally scrambled such that no word moves more than 2 slots. Probe performance for non-prototypical sentences is close to chance, indicating that general positional information (still available after local scrambling) is not enough to recover grammatical role. However, lexical semantics is preserved through layers in these scrambled instances as evidenced by the steady probe performance on prototypical sentences.

tributional effects, models have learnt to represent the ways in which syntactic word order can *independently* affect meaning.

4.3 Are these results just due to general position information?

Our results in Experiment 2 indicate that syntactic word order information can affect model representations of word meaning, even when we keep lexical and distributional information constant. A question still remains: does the divergence demonstrated in Figure 2 stem from the fine-grained ways in which word order influences syntax in English, or from heuristics based on primacy (whether a word is earlier or later in a sentence)? To further investigate this, we train and test probes on sentences where word order is locally scrambled so that no word moves more than 2 slots, and so general primacy is preserved. As shown in Figure 3, probes trained on these locally shuffled sentences do not fare better than chance on non-prototypical examples. This demonstrates that general primacy information is not sufficient to cause the non-prototypical representation we see in Figure 2.

5 Discussion

While recent work has shown that large language models come to rely on distributional semantic information, we consider a rare but important case: the representation's ability to *overcome* these distributional cues. Research showing that models rely on lexical and distributional information is not at odds with our findings that this can be overridden. In fact, even though humans can accurately understand non-prototypical sentences, human syntactic processing is often influenced by the lexical semantics of words, as evidenced by studies on human subjects (Frazier and Rayner, 1982; Rayner et al., 1983; Ferreira and Henderson, 1990) as well as by lexically-influenced syntactic processes in human languages, like differential object marking (Aissen, 2003)—a phenomenon whereby non-prototypical grammatical objects are marked.

What for human language processing is an important source of redundancy—the fact that syntactic cues are often redundant with the information supplied by word meaning—can be, for model interpretability studies, a confound. We have shown that it is easy for a straightforward probing approach to conclude that grammatical role information is available to the lowest layers of BERT. But, by separately analyzing prototypical and non-prototypical arguments, it is clear that the picture is more complicated. At lower layers, BERT representations can classify subjects and objects *most of the time*, but when a non-prototypical meaning is expressed, accurate classification is not available until the higher layers. Insofar as being able to understand non-prototypical meanings is a hallmark of human language processing (Hockett, 1960), we urge future probing studies to consider non-prototypical meanings.

References

Judith Aissen. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, 21(3):435–483.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv preprint arXiv:2102.12452*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2021. Demystifying neural language models' insensitivity to word-order. *arXiv preprint arXiv:2107.13955*.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology*, 2nd edition. University of Chicago Press, Chicago.

William A. Croft. 2001. Functional approaches to grammar. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*, pages 6323–6330. Elsevier Sciences, Oxford.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert MW Dixon. 1979. Ergativity. *Language*, pages 59–138.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Fernanda Ferreira and John M Henderson. 1990. Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):555.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, 14(2):178–210.

Richard Futrell, Evgeniia Diachek, Nafisa Syed, Edward Gibson, and Evelina Fedorenko. 2019. Formal marking is redundant with lexico-semantic cues to meaning in transitive clauses. Poster presented at the 32nd Annual CUNY Conference on Sentence Processing.

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088.

Emily Goodwin, Koustuv Sinha, and Timothy J O'Donnell. 2020. Probing linguistic systematicity. *arXiv preprint arXiv:2005.04315*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. Bert & family eat word salad: Experiments with text understanding. *arXiv preprint arXiv:2101.03453*.

Jack Hessel and Alexandra Schofield. 2021. How effective is bert without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. Conditional probing: measuring usable information beyond a baseline.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Charles F. Hockett. 1960. The origin of language. *Scientific American*, 203(3):88–96.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? experiments with jabberwocky probing. *arXiv preprint arXiv:2106.02559*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Francis Mollica, Matthew Siegelman, Evgeniia Diachek, Steven T Piantadosi, Zachary Mineroff, Richard Futrell, Hope Kean, Peng Qian, and Evelina Fedorenko. 2020. Composition is the core driver of the language-selective network. *Neurobiology of Language*, 1(1):104–134.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? *arXiv preprint arXiv:2106.08367*.

Isabel Papadimitriou, Ethan A Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual bert. *arXiv preprint arXiv:2101.11043*.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *ACL*, pages 4609–4622.

Keith Rayner, Marcia Carlson, and Lyn Frazier. 1983. The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior*, 22(3):358–374.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

A Figures for GPT-2 Experiments

We ran our experiments on both BERT and GPT-2 embeddings, and both models had similar behaviors that we discuss in the paper. For clarity, figures in the paper only visualize the BERT results, and we're including the GPT-2 versions of those same figures for comparison. Figure 4 shows the GPT-2 results of Figure 1, Figure 5 shows the GPT-2 results of Figure 2, and Figure 6 shows the GPT-2 result of Figure 3.

B Sample of argument-swapped sentences

A random sample (not cherry-picked) of our argument-swapped evaluation set, where the subject and the object of clauses are automatically swapped. The original subject is in **bold** and the original object is in ***bold and italics***. The process for creating these sentences is detailed in Section 4.1

On Thursday, with 110 days until the start of the 2014 Winter Paralympics in Sochi, Russia, ***Professor*** interviewed Assistant **Wikinews** in Educational Leadership, Sport Studies and Educational / Counseling Psychology at Washington State University Simon Ličen about attitudes in United States towards the Paralympics.

This ***approach*** shows a more realistic **video** to playing Quidditch.

Second, aggregate ***view*** provides only a high-level **information** of a field, which can make it difficult to investigate causality [23].

A ***hand*** raises her **girl**.

area of the Mississippi River and the destruction of wetlands at its mouth have left the **Alteration** around New Orleans abnormally vulnerable to the forces of nature.

It was known that a moving ***energy*** exchanges its kinetic **body** for potential energy when it gains height.

Thus, when ACPeds issued a statement condemning gender reassignment surgery in 2016 [21], many ***beliefs*** mistook the organization 's political **people** for the consensus view among United States pediatricians — although the peak body for pediatric workers, the American Academy of Pediatrics, has a much more positive view of gender dysphoria [22].

His ***painting*** perfectly combines **art** and Chinese calligraphy.

When the ***inches*** become a few **plants** tall and their leaves mature, it 's time to transplant them to a larger container.

Since the television series' inception, ***reviews*** at The AV Club have written two critical **writers** for each episode:

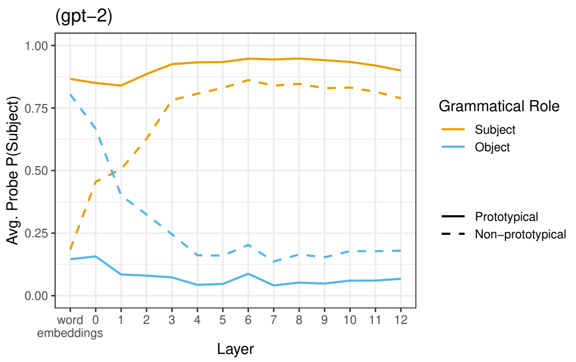

Figure 4: Equivalent to Figure 1 from the main paper, on GPT-2 embeddings

.

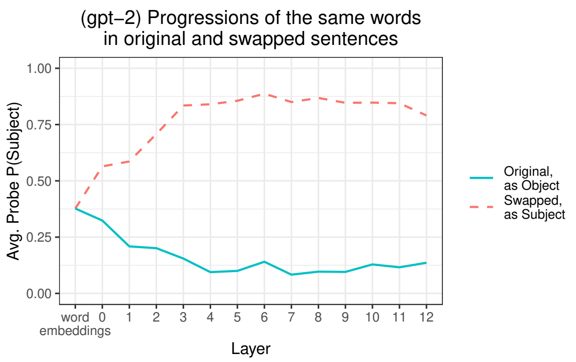

Figure 5: Equivalent to Figure 2 from the main paper, on GPT-2 embeddings. Grammatical representation in GPT-2 embedding also diverges for the same words in the same distributional contexts.

.

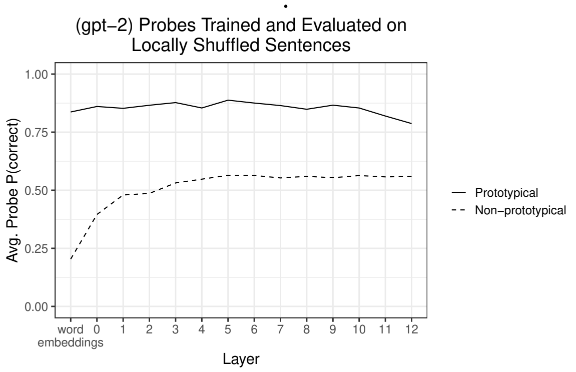

Figure 6: Equivalent to Figure 3 from the main paper, on GPT-2 embeddings. As shown by the dashed line being close to chance, grammatical role information is not extractable from locally shuffled sentences in the non-prototypical cases where lexical semantics do not help

.