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

Isabel Papadimitriou
Stanford University
isabelvp@stanford.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Kyle Mahowald
University of Texas at Austin
mahowald@utexas.edu

While 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., 2020; 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). We argue that a linguistically informed understanding of the role of word order information in human language can illuminate the role of context in contextual embedding models.

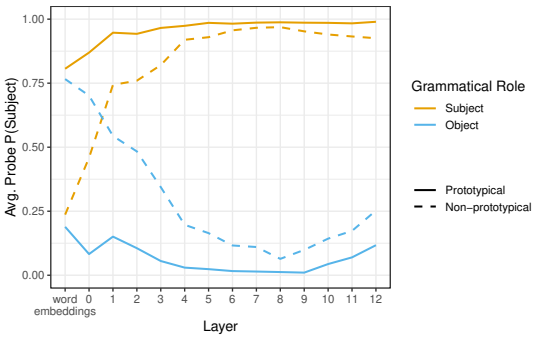

Figure 1: Probabilities of probes trained to differentiate subjects from objects in BERT embeddings. We separate our evaluation examples by prototypicality: whether the grammatical role is what we would expect given the word out of context. The majority of natural examples are prototypical (solid lines), 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.

Our primary experimental method consists of training probing classifiers on the hidden layer embeddings of English BERT (a separate classifier for each layer), to identify whether a noun token is the subject or the object of a transitive sentence. Our experiments rest on comparing the **grammatical subjecthood** of a noun in a sentence (as annotated in the English GUM and EWT UD treebanks) with the **type-level (fixed embedding) subjecthood** prediction: what role we would expect that noun to have if we did not have any context. This allows us to separate prototypical sentences, where the subjects are animate, agentive words (eg, "The man held the umbrella") from

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 203-205.
Held on-line February 7-9, 2022

non-prototypical sentences where the subjects are words generally more likely to be objects (eg, "The umbrella protected the man"). We also perform word order ablations to further understand how structural information arises in the embeddings of non-prototypical examples.

Result 1: Subjecthood is recovered at different layers of BERT, depending on context Prototypical and non-prototypical subjects differ in their probing behaviors between layers (the solid lines in Figure 1). For prototypical subjects, syntactic information is conflated with type-level information and so probe accuracy is high starting from layer 0 (word embeddings + position embeddings), and this stays consistent throughout the network. However, when we look at non-prototypical subjects, 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 later. Since prototypical subjects dominate in frequency in any corpus, if we were to take the average of all examples, we would see a very moderate change in accuracy through layers. Separating out non-prototypical examples clearly illustrates how the syntax of a phrase arises independently from type-level information through transformer layers, while also showcasing the importance of lexical semantics in forming early layer embedding spaces.

Result 2: Word-order information influences grammatical embedding In our first set of results, we do not differentiate between grammatical information that comes from syntactic word order, and that which is derived from distributional co-occurence information. To address this confound, we repeat our experiment with sentence pairs of the type "The chef cut the onion" → "The onion cut the chef", where we take a sentence from the treebank data and swap the positions of the subject and the object, thus swapping their roles. As shown in Figure 2, it is possible to extract accurate subjecthood information from these examples, which consist of the same words in the same distributional context. This shows how grammatical-semantic information in embeddings is in fact independently influenced by syntactic word order.

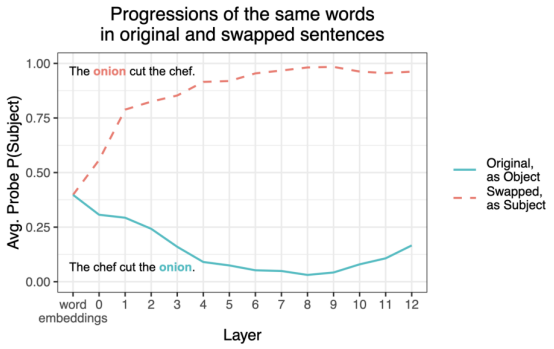

Figure 2: Probe probabilities for the same words when they are the object of an original treebank sentence (blue line) versus being the subject of that sentence after manual swapping (dashed red line). *The same words in the same distributional contexts* are clearly differentiated throughout contextualization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

Result 3: Fine-grained position information matters for the difficult cases A question still remains: does grammatiacal subjecthood embedding stem from the fine-grained ways in which word order influences syntax in English, or from heuristics based on general primacy (whether a word is earlier or later in a sentence)? To further investigate this, we train and test probes on treebank sentences where we randomly scramble the local word order so that no word moves more than 2 slots, and so general primacy is preserved. For non-prototypical cases, probes trained on these locally shuffled sentences cannot fare better than chance (prototypical cases can be classified with relatively high accuracy from just word identity). This demonstrates that general primacy information is not sufficient to cause the grammatical representation of non-prototypical cases that we demonstrate in our previous results.

Conclusion BERT takes advantage of type-level information when it is available, in order to represent information about grammatical role. But, just as humans can understand sentences like "Man bites dog," our probing task on non-prototypical subjects and objects reveals that, in higher layers of BERT, contextual information can override type-level biases using fine-grained syntactic word order information.

References

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2021. Demystifying neural lan-

guage models' insensitivity to word-order. *arXiv preprint arXiv:2107.13955*.

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

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.

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.

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*.

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180*.

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.

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*.