Subject BERT Object

Quinn Langford, Ray Wang, Jessica Boye-Doe, Paul Ihim

Abstract

Papadimitriou et al.'s, "When classifying grammatical role, BERT doesn't care about word order... except when it matters" finds that BERT distinguishes between different word orderings when prototypical sentences (correct semantics) have their subject and object swapped to create non-prototypical sentences (incorrect semantics). We replicated this paper's results, obtaining the probabilities with which the model would categorize nouns in a sentence as the subject or object at each layer. We then proceeded to test BERT on prototypical sentences where the subject and object can be plausibly swapped to create other prototypical sentences. We determined that BERT is still sensitive to word order in early layers even when a sentence's word order does not matter for plausibility. We then evaluated our datasets on a larger BERT model, which initially showed uncertainty but converged to the correct categorization of subject and object with greater certainty than the smaller BERT model. Overall, BERT is found to have some regard for word order in more cases than when word order matters for semantics.

1 Introduction

The advancement in modern natural language processing has greatly enhanced the capability to perform intricate natural language tasks, such as summarization and word prediction, using a variety of specialized models. As these models accomplish these complex tasks, it is essential to evaluate how effectively they capture and represent nuances of language, especially as they continue to grow in size, complexity, and capability. Accordingly, our work examines BERT's representation of word order in the nuanced situation where words can be rearranged without affecting a sentence's plausibility.

While the statistical nature of language empowers these models to excel at tasks trained on vast amounts of data and parameters, some critical questions remain open in the field, including whether these models genuinely comprehend natural language (Bender, 2021). Without a clear understanding of how these models represent language, it is difficult to confirm whether they truly grasp its intricacies, especially in scenarios that challenge established statistical conventions that modern models operate with. This sentiment is expounded upon in Bender et al.'s "On the Danger of Stochastic Parrots", which presents the idea that language models that lack fundamental understanding of language are having an impact on the field of NLP as a whole by not allowing us to understand the limitations of these models. The paper also implies that models without a true understanding prevent new findings from being able to properly address their contribution to NLP. We can examine this directly in the findings of Papadimitriou and Futrell's, "When classifying grammatical role, BERT doesn't care about word order. . . except when it matters", which do not demonstrate our knowledge of BERT's representation of word order in more normal conditions. With these concerns in mind, we intend to examine the basis of language understanding of the BERT model in comparison to the results seen in Papadimitriou and Futrell's paper.

BERT's contextual learning is a powerful tool that enables it to learn representations that encompass their context, but it does not fully address how contexts of different domains (grammatical and semantic) influence word representations. We will investigate BERT's developed representations when the subject and object, separated by a transitive verb, can be transposed, creating a prototypical sentence, a sentence that makes sense grammatically and syntactically. For example, in sentences like '*mother kisses child*' versus '*child kisses mother*', the grammatical order may be inconsequential, but word order has significant implications in regard to the meaning. Our research, which will probe BERT's representation of analogous sentences where word order significantly impacts meaning, aims to pinpoint when BERT accounts for word order, as Papadimitriou and Futrell only counted for cases of prototypical sentences becoming non-prototypical, where semantic properties alone are not sufficient to understand the sentence meaning.

2 Related Work

The syntax of a language is often redundant in practice. For example, word order can determine whether a noun takes the subject or object role, but this information is often already evident from the lexical meanings of the nouns. In a sentence such as '*The person enjoyed the song*', it is clear that the *person* must be the subject because *song* is an inanimate object. As a result, word order is not necessary to assign subject or object roles. In fact, it has been shown that the language regions in the human brain represent lexical information more robustly than syntactic information, likely because lexical information often provides enough information to process language (Fedorenko, 2012). Similarly, in multiple language understanding tasks, BERT is largely resilient to the shuffling of the order of input tokens, suggesting that language models can also predominantly rely on language semantics to attain proficient understanding (Hessel and Schofield, 2021). This implies that fixed embeddings in language models contain grammatical knowledge normally associated with contextualized embeddings.

In their paper, Papadimitriou et al. investigated how BERT handles both prototypical and non-prototypical sentences. The authors used prototypical sentences such as '*The chef chopped the onion*' and rearranged the nouns in the subject and object positions to construct non-prototypical sentences like '*The onion chopped the chef*'. Both types of sentences are grammatically correct, but the non-prototypical sentences place nouns in syntactic positions that are atypical, so attention to syntactical structure is required to understand them. Papadimitriou et al. tracked this understanding in BERT via syntactical probes at different layers. Their findings indicate that BERT focuses on semantic information in earlier layers and progressively integrates syntactic information in later layers. They concluded that BERT accounts for word order in cases where it impacts meaning. One intriguing case not mentioned in Papadimitriou et al.'s study is when the subject and object of a prototypical sentence are switched and yield a plausible sentence. Our primary inquiry is to understand how BERT interprets sentence pairs like these. We investigate how BERT represents word order through 3 experiments in which we replicate Papadimitriou et al. results which evaluated BERT's performance on sentences that were non-prototypical, evaluate BERT's performance on a dataset of prototypically swapped sentences, and use a larger BERT model on the original Papadimitriou et al. dataset as well as our new dataset.

3 Experiment 1: Replicate Original Experiment

We replicated data from the original experiment to evaluate its claims.

3.1 Methods

In order to see if the results of previous work can be easily replicated and are viable as comparison to our own results, our first experiment was a data replication experiment in which we trained and tested a classifier probe on the data from the Papadimitriou et al. paper. As in the original paper, we trained a 2-level perceptron classifier probe with 64 hidden units on the training dataset given in the paper. We then evaluated the probe on the dataset of prototypical and non-prototypical sentences given in the paper. The probe takes the bert-base-uncased embedding of a noun at a particular layer and predicts the likelihood of that noun being classified as a subject. The probe training data came from the Universal Dependencies treebanks and included the GUM-train and EWT-test sets concatenated together. After training the probe, we evaluated it on the dataset.

3.2 Results

The results of Experiment 1 (Figure 1) form a graph similar to the results presented in Papadimitriou et al.'s paper. In the figure, we see that non-prototypical sentences have the subject and object predicted as the opposite category with a high prob-

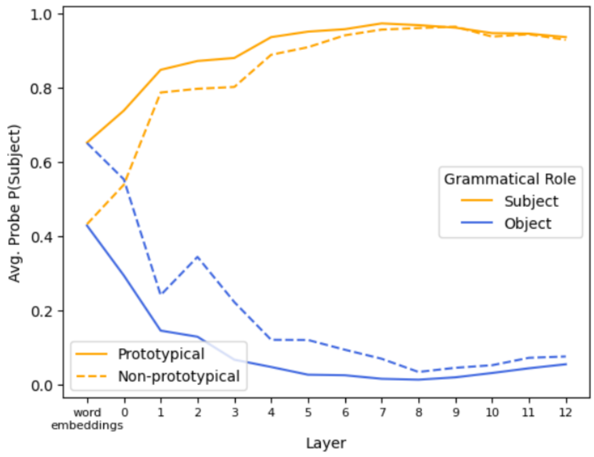


Figure 1: Data replication shows a similar graph to what was initially presented in Papadimitriou et al.'s paper. Solid lines are data from prototypical sentences while dashed lines are data from non-prototypical sentences. For prototypical sentences, BERT has a slightly higher probability of choosing the correct label for each category while for non-prototypical sentences, there is initially a slightly higher probability of choosing the incorrect label. For both, the probabilities of choosing the correct label increase with the number of layers.

ability in the embedding layer, followed by an increase in the correct probability in following layers. More concretely, for non-prototypical sentences there is only slightly greater than a 40% chance that subjects are correctly categorized as subjects in the embedding layer, while there is a greater than 60% chance that objects are categorized as subjects. By layer 3, for these same non-prototypical sentences, there is a greater than 80% chance of subjects being categorized as subjects and less than a 20% chance of objects being categorized as subjects. Prototypical sentences start off with higher probabilities of subjects being categorized as subjects and objects being categorized as objects, which continue to increase in further layers.

4 Experiment 2: Grammatical Subjecthood Probes on Plausible Swapped Sentences

If the BERT model mainly integrates word order in later layers (as previous work suggests), we would expect that if the subject and object of a sentence could be plausibly switched, there would not be much difference between the initial probe classification of the nouns. We would also expect that there may be some fluctuation as BERT sorts out the correct part of the speech label.

4.1 Methods

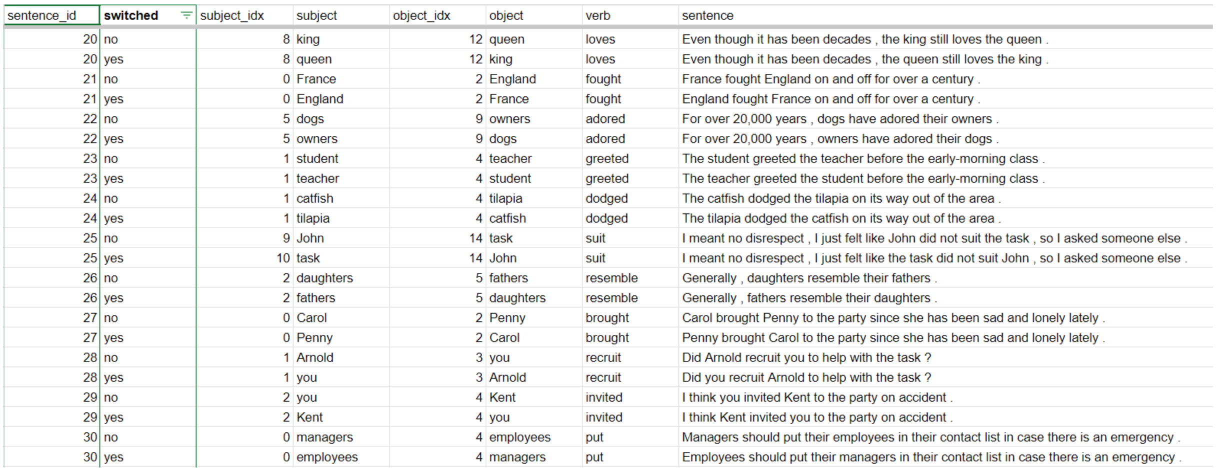
We created a dataset as shown in Figure 2, featuring 106 sentences whose subject and object could be swapped without affecting the plausibility of the sentence. For each sentence, we also included its swapped version in the dataset, so we had 53 pairs of sentences that could be swapped to yield each other. Including a swapped version of every sentence normalizes the data to avoid our own biases in picking which word to designate as the subject or object. As in Experiment 1, we trained a 2-level perceptron classifier probe with 64 hidden units. The probe training data came from the Universal Dependencies treebanks and included the GUM-train and EWT-test sets concatenated together. We evaluated the probe on our dataset of plausibly swapped sentences.

4.2 Results

The results of Experiment 2 (Figure 3) demonstrate that when evaluated only on plausibly swapped prototypical sentences, BERT rapidly gains certainty on the classification of the nouns. There is not much oscillation across the layers in BERT's decision, so it does not appear to change its opinion. In contrast to sentences where the subject and object could not be plausibly swapped, there is virtually no difference between the initial word embedding classification of the nouns in the subject and object positions. Even though the subjects and objects start out at the same value, BERT quickly begins to assign higher probabilities to the words. By as early as layer 1, it is already more than 80% sure about the correct classification, implying that BERT is already paying significant attention to word order very early. Word order does not matter for the plausibility of our sentences, yet BERT was sensitive to word order even at earlier layers, which challenges the previous work's claims that BERT is only sensitive to word order "when it matters."

5 Experiment 3: Larger BERT Model

If increasing the size of the BERT model improves its ability to classify sentence elements, we would expect production of quicker or better representations of the subject and object of a sentence. We would expect this to occur when also testing on non-prototypical sentences.



sentence_id	switched	subject_idx	subject	object_idx	object	verb	sentence
20	no	8	king	12	queen	loves	Even though it has been decades , the king still loves the queen .
20	yes	8	queen	12	king	loves	Even though it has been decades , the queen still loves the king .
21	no	0	France	2	England	fought	France fought England on and off for over a century .
21	yes	0	England	2	France	fought	England fought France on and off for over a century .
22	no	5	dogs	9	owners	adored	For over 20,000 years , dogs have adored their owners .
22	yes	5	owners	9	dogs	adored	For over 20,000 years , owners have adored their dogs .
23	no	1	student	4	teacher	greeted	The student greeted the teacher before the early-morning class .
23	yes	1	teacher	4	student	greeted	The teacher greeted the student before the early-morning class .
24	no	1	catfish	4	tilapia	dodged	The catfish dodged the tilapia on its way out of the area .
24	yes	1	tilapia	4	catfish	dodged	The tilapia dodged the catfish on its way out of the area .
25	no	9	John	14	task	suit	I meant no disrespect , I just felt like John did not suit the task , so I asked someone else .
25	yes	10	task	14	John	suit	I meant no disrespect , I just felt like the task did not suit John , so I asked someone else .
26	no	2	daughters	5	fathers	resemble	Generally , daughters resemble their fathers .
26	yes	2	fathers	5	daughters	resemble	Generally , fathers resemble their daughters .
27	no	0	Carol	2	Penny	brought	Carol brought Penny to the party since she has been sad and lonely lately .
27	yes	0	Penny	2	Carol	brought	Penny brought Carol to the party since she has been sad and lonely lately .
28	no	1	Arnold	3	you	recruit	Did Arnold recruit you to help with the task ?
28	yes	1	you	3	Arnold	recruit	Did you recruit Arnold to help with the task ?
29	no	2	you	4	Kent	invited	I think you invited Kent to the party on accident .
29	yes	2	Kent	4	you	invited	I think Kent invited you to the party on accident .
30	no	0	managers	4	employees	put	Managers should put their employees in their contact list in case there is an emergency .
30	yes	0	employees	4	managers	put	Employees should put their managers in their contact list in case there is an emergency .

Figure 2: 10 examples of prototypical sentences we have created. The sentences are in pairs where the subject and object switch places to create a sentence of slightly different meaning. The index of where the subject and object are in each sentence is specified. Punctuation is separated from words to have its own index and not interfere with representations. Find the rest of the dataset linked here.

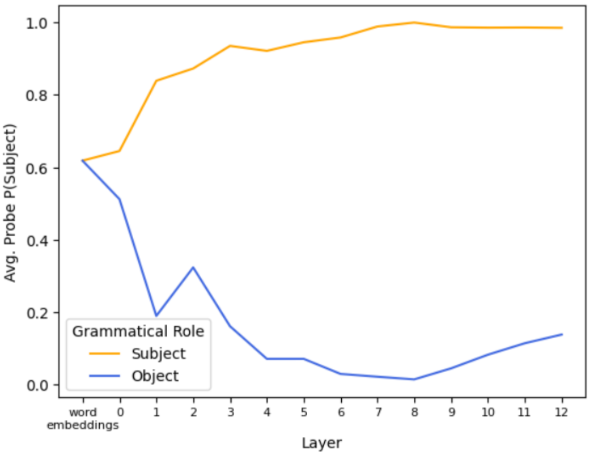


Figure 3: The results of how BERT considers word order for plausibly swapped prototypical sentences. BERT quickly identifies the object and subject and does not change decisions.

5.1 Methods

In order to demonstrate if more tokens impact BERT's ability to choose the subject and object, we increased the size of BERT, evaluated on prototypical and non-prototypical sentences, and analyzed how the graphs we have obtained from Experiments 1 and 2 changed. We tested our larger model on both the set of prototypical and non-prototypical sentences of Papadimitriou et al. and the set of plausibly swapped prototypical sentences from Experiment 2. In the case of prototypical sentences, we examined if BERT was able to come to a greater confidence of the subject and object in earlier layers than it had previously given the change in size. For plausibly swapped and non-prototypical sentences, we examined if a larger BERT model switches between its choices of subject and object in earlier layers, weighing the importance word order and semantic meaning.

5.2 Results

The results of Experiment 3 are shown in Figures 4 and 5 which emphasize that adding more tokens to make the BERT model larger does not make it more certain in its classification of the subjects versus the objects in very early layers. In Figure 4, we tested the Papadimitriou et al. dataset on our larger model to find that BERT still incorrectly classifies the subject and object in early layers for non-prototypical sentences and then switches the classification. However, the larger BERT model seems to take longer to reach a confident decision compared to Experiment 1's Figure 1. In Figure 5, we tested our larger model on our dataset of plausibly swapped prototypical sentences to find that classification percentages also don't move away from uncertainty as fast as they did in the early layers of Experiment 2's Figure 3.

In both Figures 4 and 5, however, there is more certainty about which noun is the subject and the object from layer 9 onwards than there was with the smaller model. This result is most clearly seen in comparison from layers 9 onwards of evaluation of the plausibly swapped dataset. We see that around layer 12 the larger model allows a greater than 95% chance of correctly classifying the subject (and less than 5% chance of misclassifying the object) (Figure 5), while by layer 12 the smaller model results in around a 90% chance of correctly classifying the subject (and a bit more than 15% chance of misclassifying the object) (Figure 3). Therefore, a larger BERT model is more uncertain on classifying objects and subjects in earlier layers, but allows more certainty about the correct classification in later layers.

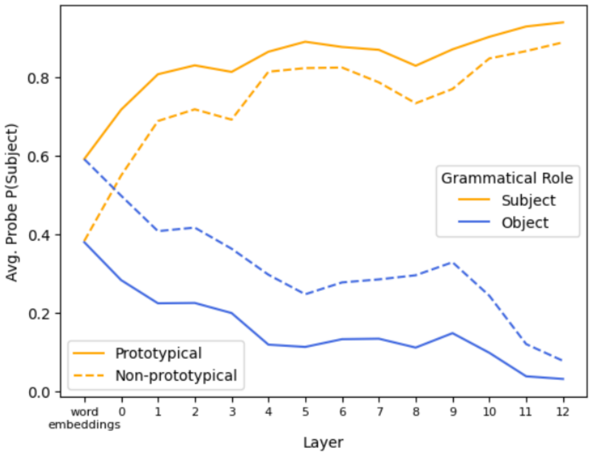


Figure 4: The probability of a word being categorized as a subject using our larger BERT model on the original paper's data. There seems to be more uncertainty when the larger model is used but the final layers here are of greater certainty on the true subject and object than the final layers of the results of the original model.

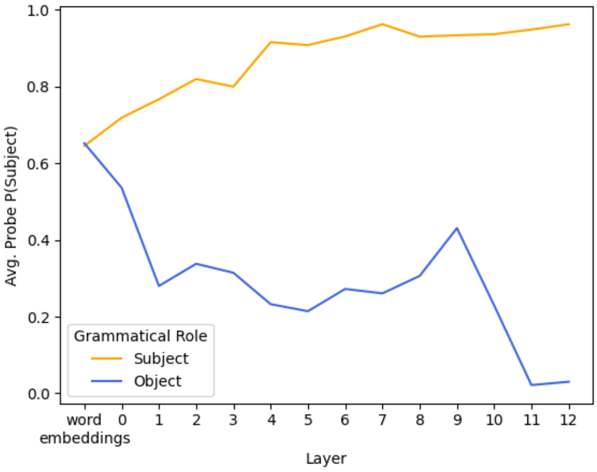


Figure 5: The probability of being categorized as a subject using a larger BERT model on our plausibly swapped prototypical sentences. Similar to results of the large model when evaluated on the original dataset, there is more uncertainty in what is the subject and what is the object in earlier layers than when the original model was used. It can be noted that after a sharp increase in probability of categorizing an object as a subject in layers 7-9, that this probability quickly decreases, showing increased certainty in later layers that outperforms the certainty reached by the original model.

6 Discussion

We have determined that even after a few layers, BERT can confidently decide what the object and subject of a sentence are, even if the subject and object could be plausibly swapped. In Experiment 2, BERT is more uncertain about the subject and object at the beginning, as it evaluates both the subject and object of a sentence as having a 62% chance of being a subject in the embedding layer. This demonstrates that the plausibility of the subject and object being swapped did have an impact on BERT's ability to determine which noun was the subject or the object. However, we see that BERT quickly comes to a conclusion of the proper subject and object without oscillations which demonstrates that even in early layers, word order contributed to BERT considering one as the subject and the other as the object. Since representations in further layers do not show much change in BERT's classification of which item is the subject versus the object, we can conclude that when evaluated on plausibly swapped prototypical sentences, BERT considers word order throughout all layers. We know that changing the word order for the plausibly swapped sentences does not affect the grammatical or semantical correctness of the sentences, so BERT does not necessarily only track word order "when it matters", as previous work claims.

With Experiment 1, we replicated the original Papadimitriou et al. results. This served as a baseline to show that our results from Experiment 2, which disagree with the original paper's, are not a result of technique differences or fundamental differences in evaluating results. Our results, therefore, disagree because they expose the original paper's shortcomings. Papadimitriou et al. did not consider that BERT would need to encode aspects of order for swapped sentences that remain prototypical. Non-prototypicality is not the only way to construct sentences where BERT cares about word order, which was not considered in the original paper.

With Experiment 3, we found that increasing the size of BERT has an impact on the evaluation of prototypical and non-prototypical swapped sentences in earlier layers by decreasing the certainty with which BERT classifies the subject and the object. In cases of both prototypical and non-prototypical sentences, this means that the order in which the words are presented is seen to have less impact on which word to consider as the subject or the object when the model is larger. This could be based on the fact that representations are stronger with a larger model, allowing other properties of a word to be highlighted that may distract from its position, impacting the categorization of subject and object. We still see that there is a consideration of word order, as BERT moves toward the correct choices for subjects and objects. But the lack of certainty shows that the larger model dilutes the impact of word order on BERT. However, by later layers BERT shows more certainty about the correct categorization of subject and object. We conclude that a larger BERT model is more unsure at first and therefore has more flexibility in cases where the subject and object can be plausibly swapped, but also utilizes word order more effectively to come to a more confident decision in the end.

With these results altogether, we can conclude that the BERT model does consider word order in more situations than when sentences can become non-prototypical. BERT is still able to represent word order when we create a new dataset of only plausibly swapped prototypical sentences, showing that word order is represented even when plausibility of meaning isn't drastically changed. Finding that a larger BERT model showed a decrease in certainty of noun classification for earlier layers and an increase in later layers may point to the reliable use of the later layers of larger models in representing word order if this information is desired for some application.

Future research can explore more avenues to consider word order in natural language processing as a whole. Do different parts of speech like prepositions and conjunctions have a similar impact of word order on their classification? If no classification difference is found, how does their word order impact the representation of meaning in the sentence? It may be interesting to test representations of a whole sentence given change in word order by sentiment analysis. It could be determined that sentiment analysis produces drastically different results when word order is changed, prompting more research on how to take word order into account when determining sentiment. With conjunctions, for example, the sentence "*Because the boy went to the market without telling anyone, everyone asked him to buy things although he had already left*", with conjunctions '*because*' and '*although*' switched to create the sentence "*Although the boy went to the market without telling anyone, everyone asked him to buy things because he had already left*", the switch in conjunctions allows a more positive sentiment than the previous sentence. Overall, more research can be done into word order's impacts on many models and strategies of natural language processing.

7 Impact Statement

Our work allows for an increased understanding of the BERT model and its abilities in representing word order. More understanding about the large models we are using allows results obtained with the use of BERT to be more specific and more informed in regards to word order than they had been in the past. Since language models are increasingly infiltrating every aspect of our lives and now have a tremendous potential to harm people, it is imperative that we have a deep understanding of how these models work. Our work is a step towards dismantling the mystery surrounding language models.

We do not anticipate any major ethical issues associated with the language processing techniques described in this paper. We have not crossed major ethical lines, given the model we used takes in random sentences without context made by the previous author and us, with proper credit given. Categorization is given by word type which should not violate major ethical guidelines as it is grammar-based. One point to consider is that our work focuses only on standard English sentences. It does not address any other languages. Languages vary greatly in how word ordering is structured. By leaving out other languages, we acknowledge that our results may over-generalize and may not be applicable to a lot of people. Future work should aim to include other languages in this discussion, especially languages with more rare word orders.

References

Bender et al. “On the Dangers of Stochastic Parrots”. *FAccT ’21*, March 2021, pp. 610-623

Fedorenko E, Nieto-Castañon A, Kanwisher N. Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*. 2012 Mar;50(4):499-513.

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.

Papadimitriou I, Futrell R, Mahowald K. When classifying grammatical role, BERT doesn’t care about word order. . . except when it matters. *arXiv preprint arXiv:2203.06204*.