Garden-Path Traversal within GPT-2

William Jurayj
Brown University
william_jurayj@brown.edu

William Rudman
Brown University
william_rudman@brown.edu

Carsten Eickhoff
Brown University
carsten@brown.edu

Abstract

In recent years, massive language models consisting exclusively of transformer decoders, led by the GPT-x family, have become increasingly popular. While studies have examined the behavior of these models, they tend to only focus on the output of the language model, avoiding analyzing their internal states despite such analyses being popular tools used within BERTology to study transformer encoders. We present a collection of methods for analyzing GPT-2’s hidden states, and use the model’s navigation of garden path sentences as a case study to demonstrate the utility of studying this model’s behavior beyond its output alone. To support this analysis, we introduce a novel dataset consisting of 3 different types of garden path sentences, along with scripts to manipulate them. We find that measuring Manhattan distances and cosine similarities between hidden states shows that GPT-2 navigates these sentences more intuitively than conventional methods that predict from the model’s output alone.

1 Introduction

OpenAI’s release of GPT-3 in 2020 heralded a major step in the field of massive language models, whose ability to generate news articles indistinguishable from those written by humans provides a salient example of their myriad social and political implications, especially when compounded with their propensity to generate bigoted or conspiratorial content (Brown et al., 2020) (Wallace et al., 2019) (Heidenreich and Williams, 2021). Within 2 years of BERT’s release, over 150 studies on had investigated BERT’s structure, exploring how its internal representations enable its powerful and flexible language comprehension (Coenen et al., 2019) (Kovaleva et al., 2019) (Tenney et al., 2019) (Rogers et al., 2020). Studies exploring GPT, however, tend to focus on its external behavior alone, and often only visit it as one of a suite of language modeling architectures.

The few studies that begin to explore its hidden dimensions report concerning results, suggesting under-utilization of its massive latent space (Ethayarajh, 2019) (Cai et al., 2021) (Rudman et al., 2021). One such study finds that a few dimensions exhibit much higher variance than the rest, but while metrics such as Euclidean distance or cosine similarity are dominated by a small subset of the dimensions in the latent space, the behavior of the model can not derived from these dimensions alone (Timkey and van Schijndel, 2021). As massive language models become more ubiquitous and powerful, it will become ever more important to understand the internal processes by which they generate content so they can be streamlined and improved upon. This will require the research community to ask insightful questions and develop innovative methods, since the properties of BERT that researchers investigate often have little to no analog in a transformer decoder model.

In this paper, we use garden path traversal as a case study to demonstrate the value of directly analyzing properties of the embedding space in transformer decoder models. A garden path sentence is one where the most likely parse that a reader expects at some point within the sentence, is proven incorrect by the end of the sentence. By analyzing how GPT-2 sequentially embeds tokens in space, we are able to identify how GPT-2 handles different garden path effects. The contributions of this study are as follows:

- to introduce a robust and diverse dataset of garden path sentences, along with construction functions to negate or extend the effect within each sentence
- to provide methods of analyzing syntactic properties such as garden path effects by ex-
amine geometric relationships between vectors in GPT's hidden states such as Manhattan distance and cosine similarity

- to motivate further study of the hidden states
  GPT and other decoder models maintain as a more thorough alternative to the surprisal-based methods that is typically used to analyze language models.

1.1 Related Work

Many studies into GPT or BERT involve fine-grained analyses of how the model handles specific syntactic phenomena. One such phenomenon is the garden path effect, which occurs when the reader’s preferred parse of the beginning of a sentence is proven inaccurate when they reach the sentence end. For instance, consider the sentence:

1. Even though the girl phoned the instructor was very upset with her for missing a lesson.

2. Even though the girl phoned, the instructor was very upset with her for missing a lesson.

In the first sentence, most readers will assume “the instructor” is the direct object of the verb “phoned”, rather than the subject of the main clause’s verb phrase, “was very upset” (van Schijndel and Linzen, 2019). However, the comma in the second sentence immediately disqualifies the incorrect parse, negating the garden path effect present in the first sentence.

This analysis is typically done by comparing the surprisal, or negative log likelihood, of the token triggering the garden path effect between garden path and negated sentences. Studies conducted on language models’ navigation of these sentences find that sufficiently large models’ relative surprisals at the disambiguating token between garden path and negated sentence forms demonstrate recognition of the garden path effect, but systematically underestimate the magnitude of the effect observed in humans, suggesting that human recovery from an incorrect parse involves more than just the triggering token’s lack of predictability (van Schijndel and Linzen, 2021) (van Schijndel and Linzen, 2018). However, Hu et al. (2020) use surprisal comparisons to show that GPT-2 seems to tackle these effects less successfully than smaller recurrent language models.

Aina and Linzen (2021) take a unique approach using a beam search beginning from the disambiguating token, from which they count the percentage of completions that follow either possible parse to show that both an LSTM model and GPT-2 implicitly consider both possible parses, but ultimately prefer the same parse that human readers do. Their innovative method allows for analysis without requiring a mirror sentence with the garden path effect negated, but still relies on the output of the language model to gain insight into how these models operate.

OpenAI has yet to release GPT-3’s source code, so we instead analyze its predecessor GPT-2, which uses an almost identical architecture at a much smaller scale. Nevertheless, we believe that the methods we use to explore GPT-2’s traversal of garden path constructs can be easily generalized to study any large decoder-based model, including GPT-3, once the model’s hidden states are made accessible.

2 Methods

2.1 Garden path sentence generation

The dataset used for these experiments builds on the combination of the NP/Z and NP/S sentences from van Schijndel and Linzen (2018) and the NP/Z and MV/RR sentences from Futrell et al. (2019). Instead of building out side-by-side datasets of each type of sentence, however, we store the components of these sentences in .tsv files, and include scripts to construct these sentences in various forms similar to those used by Futrell et al. (2019). Examples of each sentence type’s possible forms can be found along with a detailed description of these effects in Table 1 in the appendix.

2.1.1 NP/Z sentences

The garden path effect in these sentences is caused by ambiguity about whether the verb of the leading subordinate clause has a direct object. This is considered one of the stronger types of garden path effects, with an average increase in human reading time of 152 ms (Sturt et al., 1999).

2.1.2 NP/S sentences

The garden path effect in these sentences is caused by ambiguity about whether the noun following the main clause’s verb is that verb’s direct object. This is considered one of the weaker types of garden path effects, with an average increase in human reading times of 50 ms (Sturt et al., 1999).
2.1.3 MV/RR Sentences

The garden path effect in these sentences is caused by ambiguity about whether the past-tense verb of the leading subordinate clause is a past participle or the main verb of the sentence. This effect is considered stronger than that of an NP/S sentence, but reading time data to compare it with the other sentence types is not available.

2.2 Experimental design

The general structure of the tests we run is inspired by Futrell et al. (2019) and Hu et al. (2020), the key difference being that where they compare the model’s surprisal, or negative log likelihood, at the disambiguating word, we examine the hidden state the model has constructed from the sentence up until the previous token. In Figure 1, for instance, where the underlined “was” triggers the garden path effect, we examine the model’s hidden state at the previous word “instructor”, comparing the vector representations of the sentence with and without the negating token, which in this case is the bracketed comma. This is the model’s representation of the entire sentence prior to the trigger, from which the trigger token’s surprisal is calculated, so we are inspecting these sentences at essentially the same inflection point that previous studies have examined.

We compare each sentence to its negated form, computing the vector differences and cosine similarities between each token and its counterpart in the negated form (omitting the token[s] that were added to negate the garden path effect in that sentence type from the pairing process) after re-centering embeddings around the origin. We use Manhattan distance over Euclidean distance to compute scalars from the vector differences between sentences as is generally preferred in high dimensional spaces, and in our specific case moreover because its lack of the quadratic terms present in the Euclidean distance helps preserve the impact of non-rogue dimensions on the metric (Aggarwal et al., 2001). Cosine similarities are computed after re-centering all vectors so that the distribution has a mean of zero, which prevents the metric from defaulting to near-maximum values and allows it to measure the true directional variance between vectors. These side-by-side metrics are generated for all sentences’ garden path and unambiguous forms, as well as for the blocked form of the NP/Z sentences. We expect to see larger distances and smaller similarities between negated and non-negated forms of garden path sentences than between negated and non-negated forms of unambiguous sentences, because in the garden path sentences the negating tokens help to resolve some ambiguity, while in an already unambiguous sentence they will contribute minimally to the sentence’s meaning prior to the triggering token.

3 Results & Discussion

Our analysis reveals several properties of GPT-2’s experience of the garden path effect. Across all sentence types, Manhattan distances and cosine similarities show that the model reacts more heavily to negation of garden path sentences than it does to these sentences’ unambiguous counterparts, as is reflected by surprisal analyses in previous work on recurrent language models. However, this trend in the hidden state metrics is not reflected in GPT-2’s surprisal at the triggering token, highlighting the benefit of examining the model’s behavior beyond its output alone.

Our surprisal baselines showed no consistent trend and suffered from extreme variance between individual sentences of the same type. This finding is in line with previous work; Hu et al. (2020) use surprisal comparisons to score various language models’ ability to generalize syntactic, and find that GPT-2 performs especially poorly on garden path effects, while demonstrating highly variable performance across all categories. On the other hand, the high-level trends we expected to see are present when using either cosine similarity or Manhattan distance, with negation causing a less pronounced difference in unambiguous and blocked sentences than it does in garden path sentences. Whereas Figure 2 shows Manhattan distances to have relatively low variance compared to the other metrics we examine, cosine similarity suffers from very high variances within each sentence form, albeit not quite so extreme as surprisal does. We believe that this is due to Manhattan distance’s resistance to GPT-2’s rogue dimensions, which dominate Eu-
clidean distance and exert a strong effect on cosine similarities as well (Timkey and van Schijndel, 2021) (Aggarwal et al., 2001).

Our analysis revealed a few unexpected results. Most prominent among these is the extent to which the addition of the negating token (“that”) to unambiguous NP/S sentences leaves the hidden representation of the sentence unchanged. The dimension activation diagrams in Figure 3 offer a novel visualization approach for this phenomenon, highlighting the extreme scale of the few rogue dimensions, the consistency in direction that negation of the garden path effect moves the hidden state right before decoding the disambiguating token, and the lack of this distance in the unambiguous forms of NP/S sentences. It is particularly surprising that this difference is most visible in NP/S sentences because in cognitive science and computing literature, this is seen as the weakest of the effect both in humans and in language models. However, this does not contradict previous findings that NPS effects are less surprising than others, but rather complements it—indeed, both in Manhattan distance and cosine similarity, the negated and garden path forms of NP/S sentences are closest together. Rather, this shows that, except in cases where it resolves some clear ambiguity, the negating token in these sentences contributes very little to the model’s internal representation.

Another unexpected observation is that the model finds the blocked form of NP/Z sentences to be far less impacted by negation (in this case, the addition of a comma) than either the unambiguous or the garden path forms. Although neither gap is so stark as it is in the NP/S case shown in Figure 3, these differences can be visualized in Figure 5 in the appendix. A similar difference between garden path and unambiguous forms of MV/RR sentences can be seen in Figure 6. These diagrams help illustrate the common geometric impact that negating the garden path effect has on GPT-2’s hidden state, which is less pronounced in the unambiguous and blocked forms of these sentences.

4 Conclusion

This paper presents a suite of methods to analyze large language models such as GPT-2, taking advantage of a richer reflection of the model’s internal process than can be ascertained from the output of the language modeling head alone. We use novel metrics to show that GPT-2 is affected by garden path sentences in ways that are predictable based on human readers’ difficulty with these sentences. However, the model’s intuitive reaction to these effects is not reflected by the conventional surprisal analysis that has been favored by previous studies on recurrent language models. We hope that these
early insights will help inspire a deeper exploration of the hidden states of decoder-only language models. Possible directions for future work include fine-grained analyses of what information is encoded in the model’s rogue dimensions and what is encoded in the remainder of the dimensions, could more closely examine how information is transformed across different decoder blocks within GPT-2, and might focus on developing visualization techniques such as those used in Figure 4 in the appendix. The methods introduced in this study could also be used to explore decoder models’ handling of syntactic phenomena beyond garden path effects, such as verb subordination or negative polarity item licensing.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In Database Theory — ICDT 2001, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.

Laura Aina and Tal Linzen. 2021. The language model understood the prompt was ambiguous: Probing syntactic uncertainty through generation. CoRR, abs/2109.07848.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Xingyu Cai, Jiiai Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In International Conference on Learning Representations.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viegas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. CoRR, abs/1906.02715.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. CoRR, abs/1909.00512.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state.

Hunter Heidenreich and Jake Williams. 2021. The earth is flat and the sun is not a star: The susceptibility of gpt-2 to universal adversarial triggers. pages 566–573.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1725–1744, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. CoRR, abs/1908.08593.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. CoRR, abs/2002.12327.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2021. Isoscore: Measuring the uniformity of vector space utilization. CoRR, abs/2108.07344.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. Cognitive Science.

Marten van Schijndel and Tal Linzen. 2019. Neural network surprisal predicts the existence but not the magnitude of human syntactic disambiguation difficulty.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. Cognitive science, 45 6:e12988.

Patrick Sturt, Martin Pickering, and Matthew Crocker. 1999. Structural change and reanalysis difficulty in language comprehension. Journal of Memory and Language.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. CoRR, abs/1905.05950.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. CoRR, abs/2109.04404.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp.

A Appendix

A.1 Garden Path Sentence Types

A.1.1 NP/Z sentences

The garden path effect in these sentences is caused by ambiguity about whether the verb of the leading subordinate clause has a direct object. This is
Table 1: Forms of NP/Z, NP/S, and MV/RR sentences included in our dataset, with the verb that triggers or would trigger the garden path effect underlined. Note that all of the perturbations can be combined to avoid the garden path effect, except for the blocked and unambiguous forms of the NP/Z sentence.

| Sentence Type | Sentence form | Sentence |
|---------------|---------------|----------|
| **NP/Z**      | Garden Path   | When the dog scratched the vet took off the muzzle. |
|               | Negated       | When the dog scratched, the vet took off the muzzle. |
|               | Blocked       | When the dog scratched his owner the vet took off the muzzle. |
|               | Unambiguously | When the dog struggled the vet took off the muzzle. |
|               | Extended      | When the dog scratched the vet who was attending to him took off ... |
| **NP/S**      | Garden Path   | The coach discovered the player tried to show off all the time. |
|               | Negated       | The coach discovered that the player tried to show off all the time. |
|               | Unambiguously | The coach thought the player tried to show off all the time. |
|               | Extended      | The coach thought the player that his club was eager to sign tried ... |
| **MV/RR**     | Garden Path   | The horses raced past the barn fell into a ditch. |
|               | Negated       | The horses that were raced past the barn fell into a ditch. |
|               | Unambiguously | The horses ridden past the barn fell into a ditch. |
|               | Extended      | The horses raced past the barn in the countryside fell into a ditch. |

The first sentence evokes a garden path effect because the reader initially expects that “the vet” is the direct object of “scratched” before the word “his owner” reveals that it is the vet’s direct object. The negated form avoids the effect by using a comma to indicate the separation between the two clauses. The blocked form avoids the effect by adding the direct object “his owner” to block the ambiguity that triggers the effect, while the unambiguous form avoids the effect by replacing the transitive verb “scratched” with the intransitive verb “struggled” to avoid ambiguity around the verb’s direct object. Our dataset includes 43 distinct NP/Z sentences, and includes scripts allowing a user to easily transform these into unambiguous or blocked sentences. Moreover, each sentence has the option to include a negation, and an extension as illustrated in the final example, to increase the duration of the ambiguity.

### A.1.2 NP/S sentences

The garden path effect in these sentences is caused by ambiguity about whether the noun following the main clause’s verb is that verb’s direct object. This is considered one of the weaker types of garden path effects, with an average increase in human reading time of 50 ms (Sturt et al., 1999).

The first sentence evokes a garden path effect because the reader expects that “the player” is the direct object of the verb ‘discovered’ before the word “tried” reveals that it is her propensity to show off that the coach is discovering. The negated form avoids the effect by adding “that” before “the player” to eliminate the possibility that ‘the player’ is the verb’s direct object. The unambiguous form avoids the effect altogether by using the verb “thought”, which could not allow a person to be its direct object. The extended form lengthens the effect by adding information between the presumptive object, “the player”, and the verb “tried” that triggers the effect. Our dataset includes 20 distinct NP/S sentences, each of which can be negated, unambiguous, extended, or any combination thereof.

### A.1.3 MV/RR sentences

The garden path effect in these sentences is caused by ambiguity about whether the past-tense verb of the leading subordinate clause is a past participle or the main verb of the sentence. This effect is considered stronger than that of an NP/S sentence, but reading time data to compare it with the other sentence types is not available.

The first sentence evokes the garden path effect because the reader assumes ‘raced’ is the main verb of the sentence, while the negated form negates this ambiguity by clarifying that ‘raced past the barn’ is a descriptor for the horses rather than the main clause itself. The unambiguous form avoids ambiguity altogether by replacing the ambiguous ‘raced’ with the unambiguously passive ‘ridden’, and the extended form illustrates how the effect can be lengthened in this type of sentence. Our
dataset includes 20 distinct MV/RR sentences, each of which can be negated, rendered unambiguous, extended, or any combination thereof.

Figure 4: The cosine similarity between the negated and non-negated forms of a MV/RR garden path sentence. The negating tokens 'that were', which are injected between 'horses' and 'raced' have been omitted. The spikes correspond to the region following the negating tokens and the token directly before the effect’s trigger. Notice the relatively high similarities within the subordinate clause between the dips, and after the effect has been triggered, which correspond to regions where the negating tokens have little effect on the next word prediction objective.

Figure 5: Top: differences in activation of dimensions in GPT-2’s hidden space between negated and non-negated NP/Z garden path sentences. Middle: differences in activation of dimensions in GPT-2’s hidden space between negated and non-negated unambiguous NP/S sentences. Bottom: differences in activation of dimensions in GPT-2’s hidden space between negated and non-negated blocked NP/Z sentences.

Figure 6: Above: differences in activation of dimensions in GPT-2’s hidden space between negated and non-negated MV/RR garden path sentences. Below: differences in activation of dimensions in GPT-2’s hidden space between negated and non-negated unambiguous MV/RR sentences.