Does directionality influence a Language Model's ability to predict garden path effect?

Sarah Cryan
Colgate University
scryan@colgate.edu

Abstract

In this paper, we investigate whether bidirectional or unidirectional language models do a better job at predicting how humans experience garden-path sentences, such as "the old man the boat." We tested 300 sentences on 2 language models—the bidirectional masked model multiBERT and the unidirectional causal model GPT-2—and 5 random seed variations of these specific models. We compared their surprisal outputs to word-by-word human reading times to see which model more closely resembled human results. We found that GPT-2 had greater qualitative similarities to human data, suggesting its closer similarity to replicating the human garden-path reading experience. Further research using a larger data set and more types of unidirectional and bidirectional models could provide stronger conclusions on the relationship between causal models and human reading behavior.

1 Introduction

Our study was designed to answer the following question: what type of language model more closely predicts human language processing? More specifically, we focused on how causal unidirectional and masked bidirectional models process garden-path sentences. The results from our study and final answer to this question can help guide future natural language processing research design, such as model choice, particularly for studies interested in mimicking human behavior.

To answer our question, we used the NLPScholar toolkit to test 5 unidirectional GPT-2 Small models and 5 bidirectional BERT-based models on 50 ambiguous garden-path sentences and 50 correlating unambiguous sentences. All 50 pairs of sentences were further separated into the three main subtypes of garden-path sentences. We compared each model's surprisal difference between these two types of sentences across the three subtypes at the disambiguating word. We compared the differences to results from a study on word-by-word human reading times for garden-path sentences to qualitatively compare model behavior to human behavior, allowing us to answer our main research question.

Since human's read sequentially in one direction, we expected the unidirectional model to behave more similarly to humans. As hypothesized, we found that GPT2's surprisal results across the three subtypes of garden-path sentences more closely resembled human reading times across the same groups of sentences than BERT. Our results suggest that GPT2 and other causal models are better equipped to process text like humans.

2 Background

The term "garden-path" describes sentences that may seem grammatically incorrect despite being grammatically correct. A common example of these sentences is "the horse raced past the barn fell." Upon initial reading, one may want to read the example as two sentences, "the horse raced past" and "the barn fell." However, the lack of punctuation forces them to reevaluate the sentence until the intended meaning becomes clear (additional words can clarify the meaning, for example, "The horse that was once raced past the barn fell down"). Figure 1 provides a syntax tree to illustrate this process with a different example sentence. For linguistic and natural language processing researchers, the confusing garden-path sentences create interesting datasets to explore how people and language models deal with ambiguity. We designed our study based on the practices and previous insights from these published studies.

Human Data on Garden-Path Parsing Because of time and resource limitations, we used the 2024 study by Huang et al. as our comparative human data. They tested 2000 participants across

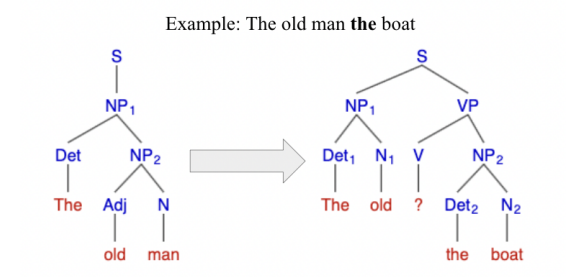

Figure 1: Garden-Path Sentence Correcting Syntax Tree

many garden-path sentence samples, including the main 3 forms we are interested in, MV/RR, NP/Z verb transitivity, and NP/Z overt object. Moreover, they attempted to create direct comparisons between actual reading times and surprisal rates, but only found moderate support (Huang et al., 2024). Therefore,as we conducted our study and drew conclusions, we focused on qualitative comparisons as opposed to quantitative comparisons.

Surprisal as a Measure in Garden-Path Research To determine our evaluation metric, we took inspiration from Futrell et al.'s 2019 study which also compared how bidirectional and unidirectional models processed garden-path sentences. They used surprisal, the $-log_2$ of the probability of a word given the word before it, because it provided word-specific data similar to how human garden-path studies use word-specific reading time data (Futrell et al., 2019). Our study incorporates word-specific human reading time data, further supporting surprisal as an appropriate metric.

Choosing Garden-Path Sentence Subtypes In both of the previously mentioned studies, the researchers looked at Main-Verb/Reduced-Relative, Noun-Phrase/Zero (Overt Object), and Noun-Phrase/Zero (Verb Transitivity) subtypes of garden-path sentences (Futrell et al., 2019; Huang et al., 2024). These sentences provide a comprehensive dataset, encompassing a majority of garden-path sentences that addresses a variety of grammatical ambiguities.

3 Methods

Our general approach to answering our question started with selecting two models, one bidirectional masked model and one unidirectional causal model, and creating five random seeds of each. We then reformatted and preexisting dataset of garden path sentences, so we could input them into the NLP-

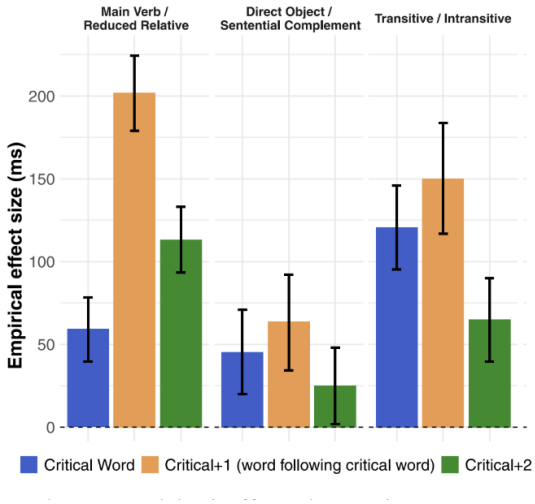

Figure 2: Human Reading Data

Scholar toolkit. The toolkit output surprisal values for each sentence at the disambiguating word, and we compared each model's average results against human reading times for garden-path sentences at the disambiguating word.

3.1 Models

To compare casual and masked model behavior, we used the unidirectional GPT2-small causal model and the bidirectional BERT-base uncased multiBERT masked model. We tested our dataset on 5 unique random seeds of each model, creating slight variation across models while keeping the same hyperparameters. Our approach reduces bias by using these random initialization seeds, allowing for more reliable results to compare against human reading results, and lets us work with more familiar model architectures.

3.2 Datasets

We used a dataset of about 300 sentences from the Zhou et. al's 2024 tree transformer study, containing 50 pairs of ambiguous garden-path sentences and their correlating unambiguous sentences for each garden-path sentence type (MV/RR, NP/Z, and NP/Z) (Zhou 20204). Using a python script, we reformatted these sentences so they were compatible with NLPScholar, a toolkit for running natural language processing experiments (Prasad and Davis, 2024). The toolkit allowed us to run an evaluative MinimalPair analysis, producing the surprisal measures we used to generate our results and compare to human garden-path sentence reading. he scripts and config files used to transform

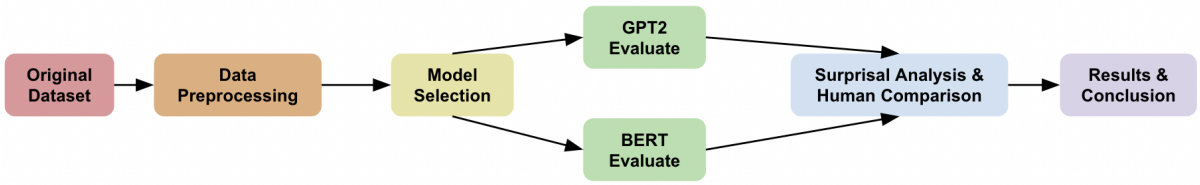

Figure 3: Pipeline for the NLPScholar Experiment

the dataset can be found in our GitHub repository linked here.

3.3 Evaluation

After running NLPScholar on all 6 sets of sentences, ambiguous and unambiguous sentences each sentence type, we found the average surprisal difference between sentence pairs for each model seed and garden-path sentence subtype. Positive differences indicate the model had higher surprisals for the ambiguous garden-path sentences. Additionally, we combined seed surprises to calculate the average surprisal difference between sentence pairs for the two models GPT-2 small and multiBERT overall.

Our analysis focused on comparing GPT-2 small and multiBERT's average surprisal differences across the three sentence subtypes, so we could qualitatively compare their outputs to data on human reading time, helping answer our guiding research question. We interpreted a model's surprisal difference in the same region of interest for a given sentence pair as similar to a human's difference in reading time. When a human reads an unexpected word in a sentence, we expect it would take them longer to read and process the word—the word seems out of place and surprises them. Therefore, we looked at both model's surprisal differences for the garden-path sentence types and observed which more closely resembled human reading times. To more easily compare human-data and model-data, we created bar charts of the averages across seeds and models, as seen in Figure 2 and Figure 3 respectively.

4 Results

Both Models Surprised, but GPT-2's Results More Similar to Human Our results, as illustrated in Figure 4, show that both models had positive suprisal differences across all sentence subtypes, indicating how both models tend to not expect garden-path sentences, like people. GPT-2's suprisal difference average from highest to lowest

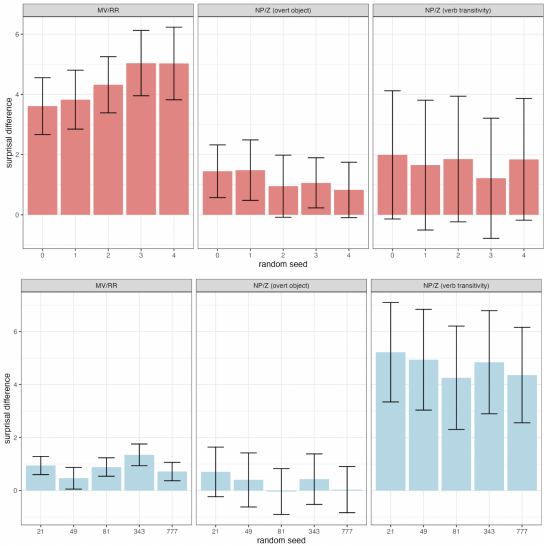

Figure 4: ROI surprisal difference across different conditions predicted by BERT (pink) and GPT-2 (blue) with different seeds

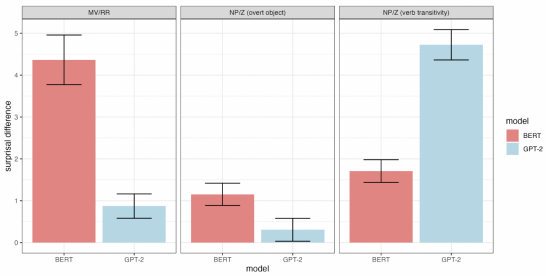

Figure 5: Overall ROI surprisal difference across different conditions predicted by BERT (pink) and GPT-2 (blue)

was NP/Z verb transitivity then MV/RR then NP/Z overt object. BERT's suprisal difference average from highest to lowest was MV/RR then NP/Z verb transitivity then NP/Z overt object. By comparing these results to the human reading-time data for the critical word seen in Figure 1, we observe that GPT-2's data's shape more closely resembles human data, aligning with our hypothesis.

Error in Surprisal for Unique Seeds vs. Overall Model The results for unique seed testing per model across the sentence subtypes, as seen in Figure 3, indicates high variability in our results. For instance, all of the error bars for GPT-2's NP/Z (overt object) seeds and for BERT's NP/Z (verb transitivity) seed extend above the main data column and below zero. The surprisal differences vary greatly for each model's seeds.

As seen in Figure 4, these error bars size relative to the main data column shrink for each model's overall average across the three sentence subtypes, with the exception of GPT-2's NP/Z (overt object) results, providing data to compare to human reading times. Because the seed data's variability contrasts so greatly but is smoothed out through averaging our results, using even more seeds could yield stronger insights.

5 Discussion

Our study aimed to determine whether the unidirectional causal model GPT-2 or the bidirectional masked model multiBERT more closely matched how human's experience garden-path sentences. To answer this broader question, we tested multiple variations of these two models on 150 pairs of ambiguous garden-path sentences and their correlating unambiguous sentence using the NLPScholar toolkit. For each model, we looked at the difference in output surprisal between the sentence pairs, with more positive numbers indicating greater garden-path surprisal.

Our results showed that the masked model output the greatest garden-path surprisal for MV/RR sentences followed by NP/Z verb transitivity followed by NP/Z overt object. the causal model output the greatest garden-path surprisal for NP/Z verb transitivity followed by MV/RR followed by NP/Z overt object. Both model's demonstrated surprisal in response to garden-path sentences, but the GPT-2 model had similar variation across the 3 types of sentences to the variation human reading times had across the 3 types of sentences. Their greater qualitative similarity compared to human reading time and the masked model suggests that unidirectional models more closely resembles human garden-path processing than bidirectional models.

We faced several limitations in our study. For example, we measured the effect size only at the disambiguating word, which may not capture the full extent of the garden-path effect. A wider region of interest, including the disambiguating word and the two following words, like in the Huang et al. study, might provide a clearer picture of processing patterns. Furthermore, we observed large error bars and variability across different random seeds, particularly in GPT-2's result, suggesting that our findings are sensitive to initial weight initialization. Additionally, the NP/Z (Overt Object) condition for GPT-2 yielded insignificant results, likely due to high variability and the limited dataset size. With only 50 sentence pairs per condition, our dataset may not be sufficient for drawing strong conclusions. Finally, we only tested one unidirectional model (GPT-2) and one bidirectional model (multiBERT), limiting our ability to generalize findings about model directionality.

Future work could correct these limitations, expanding the ROI to include the disambiguating word and the two subsequent words thus capturing more holistic processing patterns. Incorporating more random seeds for each model would reduce result variability and provide more reliable conclusions. Increasing the dataset size would give a wider variety of testing sentences leading more precise results. Additionally, testing a broader range of unidirectional and bidirectional models would allow for stronger conclusions about the role of model directionality in handling syntactic ambiguity. These improvements could provide clearer insights into how language models process garden-path sentences and better inform applications that aim to replicate human-like language processing.

Our study compared unidirectional and bidirectional language models to see which better aligns with human processing of garden-path sentences. We found that GPT-2's incremental, left-to-right processing showed greater qualitative similarities to human reading times than multiBERT's bidirectional approach. This suggests that unidirectional models may be better suited for applications requiring human-like sentence processing, though further research with more models and data is needed to confirm this.

Acknowledgements

I would like to acknowledge my research partners Anzi Wang and Brian Kherlen. They were so wonderful to work with and very motivate about our project.

I would also like to thank Forrest Davis and Grusha Prasad for all of their guidance throughout this process and for a great semester!

References

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Grusha Prasad and Forrest Davis. 2024. Training an NLP scholar at a small liberal arts college: A backwards designed course proposal. In *Proceedings of the Sixth Workshop on Teaching NLP*, Bangkok, Thailand. Association for Computational Linguistics.

Lingling Zhou, Suzan Verberne, and Gijs Wijnholds. 2024. Tree transformer's disambiguation ability of prepositional phrase attachment and garden path effects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12291–12301, Bangkok, Thailand. Association for Computational Linguistics.