BERT Shows Garden Path Effects

Tovah Irwin* and **Kyra Wilson***
Research Institute
New York University Abu Dhabi
{tovah.irwin, kyra.e.wilson22}
@ gmail.com

Alec Marantz
Dept. of Psychology and Linguistics
New York University
marantz @ nyu.edu

Abstract

Garden path sentences (i.e. "the horse raced past the barn fell") are sentences that readers initially incorrectly parse, requiring partial or total re-analysis of the sentence structure. Given human difficulty in parsing garden paths, we aim to compare transformer language models' performance on these sentences. We assess a selection of models from the BERT family which have been fine-tuned on the question-answering task, and evaluate each model's performance on comprehension questions based on garden path and control sentences. We then further investigate the semantic roles assigned to arguments of verbs in garden path and control sentences by utilizing a probe task to directly assess which semantic role(s) the model assigns.[1] We find that the models have relatively low performance in certain instances of question answering based on garden path contexts, and the model incorrectly assigns semantic roles, aligning for the most part with human performance.

1 Introduction

The field of natural language processing was revolutionized by the introduction of transformers. Models such as BERT and GPT (and successors) have vastly improved performance on a variety of tasks compared to previous models such as LSTMs. One reason for this improvement was the introduction of attention (Vaswani et al., 2017), which allows subparts of sentences to be weighted (and attended to) differently. Another advance in these models was having all input processed simultaneously rather than sequentially. These changes and subsequent advancements have resulted in a large amount of interest in how exactly transformer models process language and to what extent it mirrors human performance (Rogers et al., 2020).

In terms of syntactic information, investigations have found that BERT represents a variety of phenomena that are presumed to be relevant for human language processing such as parts of speech, roles, and syntactic chunks (Tenney et al., 2019; Liu et al., 2019a). Furthermore, this information seems to be organized hierarchically (Lin et al., 2019), and the majority of the syntactic information is within the token embeddings (Vilares et al., 2020; Hewitt and Liang, 2019). Probing tasks have revealed that BERT contains semantic information as well. Roles, proto-roles, entity types, and relations are contained in token representations in addition to syntactic information (Ettinger, 2020; Tenney et al., 2019).

Because much of the syntactic and semantic knowledge that humans are presumed to have and use is also present in BERT, it is potentially useful to compare the two in linguistic tasks and see if their performance is also similar. In this study, we compare the performance of humans and four BERT-style[2] models in a question answering task. Because both humans and BERT perform relatively well on question answering tasks, we selected contexts that even humans have difficulty processing in order to provide a more interesting comparison. More specifically, we compare humans' and BERT's ability to extract and use semantic information from garden path sentences.

Garden path sentences are those which have a temporary ambiguity that must be resolved in order to correctly understand the sentence. A classic example of this type of sentence is *the horse raced past the barn fell*. Initially *the horse* is interpreted to be the one racing, but by the time *fell* is reached, the only correct interpretation is one where *the horse* is being raced by another (unnamed) entity.

[1] Our code, datasets, and results are publicly available at https://github.com/kyrawilson/gardenBERT.

[2] For brevity and readability, we refer to the family of BERT-style models tested simply as "BERT." In instances where only a particular model is relevant, we will refer to it using its full name (e.g., $\text{BERT}_{\text{BASE}}$).

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3220–3232
May 2-6, 2023

Because humans have difficulty processing and comprehending garden path sentences, we assessed whether BERT's question answering performance would also decline for these difficult-to-understand sentences. Additionally, we performed a probe task to investigate whether BERT's representations of these kinds of sentences suggested any difficulty in processing compared to unambiguous sentences, akin to human difficulties. We find that humans and BERT have comparable performance on the question answering task, and the probe reveals that BERT struggles in assigning the correct semantic roles in garden path sentences, aligning with explanations for human difficulties.

2 Garden Path Sentence Processing

Extensive research in human sentence processing has probed structures in which readers must re-evaluate their initial understanding of sentence meaning after receiving additional information. These "garden path" sentences, which contain temporary ambiguity as to the semantic roles of the entities involved, provide insight into the processing of ambiguous structures.

Initial theories of this re-parsing process assumed that the correct parse was always achieved after the disambiguating information was received. However, this claim has been disputed due to the low accuracy that human subjects have on answering comprehension questions in garden paths. This gives rise to two alternatives: either the correct syntactic structure is never built (Christianson et al., 2001), or the semantic roles from the misparsed structure introduce interference in the correctly parsed sentence (Slattery et al., 2013).

Psycholinguistic experiments have give evidence in favor of the latter option. Slattery et al. (2013) performed a study where participants read sentences such as:

1. (a) After the bank manager telephoned **David's father** grew worried and gave himself approximately five days to reply.
   (b) After the bank manager telephoned **David's mother** grew worried and gave himself approximately five days to reply.

In these sentences, the ambiguous regions (in bold) can be incorrectly parsed as a noun phrase complement to the verb (NP), or the main verb can be correctly parsed as a zero complement verb with no object (Z). Eye tracking results revealed that the correct hierarchical structure was built by the time the reflexive pronoun "him/herself" was read, indicating that processing difficulties were not due to incorrect syntactic structures.

Christianson et al. (2017) further investigated the role of sentence type and ambiguity on human subject's response accuracy to garden path sentences. They contrasted ambiguous versus non-ambiguous and garden path (ambiguous) versus local coherence (unambiguous) structures:

2. **Garden Path**
   (a) Ambiguous
   The player tossed the ball interfered with the other team.
   (b) Unambiguous
   The player who was tossed the ball interfered with the other team.

3. **Local Coherence**
   (a) Ambiguous
   The other team interfered with the player tossed the ball.
   (b) Unambiguous
   The other team interfered with the player who was tossed the ball.

Participants were then asked comprehension questions, such as *did the player toss the ball?*. Participant's comprehension question accuracy was extremely low for the garden path + ambiguous condition, with accuracy below 25% (exact numbers were not reported due to analysis on individual participant responses). Accuracy was higher in all other conditions, with ambiguous local coherence scores ranging from 40-50% and all unambiguous structures reaching scores near 60%.

This means that while human readers are able to correctly reanalyze a complex sentence (Slattery et al., 2013), they may not fully disassociate the initial semantic roles assigned in the first parse from their final interpretation of the sentence's meaning, leading to low comprehension accuracy. Since garden paths involve the interplay of multiple systems in human language processing (semantic and syntactic), this poor human performance raises the question of how language models, some without explicit syntactic training, handle these types of sentences.

3 Related Work

Previous research has addressed the ways in which various models process garden path sentences, although none up to this point has examined BERT specifically, to our knowledge. Utilizing the metric of surprisal extracted from various models, Van Schijndel and Linzen (2018) modeled garden path effects in human self-paced reading. They compared probabilistic context-free grammars (PCFG) with explicit hierarchical syntax to recurrent neural network (RNN) models trained on text without syntactic annotation. Both the PCFGs and RNNs under-predicted the extent to which human readers slowed down in response to NP/Z type ambiguities, showing that these types of models may find garden path sentences less challenging than human readers.

Moving away from human comparisons, Futrell et al. (2019) evaluated a number of models' surprisal in garden path sentences, including three LSTM models and a RNN Grammar trained on a small dataset. All models evaluated showed increase in surprisal values at the disambiguating regions of the NP/Z garden path sentences, but they found that only the larger LSTM models evaluated utilized verb argument structure in their predictions, showing that explicit syntax training is not needed to model garden path effects.

Jurayj et al. (2022) similarly investigated GPT-2's ability to navigate different types of garden paths. They evaluated the change in GPT-2's hidden states before and after the disambiguating component of a garden path sentence. Utilizing Manhattan distances and consine similarities, they found a larger difference before and after the disambiguating token in garden paths compared to unambiguous sentences. Both Futrell et al. (2019) and Jurayj et al. (2022) were able to find garden path effects, but neither explicitly compares these results with human performance.

4 Question Answering

4.1 Materials

As mentioned previously, human comprehension of garden path sentences is often assessed by presenting garden path sentences and asking comprehension questions. Because BERT-style models achieve high performance in question-answering tasks (Devlin et al., 2019; Liu et al., 2019b), we are able to assess their performance in the same manner as humans'. We used the same materials presented in Christianson et al. (2017). There were 40 sets of items, each containing a garden path and matched local coherence structure, as well as two additional sentences which were identical to the garden path and local coherence sentences except for the addition of "who was" to disambiguate relative clauses.

We deviate from Christianson et al. (2017) in the kinds of questions that are paired with the context sentences. Christianson et al. (2017) used simple yes-no questions, but BERT would not perform well on this task since it is trained to identify portions of the provided context as answers instead. Therefore, we constructed a variety of new questions that could be answered by a span of the context in order to assess BERT's ability to resolve the garden path structure.

Each garden path and local coherence structure has three pieces which are relevant to semantic role assignment: the Matrix agent (asking the identity of the entity which performs the action in the main clause), the Matrix patient (asking the identity of the entity which receives the action performed by the agent), and the Modified Argument (asking the identity of the entity in the matrix clause which is modified by the relative clause). A set of example questions and answers can be seen in Figure 1.

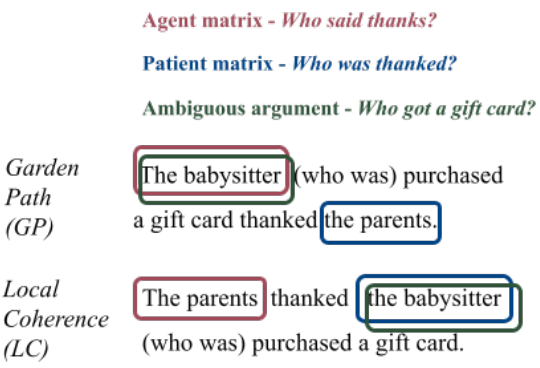

Figure 1: Example questions and answers targeting correct entity identification.

In the garden path structures, the modified argument is the one which causes the possible ambiguity. Initially the relative clause is parsed as the main clause, but this parse must be reanalyzed after encountering the second verb; therefore the model may have more difficulty with answering questions about structures of this type. In the local coherence structure, there is no ambiguity because the main verb is encountered first, so the second verb

must be correctly parsed as an embedded clause. Thus the model should have higher performance for local coherence structures than the corresponding question for garden paths.

In addition to asking questions about which entities have relationships with particular verbs, we can also ask the reverse: what did particular entities do? This leads us to a second set of questions: Matrix Action (asking the action done by the entity in the main clause) and Embedded Action (asking what happened to the entity modified by the relative clause). A set of example questions and answers can be seen in Figure 2. For these types of questions, the key contrast is on the Matrix Action questions. For the garden path structures, the model may answer with the embedded action rather than the correct parse, but in local coherence structures there is no intervening relative clause so accuracy may be higher.

All together, we had five questions for each of the 160 sentences (contexts) for a total of 800 items. Each context-question pair was provided to a publically available pre-trained transformer model ($\text{BERT}_{\text{BASE}}$-uncased, $\text{BERT}_{\text{LARGE}}$-uncased, $\text{RoBERTa}_{\text{BASE}}$-uncased, $\text{RoBERTa}_{\text{LARGE}}$-uncased) which had been fine-tuned for question-answering using the SQuAD2.0 dataset. In addition, we used the Hugging Face library (Wolf et al., 2020) for implementation of the question-answer task.

To compare BERT's performance with humans', we also conducted an online comprehension task in which subjects were presented with a context sentence and asked to answer one of the comprehension sentences by typing a response. (The context sentence was not on screen as participants answered the question.) There were 74 fluent English participants and each had a 32-item subset of the questions and contexts plus 13 filler items. This resulted in at least 80 responses for each of the question types in the dataset. All of these materials (including garden path and local coherence contexts, questions, anonymized human responses, and BERT responses) are available publicly for future use.

4.2 Results

Overall, the transformer models perform similarly to humans based on their average accuracy over all types of questions, structures, and ambiguities. As seen in Table 1, the BERT models' accuracy ranges from 2.5-100% while human accuracy ranges from

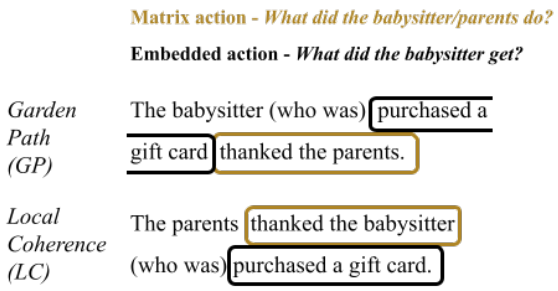

Figure 2: Example questions and answers targeting correct action identification.

32.3-95.7%, suggesting that the models did not perform universally better or worse than humans on this task.

Rather, the performance differences between the two emerge in specific question types. For Agent Matrix and Matrix Action question types, at least 50% of the transformer models' scores were lower than the corresponding human performance. Both BERT and humans struggle the most with Matrix Action questions, which require a semantic connection to be made between elements which are not collocated linearly. Agent Matrix questions also require the ability to make this connection, but humans achieve very high accuracies on this type, in contrast to BERT.

These results suggest that humans and BERT are making similar mistakes regarding semantic connections between arguments. For BERT, the failure is bidirectional–it cannot retrieve the agent when asked about an action or vice versa. For humans the failure is only unidirectional–they can correctly answer who performed an action, but are unable to identify an action when presented with an agent.

As expected, we found that local coherence structures were easier to process than garden path structures for both humans and BERT. Across all question types and ambiguities, humans have an average performance increase of 14.9%. For the transformer models the accuracy increases were between 8.8% ($\text{BERT}_{\text{LARGE}}$) and 20.5% ($\text{RoBERTa}_{\text{LARGE}}$) with an average of 15.5%. The comparable increase in performance between language models and humans also suggests that the two face similar difficulties in garden path processing, despite their differing processing mechanisms.

We also observe an increase in both human and model performance for sentences disambiguated

Question	Struct.	Amb.	BERT_{B}	BERT_{LG}	$\text{RoBERTa}_{\text{B}}$	$\text{RoBERTa}_{\text{LG}}$	Human
Agent Matrix	Garden Path	Amb.	0.65	0.675	0.55	0.575	0.907
		Unamb.	0.625	0.725	0.625	0.775	0.949
	Local Coherence	Amb.	1	0.976	0.976	0.951	0.902
		Unamb.	0.949	0.923	0.872	0.872	0.957
Patient Matrix	Garden Path	Amb.	0.925	0.975	1	1	0.946
		Unamb.	0.85	0.925	0.85	0.975	0.830
	Local Coherence	Amb.	0.683	0.805	0.805	0.902	0.93
		Unamb.	0.974	1	1	0.974	0.954
Ambiguous Argument	Garden Path	Amb.	0.825	1	0.475	0.75	0.710
		Unamb.	0.9	0.975	0.975	0.975	0.938
	Local Coherence	Amb.	0.375	0.825	0.85	0.975	0.685
		Unamb.	0.975	0.975	1	1	0.892
Matrix Action	Garden Path	Amb.	0.3	0.025	0.325	0.175	0.323
		Unamb.	0.8	0.875	0.875	0.975	0.842
	Local Coherence	Amb.	0.8	0.525	0.7	0.7	0.782
		Unamb.	0.975	0.75	0.975	0.975	0.913
Embedded Action	Garden Path	Amb.	0.725	0.825	0.675	0.475	0.639
		Unamb.	0.625	0.8	0.65	0.575	0.860
	Local Coherence	Amb.	0.825	0.95	0.875	0.975	0.777
		Unamb.	0.925	0.95	0.975	0.975	0.799
Average			**0.785**	**0.824**	**0.801**	**0.828**	**0.821**

Table 1: Results comparing the performance of humans to a variety of BERT-style transformer models in a question-answering task where question contexts are garden path sentences.

using "who was" to introduce relative clauses. Human performance increased by 13.3%, while model performance increase ranged from 13.2% ($\text{BERT}_{\text{LARGE}}$) to 15.9% ($\text{RoBERTa}_{\text{LARGE}}$) with an average of 14.9%. Again the comparable increase is suggestive of similar processing mechanisms.

Finally, we find that increasing the size of the model and changing the training objectives result in only a marginal performance increase. In model evaluations using SQuAD 2.0, $\text{BERT}_{\text{LARGE}}$ improves on $\text{BERT}_{\text{BASE}}$ by 8.5% and $\text{RoBERTa}_{\text{LARGE}}$ improves on $\text{RoBERTa}_{\text{BASE}}$ by 5.3%. This contrasts with our results of a 3.9% and 2.7% increase respectively. Additionally, RoBERTa models' performance increase over BERT models of the same size is also reduced compared to the SQuAD 2.0 evaluations (1.6% vs. 7.6% for BASE models, 0.4% vs 4.4% for LARGE models).

While the transformer models showed similar results to humans in terms of accuracy, qualitatively the performance differs between the two in terms of incorrect responses. For instance, in humans an incorrect response would likely be an incorrectly identified entity or action (depending on the question type). However, the transformer models seemed to frequently answer questions simply by repeating the sentence or a non-constituent subpart, refraining entirely from selecting a single entity or action from the sentence. This is not an error that was seen in the human data. In addition to demonstrating a lack of awareness about particular semantic relationships in the sentence, this also suggests a lack of understanding of what a felicitous question response entails.

4.3 Discussion

The fact that BERT's performance is comparable (rather than superior) to humans is somewhat surprising given the vastly different way the two process language. For humans, the difficulty in processing garden path structures arises from the fact that language is presented sequentially: when encountering the first verb, people are unaware that there will be a second verb later in the sentence and thus are likely to parse the sentence incorrectly initially. BERT, on the other hand, receives all input simultaneously, so it should face less difficulty in parsing the sentence and assigning correct semantic roles. As seen in the results however, BERT seems to struggle with forming the correct relationship

between agents and matrix verbs when there is an intervening relative clause, at rates at least as high as humans'.

Additionally, it is surprising that increasing the size of the model or changing the training procedures did not cause a corresponding performance increase in the models. This suggests that in order to understand the complex syntactic structures present in garden path sentences, one must do more than increase model parameters and the amount of training data. Rather, changing the architecture of the model itself may result in larger performance gains.[3]

5 Probe Task

Since our models perform similarly to humans on comprehension questions based on garden path sentences, we aimed to investigate precisely which semantic role each word in a garden path sentence is assigned, since human processing seems to be hampered by the misassignment of these semantic roles. In order to investigate the semantic roles assigned to the different entities in garden path sentences, we designed a probe (Alain and Bengio, 2016) trained on BERT's hidden states to better understand it's representation of the roles in question.

The linear classifier was trained on each model's embedding of a single token taken from sentences in which those words fall under the span of a semantic role of interest.

5.1 Training Materials

To create a training set for semantic roles, we used annotations from PropBank (Palmer et al., 2005). Each verb annotated by PropBank has a corresponding frame file in which verb-specific semantic relations are detailed. We chose to focus on two semantic roles. The first is PropBank's [PAG] tag, which represents a proto-agent type role across verbs. Selecting the other role was less straighforward, as the thematic roles in PropBank were not uniform across the areas of interest in the garden path sentences. For example, the two following stimuli sentences have differing thematic roles assigned to the first entity of the sentence (relevant verb in bold):

1. (The child)$_{\text{GOAL}}$ **bought** an ice cream cone smiled at the cashier.

2. (The child)$_{\text{DIRECTION}}$ **read** the story hugged the nanny.

In order to test the classifier on the largest number of stimuli possible, we chose to focus our probe on the entities tagged [GOL] (Goal role), as it has the most occurrences in our stimuli (143 instances total).

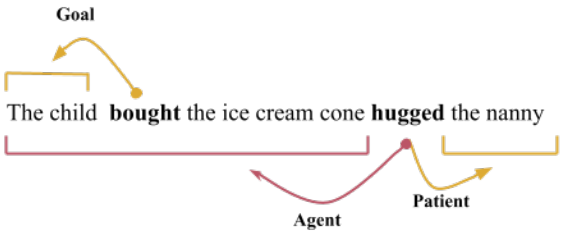

Figure 3: Here, the phrase *the child* acts as the Goal of *bought*, while the phrase *the child bought the ice cream cone* acts as the agent of *hugged*. The Patient label is presented for convenience, but not analyzed in this investigation.

We train a binary linear classifier for both Agent and Goal, on each layer of each model, in the form of a logistic regression classifier. This is due to the fact that some words may be constituents of arguments of multiple verbs, and therefore be assigned different roles, as displayed in Figure 3. This approach allows for detecting multiple different roles on each word.

5.2 Probe Design

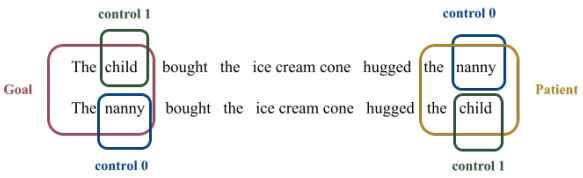

Figure 4: Words are labeled for goal and patient semantic roles (the latter of which is provided for convenience but is not addressed in this study), in addition to the control tag, which is randomly assigned to each unique word in the training data, i.e. 0 for *nanny* and 1 for *child*.

One concern when using classifiers to investigate the hidden states of a model is the possibility of the model achieving high accuracy by memorizing specific words' typical labels, rather than learning the patterns associated with the labels themselves. In

[3]Because relative clauses are the syntactic phenomena which make garden path processing difficult, finding a model which is able to correctly parse these may lead to advances in garden path understanding. While work on this is limited in English, work in other languages suggests some promising models are LSTM, PERT$_{\text{LARGE}}$, and GPT-3 (Song et al., 2022).

order to assess this level of memorization, we implemented a control task (Hewitt and Liang, 2019). This took the form of two additional classifiers which were trained on the sets of embeddings used by the Agent and Goal classifiers, but a binary label (rather than a semantic label) was randomly assigned to each unique word in the training set, as seen in Figure 4.

This allows us to assess the extent of memorization through the metric of selectivity:

$$selectivity = linguistic_{\text{ACC}} - control_{\text{ACC}}$$

The Agent semantic role is much more common in PropBank than the Goal role. In order to keep our classifiers equivalent in the number of samples they received, we created artificially split 50/50 +Role/-Role training sets for both Agent and Goal classifiers. All stimuli sets underwent a 80/20 train/test split. In order to maximize both performance and specificity in our probe of garden path sentences, we choose a regularization constant of 0.01 for our classifier, following the findings of Hewitt and Liang (2019).

To test our probe classifiers on the garden path sentences, we selected the layer from each model in which the associated classifier had the highest performance on Goal classification (which had overall lower accuracy than Agent classifiers, as shown in Table 2) for the PropBank sentences. We then applied the probe classifiers for both Agent and Goal roles to the first and second relevant entities in the same sentences that were analyzed in the question answering portion of this experiment. The sentences were sub-selected for those containing the Goal and Agent roles, leaving us with 24 sentences in each condition (garden path vs. local coherence structure, ambiguous vs. unambiguous).

5.3 Results

Overall, the linear classifier accuracies show that Agent and Goal semantic roles are decodable from token representations when trained on PropBank sentences. However, this information is not available to the same extent in each of the models. While the BERT models perform Agent classification and Goal classification accurately 80% and 70% of the time respectively, RoBERTa models' accuracy does not exceed 65% for both Agent and Goal classification as shown in Figures 5 and 6.[4]

[4]Complete results for each layer of BERT and RoBERTa are available in Appendix A.

Additionally, the information change throughout the layers is also inconsistent between models with BERT models' accuracy peaking in the later layers of the model and RoBERTa classifier accuracies staying relatively level as the layers progress.

Due to the different inconsistencies in classification accuracies for the PropBank sentences across layers, we chose to analyze the garden path sentences using only the highest performing layer from each of the models. The layers chosen as well as their accuracy and selectivity metrics can be seen in Table 2.

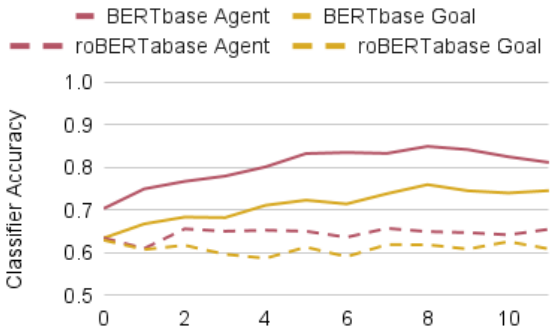

Figure 5: The average classifier accuracies from our Agent and Goal classifiers on our test set from PropBank, by model, across layers for $\text{BERT}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{BASE}}$ models.

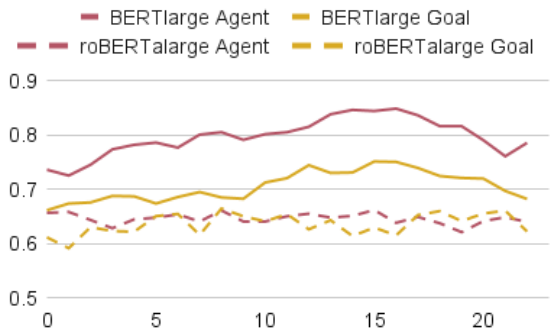

Figure 6: The average classifier accuracies from our Agent and Goal classifiers on our test set from PropBank, by model, across layers for $\text{BERT}_{\text{LARGE}}$ and $\text{RoBERTa}_{\text{LARGE}}$ models.

To analyze the garden path and local coherence sentences, both the Agent and Goal classifiers were applied to the noun associated with the first and second entities (e.g., *child* and *nanny* from Figure 3). The probabilities of each classifier assigning a given role to an entity is shown in Figure 7.

In garden path sentences, the $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ probes showed a low probability of

Model	Layer	Agent Accuracy	Agent Selectivity	Goal Accuracy	Goal Selectivity
$BERT_B$	9	84.92	34.56	75.96	28.04
$BERT_{LG}$	17	84.44	32.72	75.12	24.32
$RoBERTa_B$	11	64.2	13.56	62.6	15.88
$RoBERTa_{LG}$	10	66.08	15.36	66.44	14.8

Table 2: The accuracy and selectivity on the layers associated with the highest-performing classifier for each model, tested on the tokens from PropBank. Complete results for all layers are included in Tables 3 and 4 in Appendix A.

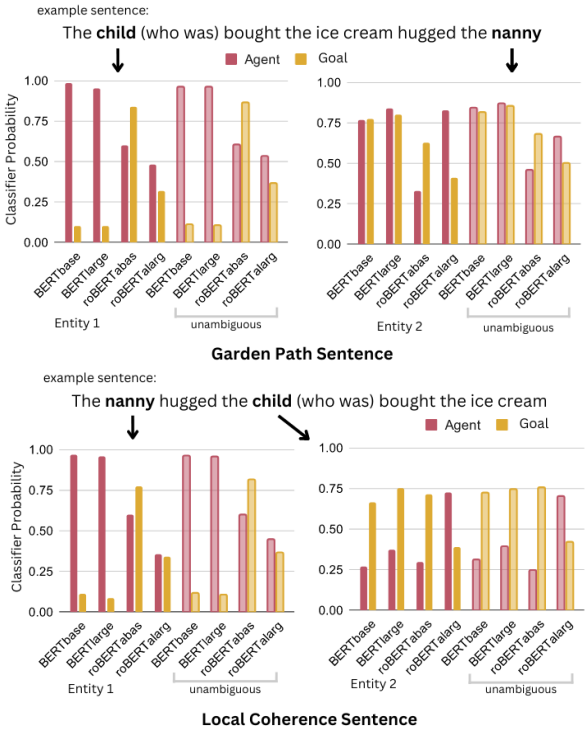

Figure 7: The average classifier probabilities assigned by our Agent and Goal classifiers to the first entity (left) and the second entity (right) in garden path and local coherence structure sentences in the selected layers. Exact numerical results are available in Appendix A.

the Goal role on the first entity in the sentence. $RoBERTa_{BASE}$, however, successfully showed high classifier probability for the Goal role in both the ambiguous and unambiguous sentence types, which could potentially suggest a better representation of the semantics of garden path sentences. However, this pattern also held for local coherence structures though (where the Goal role should not be assigned to the first entity), suggesting that $RoBERTa_{BASE}$ (nor any other model) were perfectly able to make semantic distinctions based on the differing syntactic structures. Finally, we also observe that the addition of disambiguating information into the sentence does not seem to have a large impact on classifier performance in general, for both different sentence types and models.

5.4 Discussion

Overall, the Agent classifiers were highly successful in labelling the first entity for both sentence structures (with the surprising exception of the $RoBERTa_{LARGE}$ classifier). In the case of the garden path sentences, the assignment of Goal to the first noun (i.e., *the child* in Figure 4), however, was extremely low, possibly indicating that most of the BERT models are not strongly representing *the child* as the goal of the verb *hugged*. This kind of semantic role interference is also what is hypothesized to impede human processing.

As a contrast, the second entity in the local coherence structures have low Agent and high Goal probabilities. This suggests that the models are better able to represent the Agent and Goal semantic roles in a local coherence structure, again mirroring the performance of humans and the model performance seen in the question answering task.

In contrast to the structure manipulation, ambiguity differences did not significantly change the classifiers' predictions, despite the question answering performance showing a larger gain in the unambiguous over ambiguous sentence contexts. Given that syntactic knowledge has been argued to be present in the model weights (Vilares et al., 2020; Hewitt and Liang, 2019), it is interesting that disambiguation (a lexical manipulation that clarifies syntax), does not improve performance in the probe task. This suggests that future syntactic investigations should focus not only on token representations, but also on other components of the models which may contribute to downstream task performance.

In terms of individual model performance, we find that no model perfectly aligns with human performance. In the local coherence structures, $BERT_{BASE}$ and $BERT_{LARGE}$ were most successful at assigning the correct roles to the correct entities, however they both misassigned the roles of the first entity in garden path structures. Interestingly, the semantic role predictions of the two were very similar, suggesting that increasing the model

size does not fundamentally change what semantic information is held within token representations.

RoBERTa models, on the other hand, did not perfectly assign roles in either structure condition. RoBERTa$_{\text{BASE}}$ seems to prefer assigning Goal roles to every entity, while RoBERTa$_{\text{LARGE}}$ does the same with Agent roles. In this case, it does seem that increasing the model size does lead to the acquisition of different semantic knowledge, in addition to the differences from BERT models already mentioned.

Another area where these differences are observed is the extent to which classifier accuracy increased in later levels of BERT versus RoBERTA. In BERT, we observed an increase in performance in later layers, while for RoBERTa the performance stagnated across all layers. This was not due to a difference in classifier training, as all classifiers had identical hyperparameters and training corpora. Rather, we can conclude this is due to the RoBERTa model itself–its differing training procedures must lead to a fundamental difference in how semantic roles are processed from BERT (i.e., they are less strongly represented in the token weights themselves), given that the two have very similar performance on the downstream question answering task.

Because no model was able to assign all semantic roles perfectly in every condition (based on the probe task), we are left to conclude that the semantic knowledge within token representations of all models is imperfect and not based on deep syntactic knowledge. Rather, in some cases the models seem to be relying on heuristics to assign semantic roles. Such heuristics might include word order in the sentence, frequency of a particular semantic role, and linear proximity. Future investigations should aim to discover which heuristics are most relevant to langauge models' representations and performance, as well as how token representations may interact with other model components in order to achieve performance similar to humans on downstream tasks.

6 Conclusion

Overall, BERT-style transformer models do not perform significantly better than humans on garden path sentences in question answering. This suggests that, despite the temporal amodality of BERT's language processing, it still faces the same issues of misinterpretation that human speakers do in online sentence processing. Additionally, probe results suggest that BERT fails to assign the correct semantic roles to the entities in garden path sentences, despite showing successful assignments on other corpus sentences. This error is similar to human-style garden path misinterpretations, despite the many differences between human and model language processing (i.e. temporality, working memory demands).

Additionally, we observe differences in semantic role representations between the models tested—BERT models seem to make similar role predictions regardless of model size, while RoBERTa$_{\text{BASE}}$ makes different predictions than RoBERTa$_{\text{LARGE}}$. Furthermore, both RoBERTa models seem to have different semantic representations than the original BERT models, suggesting that particular training procedures and tasks can lead to widely different internal model states yet still show negligble impact on performance of downstream tasks.

Limitations

The garden path structures presented here are a phenomenon predominantly found in English. Structural ambiguities found in other languages vary widely, and so our ability to generalize about BERT's ability to process these ambiguities cross-linguistically is limited. Additionally, these types of garden path structures are relatively scarce in natural language, and it is possible performance would be higher if BERT were fine-tuned using these structures specifically (we choose not to do this in order to approximate the levels of experience humans have with these structures to obtain more natural comparison). Finally, the number of sentences tested in the probe task is low, due to variation of stimuli thematic roles.

Ethics Statement

The authors have no ethical concerns relating to the research presented in this paper.

Acknowledgements

The first two authors (TI and KW) contributed equally to this work. The research was supported by the NYUAD Research Institute under Grant G1001. We additionally thank the anonymous reviewers for their guidance and feedback on earlier versions of this paper.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Kiel Christianson, Andrew Hollingworth, John F Halliwell, and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4):368–407.

Kiel Christianson, Steven G Luke, Erika K Hussey, and Kacey L Wochna. 2017. Why reread? Evidence from garden-path and local coherence structures. *The Quarterly Journal of Experimental Psychology*, 70(7):1380–1405.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

William Jurayj, William Rudman, and Carsten Eickhoff. 2022. Garden-path traversal within GPT-2. *arXiv preprint arXiv:2205.12302*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Timothy J Slattery, Patrick Sturt, Kiel Christianson, Masaya Yoshida, and Fernanda Ferreira. 2013. Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, 69(2):104–120.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. Sling: Sino linguistic evaluation of large language models.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Marten Van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *CogSci*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

David Vilares, Michalina Strzyz, Anders Søgaard, and Carlos Gómez-Rodríguez. 2020. Parsing as pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9114–9121.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix A

Model	Layer	Agent Acc.	Agent Selec.	Goal Acc.	Goal Selec.
bert-base-uncased	1	70.36	21.88	63.44	11.8
	2	74.96	25.8	66.76	18.72
	3	76.72	27.16	68.36	19.08
	4	77.96	29.4	68.24	20.2
	5	80.12	33.32	71.12	20.56
	6	83.24	34.68	72.32	22
	7	83.48	34.72	71.44	22.56
	8	83.28	34.08	73.84	24.28
	9	84.92	34.56	75.96	28.04
	10	84.16	36.32	74.52	26.12
	11	82.48	34.44	74	24.76
	12	81.16	32.68	74.56	26
	Avg.	**80.24**	**31.59**	**71.21**	**22.01**
bert-large-uncased	1	67.92	15.44	62.84	14.32
	2	73.6	24.84	66.12	18.6
	3	72.52	22.52	67.36	18.6
	4	74.52	25.16	67.56	19.8
	5	77.36	27.24	68.76	20.6
	6	78.2	28.4	68.68	20.24
	7	78.6	31.28	67.36	18.48
	8	77.68	28.56	68.56	19
	9	80.08	31.16	69.48	23.44
	10	80.52	30.2	68.48	20.76
	11	79.12	29.6	68.24	19.64
	12	80.16	31.12	71.24	21.44
	13	80.52	30.72	72.04	22.44
	14	81.52	32	74.44	24.28
	15	83.84	33.96	73	23.36
	16	84.64	34.68	73.08	21.44
	17	84.44	32.72	75.12	24.64
	18	84.88	34.48	75.04	24.32
	19	83.64	35.28	73.92	23.92
	20	81.64	30.92	72.44	19.72
	21	81.64	34.36	72.08	18.64
	22	79.04	29.56	71.96	20.24
	23	76.08	27.28	69.68	17.96
	24	78.56	27.52	68.2	14.48
	Avg.	**79.20**	**29.54**	**70.24**	**20.43**

Table 3: Classifier accuracy and selectivity for each layer of $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$

Model	Layer	Agent Acc.	Agent Selec.	Goal Acc.	Goal Selec.
roberta-base-uncased	1	63.48	12.36	62.96	13
	2	61	16.4	60.76	12.12
	3	65.6	17.24	61.76	12
	4	65.04	15.6	59.64	7.56
	5	65.28	17.64	58.72	6.08
	6	65.04	18.68	61.32	10.6
	7	63.64	17.4	59.08	6.32
	8	65.72	14.36	61.92	13.52
	9	64.96	18	61.88	13.48
	10	64.72	13.64	60.88	11.76
	11	64.2	13.56	62.6	15.88
	12	65.52	16.56	60.92	6.76
	Avg.	**64.51**	**15.95**	**61.04**	**10.76**
roberta-large-uncased	1	62.68	14.64	60.24	9.16
	2	65.64	13.36	61.12	10.96
	3	65.8	14.84	59.12	6.92
	4	64.32	14.28	63	14.4
	5	62.8	12.8	62.28	14.92
	6	64.44	10.48	62.16	9.48
	7	64.76	13.52	65.04	14.8
	8	65.36	15.16	65.44	17.68
	9	64.04	11.72	61.56	9.28
	10	66.08	15.36	66.44	14.8
	11	64.04	11.72	65	14.92
	12	64.04	16.68	64.08	12.76
	13	65	15.92	65.4	17
	14	65.48	16.12	62.6	16.6
	15	64.76	16.16	64.28	13.36
	16	65.08	11.16	61.36	16.28
	17	66.16	14.92	62.92	10.8
	18	63.8	16.56	61.48	10.52
	19	64.88	13.44	65.24	14.68
	20	63.72	10.96	65.96	13.76
	21	62.08	10.08	64.12	14.12
	22	64.12	12.88	65.52	17.48
	23	64.84	13.56	66.04	13.8
	24	63.96	12.28	62.2	9.8
	Avg.	**64.49**	**13.69**	**63.44**	**13.26**

Table 4: Classifier accuracy and selectivity for each layer of $\text{RoBERTa}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{LARGE}}$

Model	Layer	Sentence Type	Amb.	Entity 1 Agent Probability	Entity 1 Goal Probability	Entity 2 Agent Probability	Entity 2 Goal Probability
bert-base uncased	9	Garden Path	Amb.	0.983	0.0989	0.768	0.773
			Unamb.	0.964	0.110	0.846	0.819
		LC-A	Amb.	0.968	0.114	0.266	0.667
		LC-U	Unamb.	0.962	0.119	0.314	0.723
bert-large uncased	17	Garden Path	Amb.	0.954	0.102	0.836	0.798
			Unamb.	0.962	0.108	0.871	0.855
		Local Coherence	Amb.	0.959	0.087	0.370	0.751
			Unamb.	0.959	0.105	0.393	0.747
roberta-base uncased	11	Garden Path	Amb.	0.600	0.838	0.326	0.624
			Unamb.	0.606	0.864	0.457	0.680
		Local Coherence	Amb.	0.601	0.774	0.297	0.715
			Unamb.	0.598	0.815	0.250	0.757
roberta-large uncased	10	Garden Path	Amb.	0.479	0.320	0.827	0.411
			Unamb.	0.536	0.368	0.667	0.502
		Local Coherence	Amb.	0.356	0.337	0.727	0.389
			Unamb.	0.449	0.367	0.700	0.423

Table 5: Detailed probe classifier results on first and second entities in garden path and local coherence test sentences for highest performing layer in each model investigated.