Korean Journal of English Language and Linguistics, Vol 23, June 2023, pp. 461-481
DOI: 10.15738/kjell.23..202306.461

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Decoding BERT's Internal Processing of Garden-Path Structures through Attention Maps [*]

Jonghyun Lee (Seoul National University) · **Jeong-Ah Shin** (Dongguk University)

Received: April 12, 2023
Revised: May 17, 2023
Accepted: June 7, 2023

Jonghyun Lee (first author)
Senior Researcher,
Institute of Humanities,
Seoul National University
Email: museeq@snu.ac.kr

Jeong-Ah Shin
(corresponding author)
Professor, Division of
English Language and Literature,
Dongguk University
Email: jashin@dongguk.edu

[*]This work was supported by the Park Chung-Jip Scholarship Fund for the Next Generation in English literature and language at Seoul National University in 2022.

ABSTRACT

Lee, Jonghyun and Jeong-Ah Shin. 2023. Decoding BERT's internal processing of garden-path structures through attention maps. *Korean Journal of English Language and Linguistics* 23, 461-481.

Recent advancements in deep learning neural models, such as BERT, have demonstrated remarkable performance in natural language processing tasks, yet understanding their internal processing remains a challenge. This study employs the method of examining attention maps to uncover the internal processing of BERT, specifically when dealing with garden-path sentences. The analysis focuses on BERT's utilization of linguistic cues, such as transitivity, plausibility, and the presence of a comma, and evaluates its capacity for reanalyzing misinterpretations. The results revealed that BERT exhibits human-like syntactic processing by attending to the presence of a comma, showing sensitivity to transitivity, and reanalyzing misinterpretations, despite initially lacking sensitivity to plausibility. By concentrating on attention maps, the present study provides valuable insights into the inner workings of BERT and contributes to a deeper understanding of how advanced neural language models acquire and process complex linguistic structures.

KEYWORDS

attention map, Natural Language Processing, Psycholinguistics, Transformers, garden-path structure

1. Introduction

Deep neural language models, specifically those adopting the Transformers architecture (Vaswani et al., 2017), are at the forefront of recent advancements in natural language processing (NLP). This cutting-edge technology, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) and GPT-3 (Generative Pre-trained Transformer 3) (Brown et al. 2020) has demonstrated superior performance compared to other models and even humans in several natural language processing tasks (Koroteev 2021). They excel in many language comprehension benchmarks, seemingly acquiring linguistic competence including human-like syntactic processing during pre-training (Goldberg 2019). Despite these achievements, it is still unclear how these models acquire such skills and what specific linguistic traits are learned during the pre-training process (Rogers et al. 2020).

Recent research has attempted to explore these issues with a variety of approaches, one of which is a targeted evaluation approach (Marvin and Linzen 2018). This method involves assessing the abilities of language models (LM) by analyzing their processing of carefully-constructed sentences, similar to psycholinguistic experiments. Through this method, several studies have demonstrated that Transformer-based models exhibit syntactic processing capabilities comparable to those of humans and have learned various linguistic characteristics. For instance, Hu et al. (2020) found that GPT-2 exhibited human-like performance in several syntactic tasks, such as agreement, licensing, and long-distance dependencies. In addition, Lee et al. (2022) showed that BERT and ALBERT (A Lite BERT) (Lan et al. 2019) had similar patterns of incremental processing with those of human sentence processing and are sensitive to linguistic cues such as transitivity, plausibility and morphology.

The targeted evaluation approach employed to investigate the linguistic capabilities of LMs, however, has certain limitations. This approach involves analyzing the output of LMs on carefully-designed input sentences and uses accuracy or surprisal as the dependent variable, which is similar to the accuracy rate and reading times utilized in psycholinguistic experiments to measure human language processing. However, those are indirect measurements employed only due to the inability to access the internal processes of human language processing. As the targeted evaluation approach emulates psycholinguistic experiments, it also utilizes indirect measurement methods. While these studies have yielded important results regarding the syntactic performance of the models, they may not fully represent the internal processing of the models, which may contain more informative insights about the LMs.

The present study seeks to address the limitations of the targeted evaluation approach by combining it with an analysis of attention maps from a pre-trained model. Attention is a critical component in Transformer architectures, as they rely on an attention mechanism to establish global dependencies between input and output (Vaswani et al. 2017). Attention can be interpreted through attention weights, which indicate the level of weight or attention given to a particular word when the current word computes the subsequent representation (Clark et al. 2019). Clark et al. (2019) demonstrated that by examining attention maps of BERT through attention weights, the attention heads of BERT exhibit specific patterns in processing sentences, and certain attention heads correspond to particular syntactic characteristics. For example, the 10th head of the 8th layer is specialized in the dobj (direct object to verb) dependency, where the direct objects attend to the verbs with 86.8% accuracy. In this regard, this study tests the syntactic capacities and the internal processing of LMs such as BERT through carefully-designed sentences, but instead of using accuracy or surprisals as dependent variables, it examines attention maps.

1.1 Garden-path effect and surprisals

The current study will utilize carefully-designed sentences that have been used in psycholinguistic experiments, in accordance with the targeted evaluation approach. These sentences are specifically constructed to compel language comprehenders or models to use particular linguistic or syntactic features when processing the sentences. For instance, sentences such as "The farmer that the parents love swim." (Marvin and Linzen 2018) are useful in evaluating whether and how LMs are capable of handling problems related to subject-verb agreement. Examining how LMs process a range of sentence structures allows researchers to determine the extent to which these models have acquired diverse linguistic characteristics.

In this study, garden-path structure sentences are employed alongside other sentence structures, to assess whether an LM can detect temporary ungrammaticality and to which linguistic cues it is sensitive. Garden-path structure refers to a sentence that is grammatical as a whole but may initially appear ungrammatical during incremental parsing. For example, the sentence "When the dog scratched the vet took off the muzzle." is grammatical, but when processed word-by-word from the beginning of the sentence, it can be seen as an ungrammatical sentence with no subject for the main verb, "took off." This is because during online-processing, "the vet," the subject for "took off," can be misinterpreted as the object of "scratched." To understand the garden-path sentences correctly, a comprehender or model must first notice the temporary ungrammaticality, or misinterpretation, at or after the region of main verb, and then abandon the initial misinterpretation—"the vet" as an object of "scratched"—and reanalyze it—"the vet" as the subject of "took off." It is known that human comprehenders exhibit increased reading times at the main verb region when reading garden-path structure sentences, as compared to non garden-path sentences where the presence of comma explicitly signifies the clause boundary (Adams et al. 1998, Ferreira et al. 2001, Hopp 2015, Pickering and Traxler 1998), such as in "When the dog scratched, the vet took off the muzzle." The increased reading times have been regarded as indicating that the comprehenders may be successful in detecting the temporary ungrammaticality and reanalyzing the garden-path structure (Ferreira and Henderson 1991, Frazier and Rayner 1982).

These characteristics of the garden-path structure can serve as a valuable tool in investigating whether comprehenders rely on certain linguistic cues (Frazier and Rayner 1982) such as plausibility and transitivity when processing sentences. For example, in sentences such as (1a) and (1b), the verb "scratch" as a transitive verb requires an object, in this case, "the vet," while "struggle" does not.

(1) a. When the dog scratched the vet took off the muzzle. [Transitive, Garden-path]
b. When the dog struggled the vet took off the muzzle. [Intransitive, Garden-path]

For the garden-path effect, which is reflected in an increase in reading times, to occur, comprehenders must judge "the vet" as the verb's object. Thus, in (1b), where an intransitive verb is used, the garden-path effect should not appear because "the vet" cannot be interpreted as an object. This type of garden-path enables researchers to examine whether language comprehenders utilize transitivity when interpreting sentences, as indicated by the size of the garden-path effect. It is predicted that if language comprehenders are sensitive to transitivity, a greater reading times for the main verb will be observed in (1a) compared to (1b). Several psycholinguistic studies have demonstrated that humans take longer reading times for transitive conditions than for intransitive ones (Adams et al. 1998; Mitchell 1987; Staub 2007; Van Gompel and Pickering 2001).For LMs, surprisals function as an indicator of their ability to recognize ungrammatical structures within garden-path sentences and reanalyze them accordingly. The term 'surprisal', or the log inverse probability of a target word, draws an analogy with the processing in human comprehension, where the unexpected occurrence of a word, namely higher surprisal, tends to be correlated with greater reading times (Levy 2008, Smith and Levy 2013). The concept of surprisal enables a comparative framework between the two distinct systems. This is because surprisal, in both human comprehension and language models, is conceptualized as the cognitive or computational load experienced when integrating an unexpected input (Oh and Schuler 2023). While this does not measure loads in the exact same sense in both systems and thus the human reading time does not directly translate into processing time for LMs, it nonetheless provides a theoretical bridge, a common metric allowing comparison of the response to unexpected inputs in both humans and artificial neural networks. This analogy is substantiated by research showing that surprisals generated by LMs can predict human reading times (Goodkind and Bicknell 2018; Hao et al. 2020). Several studies have indeed harnessed surprisal as a measurement to scrutinize the predictive performance of LMs, thus creating a connection to human reading times. For instance, Futrell et al. (2018) adopted surprisal as an evaluative measure of LMs' prediction accuracy, drawing parallels to human reading times. Likewise, Lee et al. (2022) found that transformer-based models, including BERT and ALBERT, exhibited elevated surprisals, which they analogously related to increased human reading times, during the processing of garden-path sentences. Although these results implied that neural language models might exhibit human-like performance in syntactic processing, the studies mentioned above possess limitations owing to their reliance on surprisals as dependent measures. These results primarily reflect the outputs of LMs and offer limited insight into the internal mechanisms that yield these outcomes. Specifically, even if LMs demonstrate greater surprisal when processing garden-path structures, it might indicate a failure to accurately process the sentence rather than human-like syntactic processing. Essentially, LMs might fail to correctly understand the sentence, instead processing it with the subject missing. Unlike human processors, it is not feasible to ascertain whether LMs accurately comprehend a sentence, for example, through comprehension questions. Moreover, recent observations have noted that larger language models, despite their lower perplexity and higher parameter count, do not consistently yield surprisal estimates that are more predictive of human reading times (Oh et al. 2022, Oh and Schuler 2023). This counter-intuitive phenomenon has been specifically studied with variants of the GPT-2 language model, wherein it was found that the predictive power of surprisal estimates was actually less for these larger models (Oh and Schuler 2023). Such findings further highlight the complex relationship between surprisal and syntactic comprehension, reinforcing the need for more nuanced methods of evaluation.

Given these considerations, this study aims to examine the internal processing of garden-path sentences by an LM, using attention maps, which provide insight into the internal calculations of the LM. This approach aspires to provide a more comprehensive understanding of how the model manages the complexities of syntactic comprehension, and how closely its processes align with those of human readers.

1.2 Transformer and attention maps

This study aims to explore the internal mechanism of LMs, focusing on BERT (Devlin et al. 2019), a large transformer-based network (Vaswani et al. 2017). The most pivotal structure guaranteeing the performance of the Transformer architecture is the attention mechanism (Vaswani et al. 2017). BERT incorporates the multiple attention layers with multiple attention heads in its structure. Understanding how this attention mechanism works is important for fully appreciating the model (for more details, see Clark et al. 2019 and Devlin et al. 2019). The attention mechanism in BERT operates on a sequence of vectors that serve as input. These vectors, essentially numerical representations, capture the meanings of the tokens in a format that the model can process. It is worth noting that while for the initial attention layer, these input tokens are indeed the input sentence, for subsequent layers, the input becomes the comprehensive output from the preceding layer. Within each attention head, these input vectors undergo distinct linear transformations to generate corresponding query (Q), key (K), and value (V) vectors. These transformed vectors offer a means to explore various aspects of the information embedded within the input. Attention weights, denoted as Attention (Q, K, V), are then computed based on the transformed vectors. Each attention head calculates these weights by performing softmax-normalized dot products between the query and key vectors. This computation allows the model to assign scores to the relationships or 'attention' between every pair of tokens in the input sequence.

$$Attention\ \ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Following the computation of attention weights, each attention head produces its output. This is achieved by creating a weighted sum of the value vectors, where the weights are derived from the previously calculated attention scores. This sum represents a composite of the input tokens, with more emphasis placed on features that have higher attention scores. The outputs generated by each attention head within a layer are then aggregated, typically through a process of concatenation, resulting in a single, comprehensive output for that attention layer. This output, after undergoing additional transformations, serves as the input for the next attention layer, leading to further processing. This iterative mechanism allows the model to build a multi-layered, complex understanding of the input sentence (Clark et al. 2019, Devlin et al. 2019, Vaswani et al. 2017).

The process of examining attention maps involves scrutinizing the computed attention weights from each attention head across all attention layers, as opposed to only considering the final output after several processing steps summarized above. Given that the base model of BERT consists of 12 attention layers with 12 attention heads each, a total of 144 distinct attention weights can be derived. One thing to note here is that the term "attention" carries distinct connotations in the context of a Transformer model compared to its use in human psychology and cognition. In human cognition, attention often refers to the process of selectively concentrating on one aspect of information, whether sensory or conceptual, while ignoring other perceivable information (Posner and Petersen 1990). On the other hand, attention in the context of Transformer models denotes a computational mechanism for assigning different levels of relevance or importance to different parts of the input data. Attention mechanisms are used to weight the influence of different input values on the output. This is not akin to the human cognitive process of selective concentration, but rather a mathematical operation formulated to assign higher weights to more relevant or useful information and lower weights to less relevant or useful information. Thus, the attention weights indicate how much attention, or emphasis, a particular token (referred to as the query) gives to other tokens (keys) within its context. For instance, in the sentence "The cat chased the mouse," if the query word is "chased," the attention weights might assign greater weight to "cat" and "mouse." This suggests that these words have a greater influence on the understanding of the context and meaning of "chased" in this particular sentence, or that they share certain dependencies.

Interestingly, Clark et al. (2019) found that individual heads across each layer attend to distinct types of relationships between words. To simply put, in one attention head, the word "chased" might assign greater attention weights for "cat," while in another attention head, it might for "mouse." According to Clark et al. (2019), the lower layers tend to generate broader attention weights, meaning that the weights are spread across multiple words, while specific heads (2, 4, 7, and 8) in the earlier layers, predominantly focus their attention on the previous token. Additionally, their findings showed an increase in attention to certain syntactic relations within a particular head of a particular layer. For instance, direct objects attend to their verb (*dobj*) more than to other words in 10th head of 8th layer, noun modifiers such as determiners (*det*) attend to their nouns in 11th head of 8th layer, and subjects attend to their verbs (*nsujb*) in 2nd head of 8th layer. This finding illustrates the capability of each attention head in the model to learn and emphasize different types of word relationships.

The present study leverages these characteristics of attention weights to explore the processing of garden-path sentences in BERT. In this study, two syntactic relations that play crucial roles triggering and resolving the garden-path effects are *dobj* (the grammatical relation between a direct object and its verb) and *nsubj* (between a subject and its verb), these being among the syntactic dependencies outlined in the Stanford typed dependencies manual (De Marneffe and Manning 2008). Considering a sentence such as "When the dog scratched the vet took off the muzzle," the garden-path effect necessitates initially interpreting "the vet" as the object of "scratched," while the resolution of this garden-path requires the reanalysis of "the vet" as the subject of "took off." Therefore, by analyzing how much "the vet" attends "scratched" (dobj) or "took off" (nsubj) in BERT, how it handles the garden-path structure can be explored. As previously noted, the *dobj* relation can be captured in the 10th head of the 8th layer, while the *nsubj* relation can be identified in the 2nd head of the 8th layer. Hence, a high attention level from "the vet" to "scratched" in the 10th head of the 8th layer, indicating a significant focus on the *dobj* relation, represents the occurrence of a garden-path effect in the LM. In contrast, high attention from "the vet" to "took off" in the 2nd head of the 8th layer signifies that the LM is successfully interpreting "the vet" as the subject of "took off," forming the *nsubj* relation between two words.

1.3 Research Questions

Given this context, the current study is guided by the following research questions.

(1) Does BERT induce misinterpretation due to temporary ungrammaticality in processing garden-path sentences?
(2) To what extent does BERT employ linguistic cues such as transitivity and plausibility in syntactic processing?
(3) How does the syntactic processing of BERT align with or diverge from human syntactic processing in the presence of the garden-path effect and in the utilization of transitivity and plausibility?
(4) Can the examination of attention maps provide insights into BERT's internal processing mechanisms?

To explore these questions, a series of experiments with the LM was conducted as follows.

2. Method

2.1 Neural Language Models

The language model tested in this study was BERT, pre-trained from masked texts by jointly conditioning on both left and right context in all layers (Devlin et al. 2019). BERT performs two major tasks to acquire the meaning of words and sentences: "masked language modeling" and "next sentence prediction." In the "masked language modeling" task, the model predicts a certain word covered by [mask] in a sentence and learns the meaning of words and their relationships with other words by receiving feedback on the prediction results and correcting the weights. In the "next sentence prediction" task, the model takes the first sentence as input and predicts which sentence will follow, thus learning the relationship between sentences.

The version of BERT employed in this study was bert-base-uncased, which has 144 attention heads (12 layers and 12 head). The Huggingface implementation[1] was used as the pre-trained models for BERT in all experiments. Following Goldberg (2019) and Ettinger (2020), a [CLS] token was inserted at the beginning of each sentence to simulate the training conditions of the model and a [SEP] token was included after the end of each sentence to indicate the end of the sentence to the model. Attention maps were calculated with the codes adapted from the multiscale visualization tool of Vig (2019).

2.2 Materials

The materials consisted of a total of 48 garden-path sentences, distinguished by the incorporation of two distinct linguistic cues: transitivity and plausibility, with each represented in 24 instances. These sentences are characterized by an initial subject-object ambiguity, a structural feature that generates temporary processing difficulties. For instance, a sentence such as "As the woman edited the magazine amused all the reporters," initially allows a misinterpretation of "the magazine" as the object of "edited," before the introduction of "amused" necessitates a reinterpretation. The primary aim of incorporating two linguistic cues was to understand how BERT used those cues during the syntactic processing. Both cues influence on the syntactic processing of garden-path sentences, but they highlight distinct aspects. Transitivity relates to the structural aspects of sentence comprehension, reflecting the number of arguments a verb can accommodate. Plausibility, on the other hand, focuses on the semantic information or world knowledge, probing how these semantic aspects can influence syntactic processing. Through these two elements, this study sought to investigate the distinct yet complementary aspects that influence syntactic processing.

2.2.1 Transitivity

A set of 24 items with garden-path structure and transitivity manipulation was created such as Example (2) adapted from Futrell et al. (2019) and Staub (2007).

(2) a. When the dog scratched the vet (with his new assistant) took off the muzzle. [Transitive, Garden-path]
b. When the dog struggled the vet (with his new assistant) took off the muzzle. [Intransitive, Garden-path]
c. When the dog scratched, the vet (with his new assistant) took off the muzzle. [Transitive, Non garden-path]
d. When the dog struggled, the vet (with his new assistant) took off the muzzle. [Intransitive, Non garden-path]

Comprehenders might initially assume that the vet is the object of "scratched/struggled" in (2a) and (2b), resulting in increased reading times or surprisals when they encounter the main verb phrase "took off" compared

[1] https://huggingface.co/transformers/pretrained_models.html

to the Non garden-path sentences such as (2c) and (2d) where comma marks the end of the clause. However, while "scratched" in (2a) is a transitive verb that accepts an object, "struggle" in (2b) is an intransitive verb so that it is, in fact, not possible to interpret "the vet" as its object in (2b). Therefore, if BERT is sensitive to *Transitivity*, larger surprisals / greater reading times should be found in (2a), compared to (2b), which was supported by the previous studies (Adams et al. 1998; Lee et al. 2022; Mitchell 1987; Staub 2007; Van Gompel and Pickering 2001). In addition, a condition, *Length* was introduced into the material design to account for the distance between critical syntactic constituents. Specifically, a longer version of the test set was created by adding intervening words — "with his new assistant" in Example (2) — between the noun and the disambiguating verb. The inclusion of *Length* condition aimed to explore the effect of processing several intervening words before encountering the disambiguating word on BERT's syntactic representation. This strategy helped probe whether BERT recognized "the vet" as the subject of "took off" (nsubj) through sequential parsing or just by employing structural information.'

2.2.2. Plausibility

Twenty-four items were also created such as Example (3) modifying the sentences of Trueswell, Tanenhaus and Garnsey (1994).

(3) a. As the woman edited the magazine (about fishing) amused all the reporters. [Plausible, Garden-path]
b. As the woman sailed the magazine (about fishing) amused all the reporters. [Implausible, Garden-path]
c. As the woman edited, the magazine (about fishing) amused all the reporters. [Plausible, Non garden-path]
d. As the woman sailed, the magazine (about fishing) amused all the reporters. [Implausible, Non garden-path]

These sentences were examples of garden-path structures that can lead to subject-object ambiguities, similar to Example (2). Comprehenders may misunderstand "the magazine" as the object of "edited/sailed" when reading sentences such as (3a) and (3b). In terms of *Plausibility*, however, (3a) and (3b) were distinguished. The verb phrase, "edited the magazine," was plausible enough for processors to misinterpret "the magazine" as the object of the preceding verb, compared to "sailed the magazine," which was less plausible. A long version of the item was also created to examine the effect of some intervening words.

2.3. Procedure[2]

First, the input sentences were fed to the language models to collect the attention weights computed while processing garden-path sentences. Subsequently, attention maps required for the analysis were generated. To verify proper parsing of garden-path structures, attention maps for dobj and nsubj relationships were required, as described earlier. Finally, the attention weights at a specific head (8-10 for dobj, 8-2 for nsubj) were compared across the four conditions using the generated attention maps.

[2] The code for the test is available at: https://github.com/coolmintmild/bert_attention_map

2.4. Statistical analysis

The statistical analyses were performed using the R statistical programming environment (R Core Team 2023), with linear mixed-effects models (lme4: Baayen et al. 2008, lmerTest: Kuznetsova et al. 2017) employed to assess the statistical differences in attention weights within a designated head. The model assumed *Garden-path*, *Linguistic Cue* (*Transitivity* or *Plausibility*) and *Length* as fixed effects, and *Items* as random effects in order to minimize the influence of by-item variation, which resulted in 2×2×2 analysis (*Garden-path* vs. *Non garden-path* × *Transitive* (*Plausible*) vs. *Intransitive* (*Implausible*) × *Short* vs. *Long*). In the case where an interaction was observed between *Length* and *Garden-path* or *Linguistic cues*, further separate analyses were conducted for short and long sentence versions to provide a detailed examination of the impact of the added words in the sentences. For separate analyses for each *Length* group, the model assumes *Garden-path* and *Linguistic Cue* (*Transitivity* or *Plausibility*) as fixed effects, and Items as random effects. Data visualization was conducted using the seaborn (Waskom 2021) and matplotlib (Hunter 2007) packages in Python for attention heatmaps, and the ggplot2 package in R (Wickham 2016) for the other graphs.

2.5. Prediction

With this design, the predicted results are as follows:

For transitivity conditions,

(i) If BERT shows a garden-path effect or mistakenly interprets “the vet” as an object of “scratched/struggled,” it is anticipated that the attention weights for dobj relations (from "the vet" to "scratched/struggled") at the 10th head of the 8th layer in the Garden-path conditions (conditions a and b) will increase compared to the Non garden-path conditions (conditions c and d).

(ii) However, if BERT exhibits sensitivity towards transitivity or rejects an object for an intransitive verb, there will be no augmentation in the attention weights for dobj relations at the 10th head of the 8th layer under condition (b) — Intransitive, Garden-path conditions, or a lesser increase when compared to condition (a) — Transitive, Garden-path conditions.

(iii) If BERT successfully reinterprets the misperceived object “the vet” as the subject of the main verb “took off,” then it is predicted that the attention weights for nsubj relations (from “the vet” to “took off”) at the 2nd head of the 8th layer will remain relatively uniform across all four conditions.

For plausibility conditions:

(i) If BERT manifests a garden-path effect or incorrectly interprets “the magazine” as an object of “edited/sailed,” it is predicted that the attention weights for dobj relations (from “the magazine” to “edited/sailed”) at the 10th head of the 8th layer in the Garden-path condition (conditions a and b) will increase in relation to the Non garden-path conditions (conditions c and d).

(ii) However, if BERT demonstrates sensitivity to plausibility or discards “the magazine” as an object for “sailed” due to its implausibility, it is not expected that there will be an increase in the attention weights for dobj relations at the 10th head of the 8th layer in condition (b) — the Implausible, Garden-path conditions, or a more modest increase when compared with condition (a) — the Plausible, Garden-path conditions.

(iii) Lastly, if BERT successfully reinterprets the misinterpreted object “the magazine” as the subject of the main verb "amused," then it is anticipated that the attention weights for nsubj relations (from "the magazine" to "amused") at the 2nd head of the 8th layer will be consistent across all four conditions.

3. Results

Before presenting the results, here is how to read an attention map: Figure 1 provides an example of an attention map that displays all the attention weights from "the vet" to "took (off)". They are expressed by layer and head. The y-axis indicates the layer, and the x-axis indicates the head. Attention from "the vet"" to "took (off)" in the 10th head of the 8th layer (dobj) is indicated by the circled area.

3.1. Transitivity

3.1.1 dobj

The attention map in Figure 1 displays the average attention weights from "the vet" to "scratched/ struggled" across all conditions. The blue circled area marks the head and layer related with dobj (See also, Figure 2 for the bar graph of the mean attention weight for dobj). First, the statistical analysis revealed main effects for Garden-path (*estimate=.2180, SE=0.0177, t=12.29, p<.001*) as well as significant interactions between Garden-path and Transitivity (*estimate=.1770, SE=0.0355, t=4.99, p<.001*). This suggested that BERT was sensitive to the presence of a comma and to transitivity. First, the statistical analysis revealed main effects both for Garden-path (*estimate*=.2180, *SE*=0.0177, *t*=12.29, *p*<.001) and Transitivity (*estimate*=.0921, *SE*=0.0177, *t*=5.19, *p*<.001) along with significant interactions between Garden-path and Transitivity (*estimate*=.1770, *SE*=0.0355, *t*=4.99, *p*<.001). This implies that more attention was allocated to dobj relationship in the Garden-path conditions than in the Non Garden-path conditions, for both Transitive and Intransitive conditions, but less attention was devoted to the intransitive conditions. Furthermore, the analysis revealed a significant main effect of Length (*estimate=-.1029, SE=0.0177, t=-5.80, p<.001*), where the dobj relation was attended more in the long sentences than in short sentences. The difference between the short and long sentence versions was especially prominent within the garden-path conditions, as evidenced by the significant interaction between Garden-path and Length (*estimate=-.2060, SE=0.0355, t=-5.81, p<.001*), as well as the main effect of Length found in the separate post-hoc analysis including only the garden-path conditions (*estimate=-.2059, SE=0.0300, t=-6.87, p<.001*). In addition, there was a significant interaction between Transitivity and Length (*estimate=-.0737, SE=0.0355, t=-2.08, p<.05*) and a significant three-way interaction among Garden-path, Transitivity, and Length (*estimate=-.1473, SE=0.0710, t=-2.08, p<.05*). This suggested that, as shown in the Figure 3, while the long sentences produced greater attention weights both in transitive (a main effect of Length within the transitive conditions: *estimate=-.1398, SE=0.0276, t=-5.07, p<.001*) and intransitive conditions (a main effect of Length within the intransitive conditions: *estimate=-.0661, SE=0.0209, t=-3.16, p<.01*), the difference was even larger in the transitive conditions.

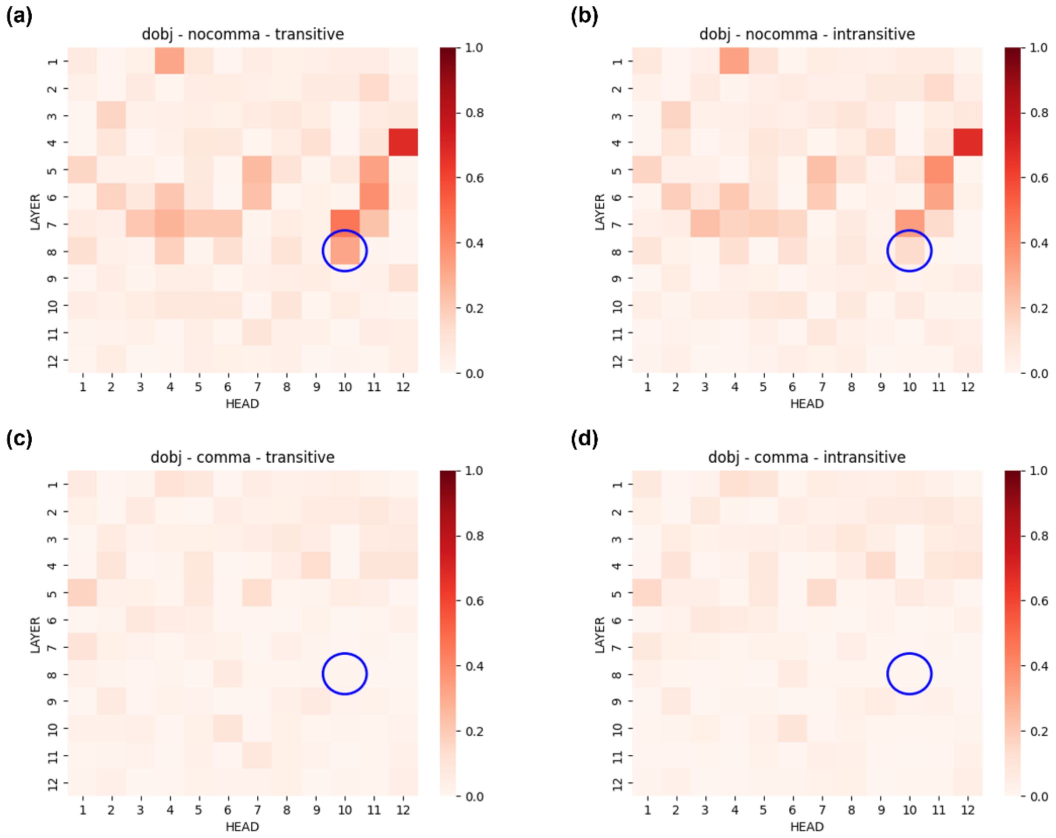

Figure 1. The Heatmap of Average Attention Weights for dobj (from "vet" to "scratched/struggled")

Here, y-axis is Layer and x-axis is Head. The circled area is the 10th head of the 8th layer. (a) Transitive – Garden-path (b) Intransitive – Garden-path (c) Transitive – Non Garden-path (d) Intransitive – Non Garden-path.

3.1.2. nsub

In the nsubj relation (Figure 4, 5), there was neither main effect nor significant interaction (p>.1) except for a main effect of Length (*estimate*=.1452, *SE*=0.0146, *t*=9.94, *p*<.001), which indicated that attention weights were greater in the short sentences than long sentences. In summary, the attention given to the nsubj relation was found to be relatively consistent regardless of Transitivity and Garden-path, while the absence of intervening words resulted in greater attention weights for this relation in short sentences where the distance between the subject and main verb was shorter.

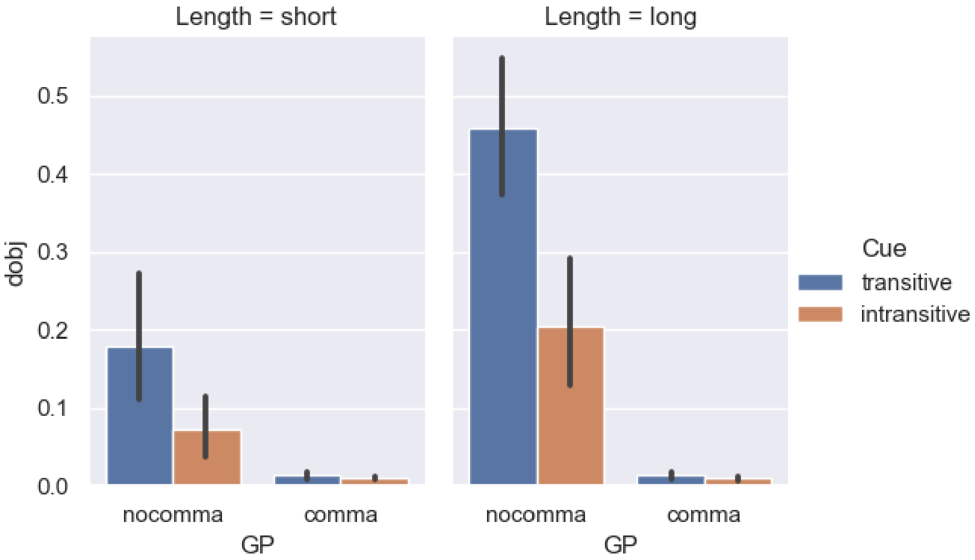

Figure 2. Bar Graphs of Mean Attention Weight for dobj at Layer 8 - Head 10

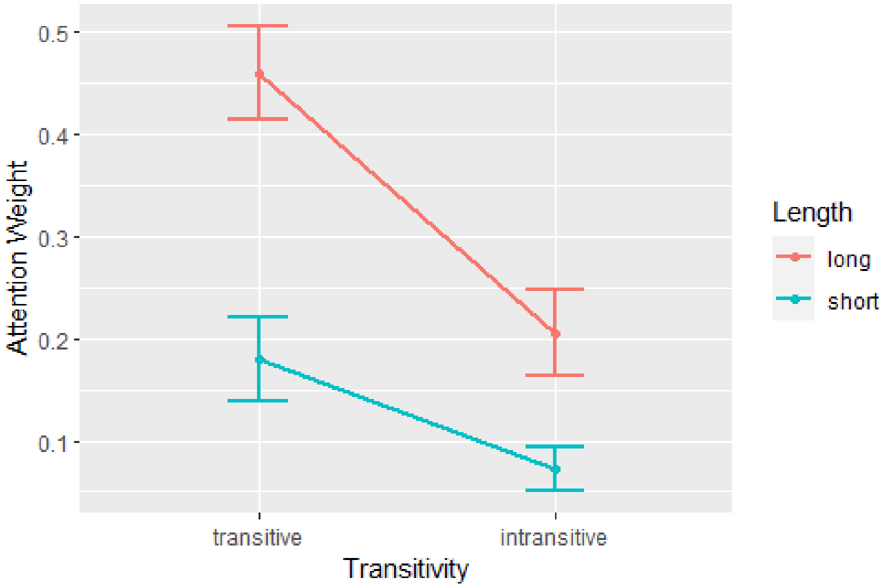

Figure 3. Interaction Plot Between Transitivity and Length only within the Garden-Path Conditions

3.2. Plausibility

3.2.1 dobj

In dobj relation (Figure 6, 7), there was a main effect of Garden-path (*estimate*=.3047, *SE*=0.0173, *t*=17.66, *p*<.001) and Length (*estimate*=-.1055, *SE*=0.0173, *t*=-6.11, *p*<.001) as well as a significant interaction between Garden-path and Length (*estimate*=-.2038, *SE*=0.0345, *t*=-5.91, *p*<.001). In short, the findings indicated that attention weights were greater in the garden path condition, particularly in the long sentence conditions, compared to the short sentence conditions.

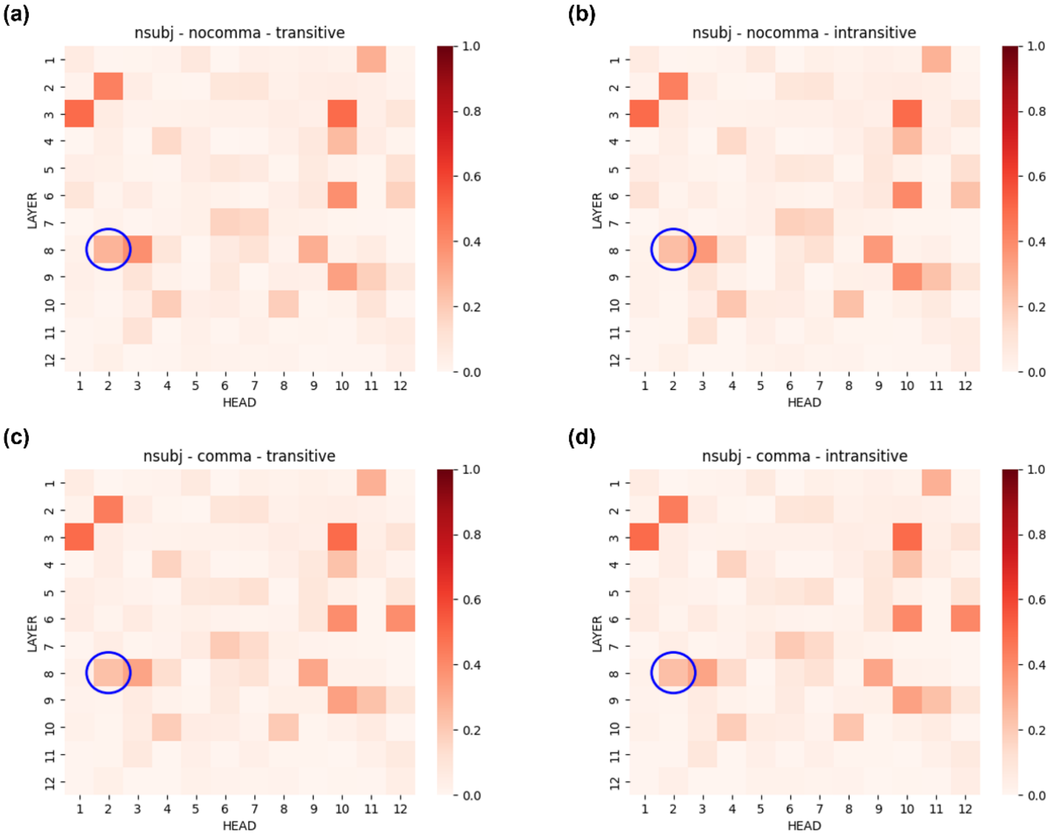

Figure 4. The Heatmap of Average Attention Weights for nsubj (from "vet" to "took (off)")

The circled area is the 2nd head of the 8th layer. (a) Transitive – Garden-path (b) Intransitive – Garden-path (c) Transitive – Non Garden-path (d) Intransitive – Non Garden-path

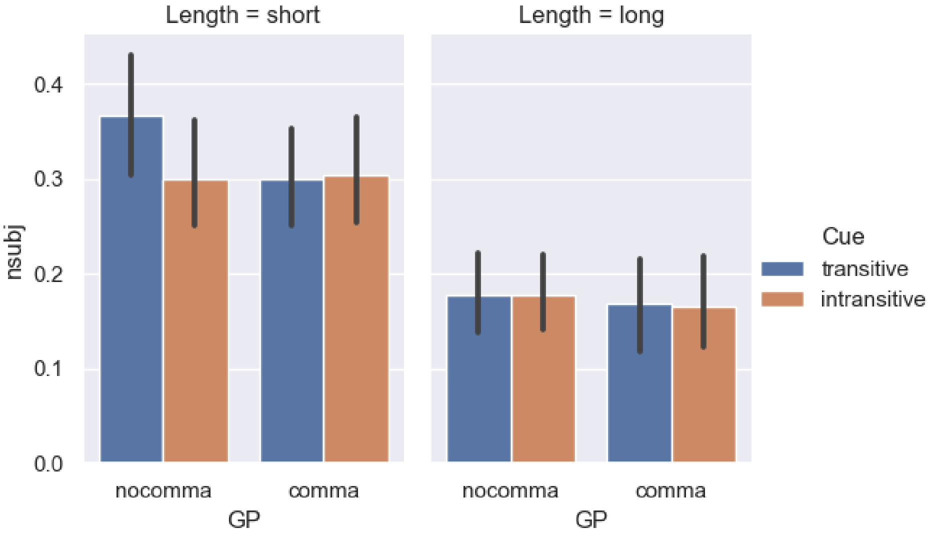

Figure 5. Mean Attention Weight for nsubj at Layer 8 - Head 2

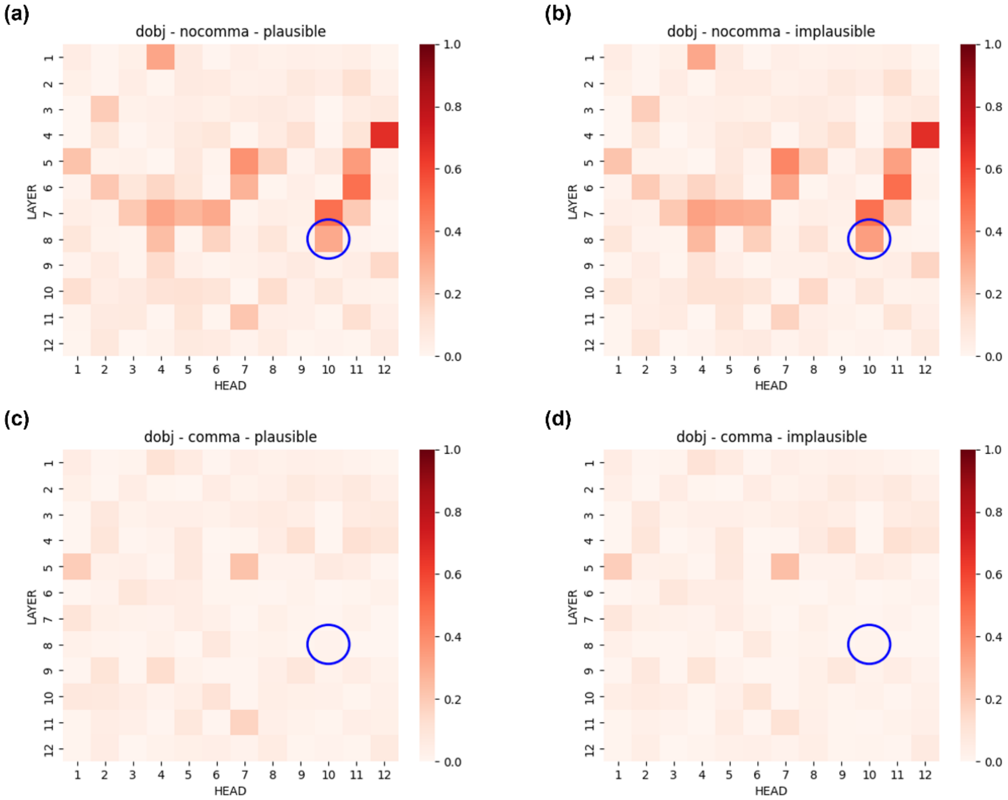

Figure 6. The Heatmap of Average Attention Weights for dobj (from "magazine" to "edited/sailed")

(a) Plausible – Garden-path (b) Implausible – Garden-path (c) Plausible – Non Garden-path (d) Implausible – Non Garden-path.

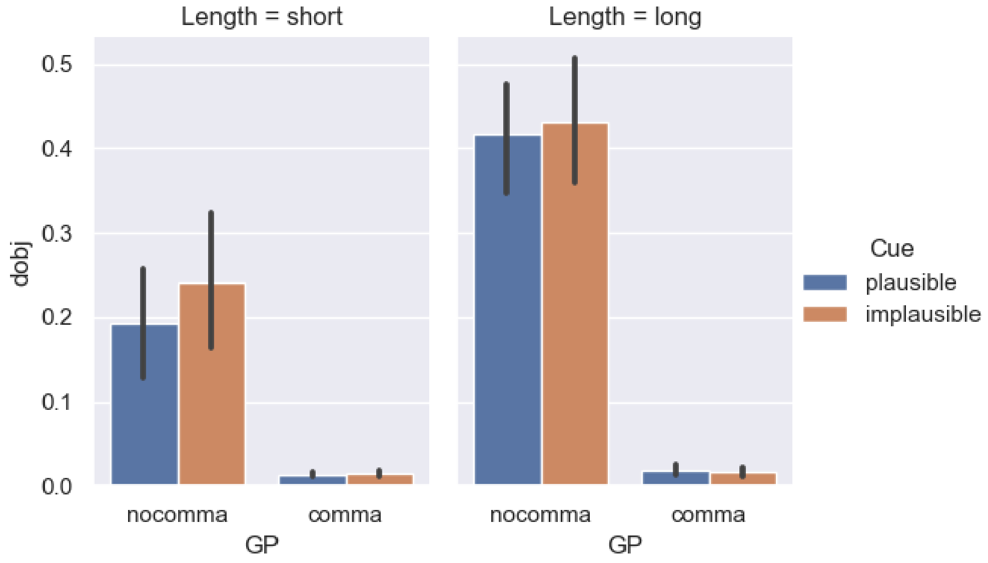

Figure 7. Mean Attention Weight for dobj at Layer 8 - Head 10

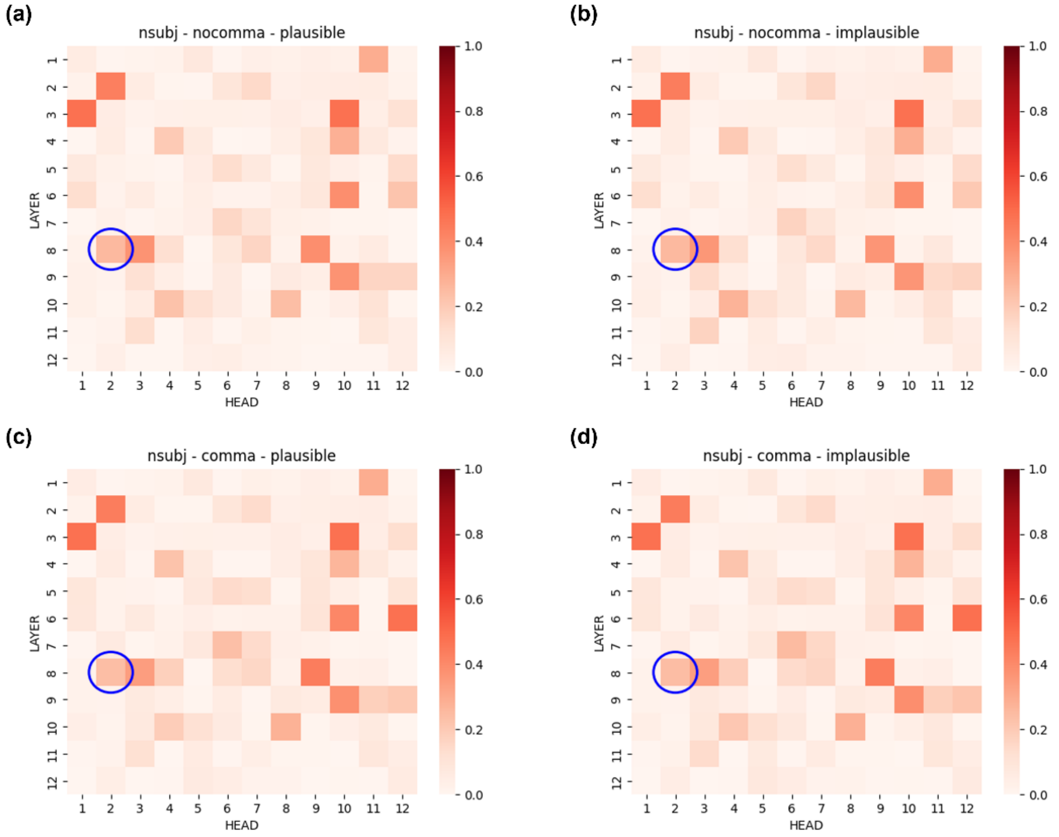

Figure 8. The Map of Average Attention Weights for nsubj (from "magazine" to "amused")

(a) Plausible – Garden-path (b) Implausible – Garden-path (c) Plausible – Non Garden-path (d) Implausible –

Non Garden-path.

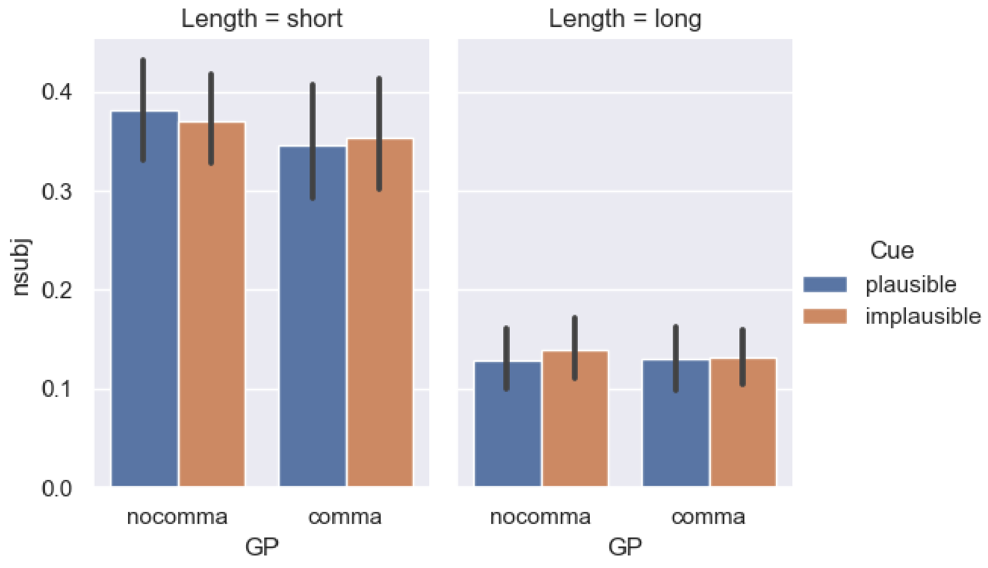

Figure 9. Mean Attention Weight for nsubj at Layer 8 - Head 2

3.2.2 nsubj

The nsubj relation (Figure 8, 9) did not yield significant interaction or main effects, with the exception of a main effect of Length (*estimate*=.2305, *SE*=0.0133, *t*=17.37, *p*<.001). This result implies that attention weights were greater in short sentences compared to long ones. Similar to transitivity, the presence of plausibility and comma did not affect the attention allocated to the nsubj relation, while the intervention of several words lowered it.

4. Discussion and Conclusion

This research aimed to explore the internal mechanism of BERT, by examining attention maps, particularly its handling of complex garden-path structures with specific linguistic cues such as transitivity and plausibility and compare this to human syntax processing. The findings revealed that BERT exhibited parallels to human syntactic processing in the manifestation of the garden-path effect and in the employment of transitivity and plausibility. However, while it was sensitive to transitivity, it was less so to plausibility, which did not entirely align with human syntactic processing or the results of surprisal. The following sections will discuss these findings in detail along with the implications of employing attention maps as a methodological tool.

First, the results revealed that BERT, due to the garden-path effect caused by subject-object ambiguity, led to misinterpretations. The main effect of Garden-path was found in both transitivity and plausibility items. In dealing with garden-path sentences, BERT often erroneously paid attention to the dobj relation at the 10th head of the 8th layer, even though "the vet" (or "the magazine") is not the object of "scratched/struggled" (or "edited/sailed"). This indicated that BERT, when a comma was absent, misinterpreted the subject of the main clause as the object of the conjunction clause when subject-object ambiguity occurred. This finding was consistent with the previous results from the studies of human syntactic processing and a surprisal experiment, where the garden-path effect comes into play due to a missing comma (Lee et al. 2022).

Secondly, BERT, analogous to human comprehenders, seemed to utilize the transitivity of the verb as a linguistic cue for sentence processing. The garden-path effect was more pronounced in transitive conditions than in intransitive conditions, which was supported by a significant interaction between Garden-path and Transitivity in attention weights of the dobj relation. In other words, BERT was less likely to misregard "the vet" as the object in the intransitive conditions than the transitive conditions. However, the main effect of the Garden-path was also observed, which suggested that the dobj relation was attended not only in the transitive condition but also in the intransitive conditions where the verb cannot accept the following noun as an object. This might indicate that BERT, in the absence of a comma, could not completely rule out the following noun being treated as an object even if it was syntactically not allowed.

The attention map pattern for transitivity in BERT exhibited similarities with human syntactic processing. Previous psycholinguistic studies have demonstrated that humans possess sensitivity towards transitivity (Adams et al. 1998, Mitchell 1987, Staub 2007, Van Gompel and Pickering, 2001), and there is also evidence that in certain situations, the subsequent nouns of intransitive verbs may also be mistaken as objects (Mitchell 1987, Van Gompel and Pickering 2001). However, there are contrasting findings regarding whether the garden-path effect occurs in intransitive conditions. While Van Gompel and Pickering (2001) identified a garden-path effect in intransitive conditions, Adams et al. (1998) provided evidence to the contrary. It is worth noting, nevertheless, that the sentences used in Van Gompel and Pickering (2001), which identified the garden path effect in intransitive verbs, contained several intervening words between the target noun ("vet") and the main verb ("took off"), mirroring the material of the long sentence condition in the present study, whereas Adams et al. (1998) employed materials that were similar to the short sentence condition in this study, where no additional words were included. This finding aligned with BERT's heightened attention towards dobj relation in intransitive conditions of the long sentences.

In contrast, BERT did not show sensitivity in plausibility in that there was only a main effect of Garden-path but no interaction with Plausibility in the analysis for dobj relation. This finding diverged from the results of the surprisal analysis (Lee et al. 2022), where the plausible conditions induced significantly greater surprisal than the implausible conditions for bert-base model, which was employed in this study. However, it is important to acknowledge that in the same study by Lee et al. (2022), a significant interaction between the Plausibility and Garden-path variables was not found for the majority of Transformer models, including bert-base. Therefore, it would be premature to assert that BERT manifests sensitivity to plausibility in the surprisal measurements, yet lacks such sensitivity when considering attention weights. Garden-path Likewise, this result appeared to be inconsistent with the findings of psycholinguistic experiments conducted on human processors which revealed longer total reading times and more regression at the disambiguating verb (e.g. amused) in the plausible conditions than in the implausible conditions (Pickering and Traxler 1998). Nevertheless, similar to the processing of transitivity, human compreheders also tended to automatically misinterpret the main subject as the object during the initial processing stage even in the implausibility conditions, as supported by increased reading times at the noun (e.g. "the magazine") in implausible conditions phrases (Pickering and Traxler 1998). Thus, while BERT failed to exhibit sensitivity to plausibility, it did not demonstrate a pattern entirely distinct from that observed in human syntactic processing.

The observations from the results invite the necessity of further exploration into BERT's insensitivity to plausibility as compared to its sensitivity to transitivity. As outlined in the Materials section, these two linguistic cues represent distinct facets that influence syntactic processing. Transitivity is predominantly associated with the structural information of verbs, whereas plausibility pertains to the semantic information affecting syntactic processing. Sensitivity towards transitivity implies that BERT has learned the structural information about whether a specific verb can accept an object or not. In contrast, sensitivity to plausibility is not limited to whether BERT has discerned that a particular verb-noun pairing can semantically form a verb phrase. It also sheds light on the extent to which BERT employs semantic information in syntactic processing. Despite the semantic implausibility of a verb-noun pair, the verb can still structurally accept an object. Consequently, while "sailed the magazine" may be semantically impractical, it remains structurally feasible (c.f. "sailed the yacht"). Hence, BERT's insensitivity to plausibility may not derive from its incapacity to comprehend that "sailed the magazine" is semantically unacceptable, but rather its decision to disregard such semantic cues in its structural resolution. This research paid special attention to the 10th head of the 12th layer concerning the dobj relation. The attention in this layer predominantly captures syntactic dependency, thus its determination of the dobj relation may be solely reliant on structural parameters. It can be conjectured that the semantic plausibility of the relationship between two words might be identified by a different head in a different layer. Interestingly, BERT exhibited greater surprisal in the implausible conditions (Lee et al. 2022). Given that surprisal is calculated based on the final output, it can be interpreted as encapsulating the aggregate information from the functionally distributed attention heads. Therefore, in addition to the syntactic dependency information exhibited in the 10th head of the 12th layer, other information was likely factored into the final output, indicating the influence of semantic information processed by other heads.

The findings above underscore the potential value of using attention maps, as well as offer insights on their effective use. The study examined the syntactic dependency specific to a particular attention head, illustrating the advantages of focusing on a particular piece of information via a specific attention head assessment. However, such a targeted strategy may also increase the possibility of neglecting information that is not processed in the designated head, highlighting the necessity of the parallel application of more comprehensive indicators such as surprisal and examination of attention weights in other heads. However, a significant challenge here is the limited knowledge about other heads. While Clark et al. (2019) have provided some insights on a number of heads processing specific information, this only represents a small fraction considering the diversity of linguistic phenomena. Many aspects of the specialized functions of each head still remain to be explored. In this context, the present study only examined the dobj and nsubj relations, representing a highly localized approach which has limitations in providing a comprehensive understanding.

Despite the initial misinterpretation, BERT appeared to be capable of abandoning this misinterpretation and reanalyzing it into an appropriate sentence structure. As shown in the results of nsubj, the degree to which the main subject attended the main verb in the attention head related with nsubj did not differ regardless of the presence of comma or linguistic cues. Although BERT had a greater tendency to mistake the main subject as the object of a conjunction clause in the garden-path conditions, this did not mean that it finished its processing while missing a subject. The relationship between the subject and the verb was attended in garden-path conditions to the same degree as in the other conditions. However, it should be noted that the absolute value of the attention weights for nsubj was relatively small, particularly in the long sentence conditions where intervening words were supposed to hinder finding the 'real' subject. The modest attention weights observed may be attributed to the characteristic of the nsubj relation. According to Clark et al. (2019), Layer 8-Head 2 is the layer and head associated with nsubj, but its accuracy was only around 60%, although this is significantly higher than the baseline. In other words, the attention weights for the nsubj relation were relatively small, as the relationship could potentially be attended to in other layers and heads. However, since this research had the limitation of relying exclusively on the nsubj relation, further investigation is necessary to accurately identify and assess other layers and heads associated with the nsubj relation.

In conclusion, this study has investigated the syntactic capacities and internal processing of BERT by examining attention maps instead of relying solely on indirect measurements such as accuracy or surprisals. The findings reveal that BERT demonstrates human-like syntactic processing when dealing with garden-path sentences, utilizing linguistic cues and correcting misinterpretations. Despite certain limitations such as reliance on a few syntactic relations, this research has provided valuable insights into the inner workings of BERT and contributes to a deeper understanding of how advanced neural language models acquire and process complex linguistic structures.

References

Adams, B. C., C. Clifton. and D. C. Mitchell. 1998. Lexical guidance in sentence processing? *Psychonomic Bulletin and Review* 5(2), 265-270.

Baayen, R. H., D. J. Davidson. and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4), 390-412.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever. and D. Amodei. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*

Clark, K., U. Khandelwal, O. Levy. and C. D. Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341.*

Devlin, J., M. W. Chang, K. Lee. and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

De Marneffe, M. C., and C. D. Manning. 2008. *Stanford typed dependencies manual* (pp. 338-345). Technical report, Stanford University.

Ettinger, A. 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.

Ferreira, F., and J. M. Henderson. 1991. Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language* 30(6), 725-745.

Ferreira, F., K. Christianson. and A. Hollingworth. 2001. Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research* 30, 3-20.

Frazier, L. and K. Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14(2), 178-210.

Futrell, R., E. Wilcox, T. Morita. and R. Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329.*

Futrell, R., E. Wilcox, T. Morita, P. Qian, M. Ballesteros. and R. Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260.*

Goldberg, Y. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287.*

Goodkind, A. and K. Bicknell. 2018, January. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, 10-18. Salt Lake City, Utah, USA, 7 January, 2018

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138.*

Hao, Y., S. Mendelsohn, R. Sterneck, R. Martinez. and R. Frank. 2020. Probabilistic predictions of people perusing:

Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954.*

Hoover, B., H. Strobelt. and S. Gehrmann. 2019. exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276.*

Hopp, H. 2015. Individual differences in the second language processing of object–subject ambiguities. *Applied Psycholinguistics* 36(2), 129-173.

Hu, J., J. Gauthier, P. Qian, E. Wilcox. and R. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692.*

Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering,* 9(03), 90-95.

Koroteev, M. V. 2021. BERT: A review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943.*

Kuncoro, A., L. Kong, D. Fried, D. Yogatama, L. Rimell, C. Dyer. and P. Blunsom. 2020. Syntactic structure distillation pretraining for bidirectional encoders. *arXiv preprint arXiv:2005.13482.*

Kuznetsova, A., P. B. Brockhoff. and R. H. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82, 1-26.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma. and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

Lee, J., J. A. Shin. and M. K. Park. 2022. (AL)BERT down the garden path: Psycholinguistic experiments for pre-trained language models. *Korean Journal of English Language and Linguistics* 22, 1033-1050.

Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3), 1126-1177.

Linzen, T., E. Dupoux. and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4, 521-535.

Marvin, R. and T. Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031.*

Mitchell, D. C. 1987. Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart ed., *Attention and Performance XII: The Psychology of Reading,* 601-618. London, Routledge

Oh, B. D., C. Clark. and W. Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence, 5, 777963*.

Oh, B. D. and W. Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?. *Transactions of the Association for Computational Linguistics* 11, 336-350.

Pickering, M. J. and M. J. Traxler. 1998. Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24(4), 940.

Posner, M.I. and S. E. Petersen. 1990. The attention system of the human brain. *Annual Review of Neuroscience* 13, 25-42.

R Core Team 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Smith, N. J. and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3), 302-319.

Staub, A. 2007. The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(3), 550.

Trueswell, J. C., M. K. Tanenhaus. and S. M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33(3), 285-318.

Van Gompel, R. P. and M. J. Pickering. 2001. Lexical guidance in sentence processing: A note on Adams, Clifton, and Mitchell 1998. *Psychonomic Bulletin and Review* 8, 851-857.

Van Schijndel, M. and T. Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of 40th Annual Meeting of the Cognitive Science Society*, 2600-2605. Madison, Wisconsin, USA, 25-28 July, 2018.

Van Schijndel, M., A. Mueller. and T. Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. *arXiv preprint arXiv:1909.00111*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, T. Kaiser. and I. Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*

Vig, J. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714.*

Vig, J. and Y. Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284.*

Waskom, M. L. 2021. Seaborn: statistical data visualization. *Journal of Open Source Software* 6(60), 3021.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis (2nd ed.). New York, NY: Springer*.

Wilcox, E., R. Levy, T. Morita. and R. Futrell. 2018. What do RNN language models learn about filler-gap dependencies?. *arXiv preprint arXiv:1809.00042.*

Wilcox, E., R. Levy. and R. Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. *arXiv preprint arXiv:1906.04068.*

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary