Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies

Jillian K. Da Costa
School of Management
University at Buffalo, SUNY
jillian.k.dacosta@buffalo.edu

Rui P. Chaves
Department of Linguistics
University at Buffalo, SUNY
rchaves@buffalo.edu

Abstract

Filler-gap dependencies are among the most challenging syntactic constructions for computational models at large. Recently, Wilcox et al. (2018) and Wilcox et al. (2019b) provide some evidence suggesting that large-scale general-purpose LSTM RNNs have learned such long-distance filler-gap dependencies. In the present work we provide evidence that such models learn filler-gap dependencies only very imperfectly, despite being trained on massive amounts of data. Finally, we compare the LSTM RNN models with more modern state-of-the-art Transformer models, and find that these have poor-to-mixed degrees of success, despite their sheer size and low perplexity.

1 Introduction

A flurry of recent work has shown that modern large-scale and general-purpose Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) achieve impressive results as computational psycholinguistic models of human language processing, such as Linzen et al. (2016), Gulordava et al. (2018), Linzen and Leonard (2018), van Schijndel and Linzen (2018), Futrell et al. (2019), and Wilcox et al. (2019a), to list only a few. Some of this work has focused on long-distance dependencies like (1), involving a linkage between a *wh*-phrase and a gap. This is one of the phenomena that Markovian language models have always been inherently bad at.

(1) I know **who**$_i$ the gardener reported the butler said the hostess believed her aunt suspected you delivered a challenge **to** $\__i$ at the party. (Wilcox et al., 2019b)

However, such long-distance dependencies are accompanied by morphosyntactic constraints which have not previously been tested, in particular, agreement constraints like those in (2).

(2) a. It was the lawyer who I think you said _ was/*were upset.
b. It was the lawyers who I think you said _ *was/were upset.
c. They wondered which lawyer I think you said _ was/*were upset.
d. They wondered which lawyers I think you said _ *was/were upset.

There are two different dependencies at work in the these examples. One is between the filler phrase *who* and the gap (i.e. the missing subject of the embedded verb) and another between the head noun *lawyer*(*s*) and the *wh*-phrase adjacent to it. It is not possible to claim that LSTM RNN models have learned English filler-gap dependencies without showing that the associated morphosyntactic constraints have also been learned. At the time of this writing, LSTM RNNs are no longer the state-of-the-art English language models. Transformer (attention-based) models have obtained lower test-time perplexity. In the present work we focus on whether any of these neural language models have truly learned long-distance agreement (filler-gap) dependencies like those in (1) and (2).

The structure of the paper is as follows. First we show that the same general-purpose LSTM RNN models that Wilcox et al. (2019b) have claimed to successfully cope with filler-gap dependencies have not learned the morphosyntactic constraints associated to such constructions, illustrated in (2). Next, we compare these results with those of three more recent transformer-based architectures that have obtained better perplexity results, namely Transformer-XL (Dai et al., 2019), BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), and OpenAI GPT-2 (Radford et al., 2019).[1]

[1] All our materials, code, and analysis are available at https://github.com/RuiPChaves/Transformers-FillerGap-dependencies.

We acknowledge that these models are not directly comparable, and that the present results should be taken with some caution because the architectures are different (transformer vs. recurrent), as are the training objectives (masked language modeling vs. non-masked language modeling), evaluation methods (use of sentences prefix + suffix vs. only prefix for language models), and the training datasets. Nonetheless, we argue that such a preliminary comparison is useful in that is sheds some light on how well extremely large neural models of English cope with perhaps of the most historically vexing syntactic phenomena in computational linguistics. As we shall see, there is a wide range of variation in how accurately the models cope with filler-gap dependencies, with LSTM RNNs fairing among the worse. Our results are consistent with those reported by Goldberg (2019), which suggest that BERT is better than LSTM RNNs at English subject-verb agreement (Marvin and Linzen, 2018).

2 LSTM RNNs

Wilcox et al. (2019b) found evidence suggesting that LSTM RNNs can maintain filler-gap dependencies across up to at least four clausal boundaries like the one in (1). Two models were used for these experiments. One was Gulordava et al. (2018), henceforth the **Gulordava model**, which was trained on 90 million tokens of English Wikipedia, and has two hidden layers of 650 units each. The second model was Jozefowicz et al. (2016), henceforth the **Google model**, which was trained on the One Billion Word Benchmark (Chelba et al., 2013), has two hidden layers with 8196 units each, and uses the output of a character-level Convolutional Neural Network as input to the LSTM. One of the trademark properties of filler-gap dependencies is that the morphosyntactic properties imposed on the gap site are preserved by the filler phrase, as already illustrated in (2). Here, the plural noun must be matched with the plural verb form and the singular noun with the singular verb. In what follows we examine how well these dependencies are learned by the Gulordava and Google models.

2.1 Experiment 1: agreement in clefts

Following basically the same experimental approach as Wilcox et al. (2018), we created 20 cleft items using a 2×2×4 factorial design, for a total of 320 sentences. All the conditions for an item are illustrated in (3). Like Wilcox *et al.*, we extracted the softmax activation of the critical verbs *were*/*was*, given the prefix sentence, using basically the same code as Wilcox et al. (2018), made available at `https://osf.io/zpfxm/`.

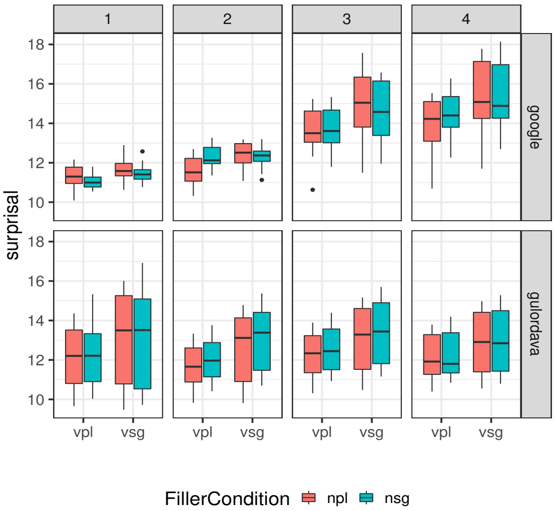

Figure 1: Surprisal of the gap-agreeing verb in ‘it’ clefts across 4 levels of embedding (LSTM RNNs)

(3) a. It was the lawyer(s) who I think was/were ... [$N_{sg/pl}$, LEVEL1, $V_{sg/pl}$]

b. It was the lawyer(s) who I think you said was/were ...
[$N_{sg/pl}$, LEVEL2, $V_{sg/pl}$]

c. It was the lawyer(s) who I think you said you thought was/were ...
[$N_{sg/pl}$, LEVEL3, $V_{sg/pl}$]

d. It was the lawyer(s) who people believe I think you said you thought was/were ...
[$N_{sg/pl}$, LEVEL4, $V_{sg/pl}$]

Finally, we converted the softmax values into surprisal (i.e. the negative log probability), following Wilcox et al. (2019b). See see Hale (2001) and Levy (2008) for more discussion.

The results were rather weak, as shown by Figure 1. Had the RNNs succeeded at this task, then the conditions where the noun and verb agree (i.e. N_{pl}-V_{pl} and N_{sg}-V_{sg}) would be lower in surprisal than the conditions where the agreement is mismatched (i.e. N_{pl}-V_{sg} and N_{sg}-V_{pl}). This was generally not the case in either model. Finally, in the larger Google model surprisal increased with the level of embedding, so that the correct verb form is more unexpected in level 4 than the incorrect verb forms in levels 1 and 2.

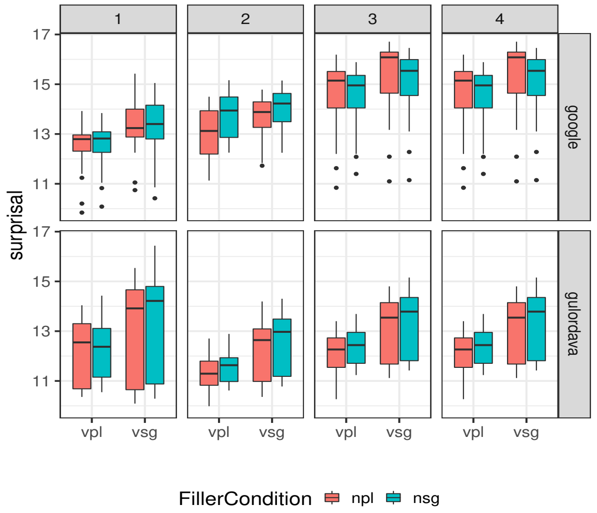

Figure 2: Surprisal of the gap-agreeing verb in 'which' interrogatives across embedding levels (LSTM RNNs)

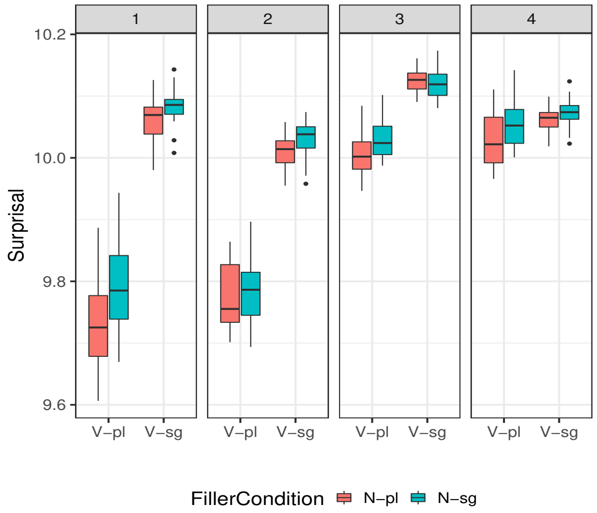

Figure 3: Surprisal of the gap-agreeing verb in 'it' clefts across embedding levels (Transformer-XL)

There is a general increase of surprisal as clausal embedding increases, which in our view may simply reflect the fact that multiple occurrences of embedded declarative clauses under verbs of indirect discourse are rare. Overall, the results suggest that these models have not learned the morphosyntax of filler-gap dependencies.

2.2 Experiment 2: agreement in indirect interrogatives

In order to assess if these results are specific to the cleft construction, we converted the 20 items into embedded interrogatives, effectively inverting the order of the *wh*-phrase and the agreeing nominal head, as (4) illustrates.

(4) a. Someone wondered which lawyer(s) I think was/were ...
[$\mathrm{N}_{sg/pl}$, LEVEL1, $\mathrm{V}_{sg/pl}$]

b. Someone wondered which lawyer(s) I think you said was/were ...
[$\mathrm{N}_{sg/pl}$, LEVEL2, $\mathrm{V}_{sg/pl}$]

c. Someone wondered which lawyer(s) I think you said you thought was/were ...
[$\mathrm{N}_{sg/pl}$, LEVEL3, $\mathrm{V}_{sg/pl}$]

d. Someone wondered which lawyer(s) people believe I think you said you thought was/were ...
[$\mathrm{N}_{sg/pl}$, LEVEL4, $\mathrm{V}_{sg/pl}$]

The outcome was the same, as Figure 2 shows, suggesting that our results are robust and not specific to the type of filler-gap construction chosen. We conclude that the Gulordava and Google models have not truly learned the morphosyntax of filler-gap dependencies. In what follows we examine how more recent transformer-based models fair at the same tasks.

3 Transformer-XL

Transformer-XL (Dai et al., 2019) has 24 million parameters, an average attention span of 640 tokens, and 16 10-word transformer layers. Transformer-XL is supposed to learn dependencies that are about 80% longer than those learned by RNNs but as Figure 3 shows, it did only marginally better than the Google and the Gulordava models when processing the same agreement in clefts dataset from Experiment 1.

In fact, only in embedding level 1 was the surprisal of agreeing N-V pairs statistically lower than their non-agreeing counterparts (for N_{pl}-V_{pl} vs. N_{sg}-V_{pl} we have t = -2.39, p = 0.021, and for N_{sg}-V_{sg} vs. N_{pl}-V_{sg} we have t = -1.83, p = 0.068). For all other levels of embedding there was no statistical difference in surprisal ($p > 0.4$), except for level 3 where N_{pl}-V_{pl} vs. N_{sg}-V_{pl} (t = -2.13, p = 0.039). The model does equally bad on the indirect interrogatives dataset from Experiment 2, as Figure 4 shows.

3.1 Experiment 3: Filler-gap surprisal in subject-inverted interrogatives

For completeness, we also tested Transformer-XL's ability to maintain a filler-gap dependency without the interacting factor of subject-verb

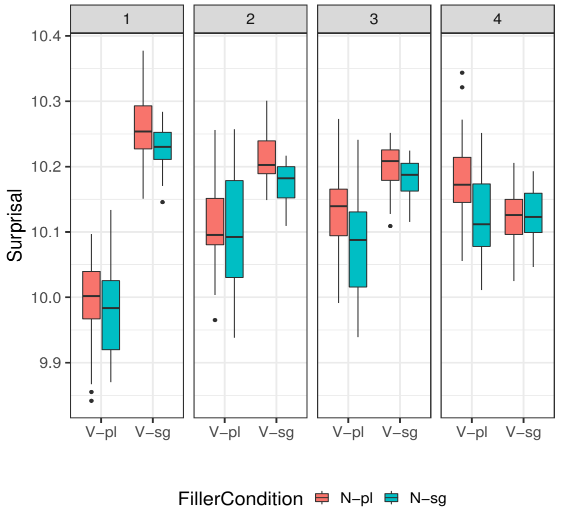

Figure 4: Surprisal of gap-agreeing verb in 'which' interrogatives across embedding levels (Transformer-XL)

agreement. We created 20 items, in a 2 × 2 × 4 design, for a total of 320 sentences, as illustrated in (5). We extracted the softmax value of the masked post-gap region item (below, the preposition *at*). This experiment serves as the counterpart of the experiments in Wilcox et al. (2019b) showing LSTM RNNs can maintain filler-gap dependencies across up to at least four clausal boundaries (diacritic '*' not included in the input).

(5) a.*What did we talk about it at the party?
[WH-NOGAP, LEVEL1]

b. What did we talk about _ at the party?
[WH-GAP, LEVEL1]

c. Did we talk about it at the party?
[NOWH-NOGAP, LEVEL1]

d.*Did we talk about _ at the party?
[NOWH-GAP, LEVEL1]

The results confirm that Transformer-XL has a poor representation for filler gap dependencies, as seen in Figure 5. Already at one level of embedding the surprisal of the (ungrammatical) nowh-gap condition is lower than the grammatical nh-gap counterpart, whereas it should be the other way around. In levels 2 through 4 there is no statistical difference between any of the four conditions.

3.2 Experiment 4: Filler-gap surprisal in uninverted indirect interrogatives

In order to determine if the results of Experiment 3 scale to other filler-gap constructions, we constructed non-inversion counterparts of the 20 items, illustrated in (6). As before, we extracted the softmax activation of the critical verbs at the end of the item, after the sentence prefix is processed. The results were similar in that in no level of embedding the correct surprisal pattern was observed. See the materials for details.

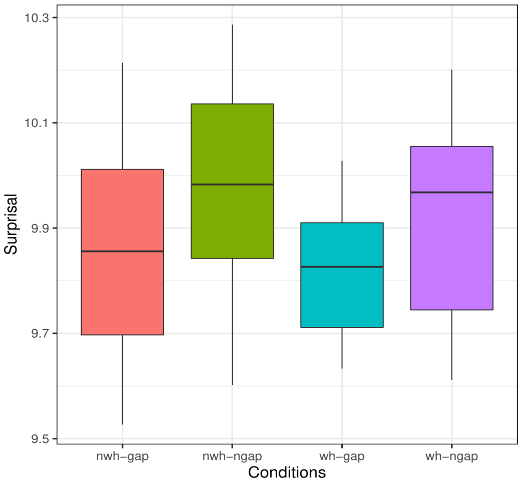

Figure 5: Surprisal of the post-gap region in inverted interrogatives at embedding level 1 (Transformer-XL)

(6) a. People wondered what we talked about it at ... [WH-NOGAP, LEVEL1]

b. People wondered what we talked about _ at ... [WH-GAP, LEVEL1]

c. People wondered if we talked about it at ... [NOWH-NOGAP, LEVEL1]

d. People wondered if we talked about _ at ... [NOWH-GAP, LEVEL1]

We conclude that the English Transformer-XL model does much worse than the English LSTM RNNs in coping with filler-gap dependencies.

4 BERT

Google's Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model that learns bidirectional encoder word representations via a masked language model training objective, using 340 million parameters, 768 hidden layers, 24 transformer blocks, and 1020 word context windows.

Using the same agreement in filler-gap dependencies dataset used in Experiment 1, we probe whether BERT assigns relative probability to plural and singular verb forms in such a way that this consistent with the agreement information of the

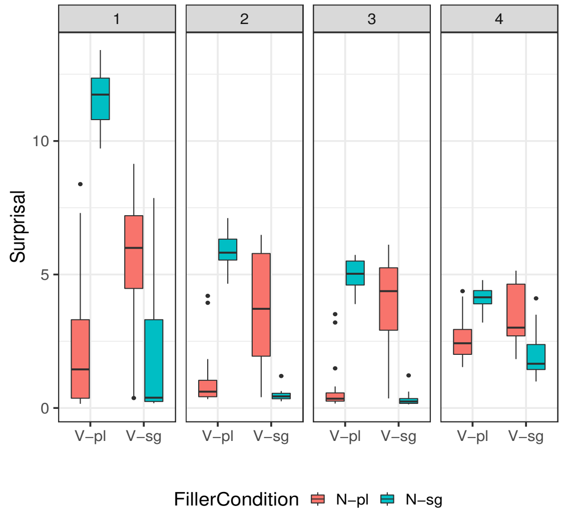

Figure 6: Surprisal of the gap-agreeing verb in 'it' clefts across 4 levels of embedding (BERT)

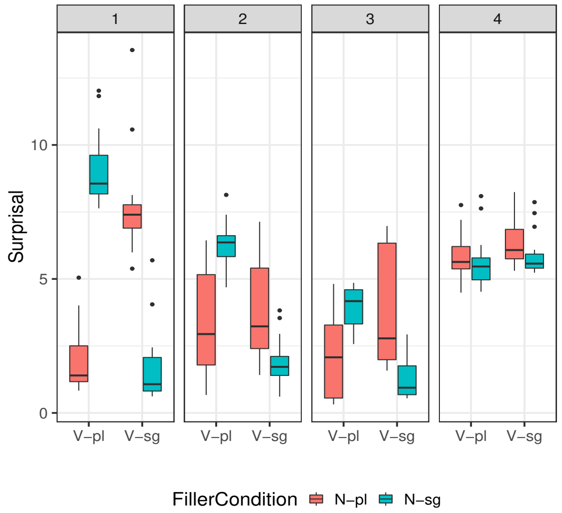

Figure 7: Surprisal of the gap-agreeing verb in 'which' questions across embedding levels (BERT)

nominal antecedent at the top of the filler-gap dependency. Like Goldberg (2019) and Wolf (2019), we masked the verb and then extracted the softmax values for both *was* and *were*, as shown in (7).

(7) a. It was the lawyer(s) who I think [MASK] upset. [N$_{sg/pl}$, LEVEL1]

b. It was the lawyer(s) who I think you said [MASK] upset. [N$_{sg/pl}$, LEVEL2]

c. It was the lawyer(s) who I think you said you thought [MASK] upset. [N$_{sg/pl}$, LEVEL3]

d. It was the lawyer(s) who people believe I think you said you thought [MASK] upset. [N$_{sg/pl}$, LEVEL4]

The results are much better than those obtained by LSTM RNNs on the same items, as Figure 6 shows. The surprisal of the agreeing conditions is systematically lower than that of the non-agreeing conditions in all embeddings (all $ps < 0.0001$).

In the next experiment, the 20 items were converted the *which* interrogative counterparts, analogously to Experiment 2 above, where the agreeing verb were masked, as seen in (8).

(8) a. Someone wondered which lawyer(s) I think [MASK] upset. [N$_{sg/pl}$, LEVEL1]

b. Someone wondered which lawyer(s) I think you said [MASK] upset. [N$_{sg/pl}$, LEVEL2]

c. Someone wondered which lawyer(s) I think you said you thought [MASK] upset. [N$_{sg/pl}$, LEVEL3]

d. Someone wondered which lawyer(s) who people believe I think you said you thought [MASK] upset. [N$_{sg/pl}$, LEVEL4]

The results are in Figure 7, and are only weak in embedding level 4, where neither condition is statistically different in the V$_{sg}$ (t = 0.91, $p = 0.36$) nor in the V$_{pl}$ (t = 1.93, $p = 0.06$) conditions.

If BERT's ability to maintain filler-gap dependencies in memory is too superficial and eager, then it may ignore the presence of a local subject, and not recognize that a subject gap is grammatically impossible, as in (9).

(9) a.*It was the boys who I think she/he were lost [N$_{pl}$, V$_{pl}$, LEVEL1]

b.*It was the boy who I think we/they was lost. [N$_{sg}$, V$_{sg}$, LEVEL1]

For example, if the model attempts to link *boys* to the copula verb in (9a) despite the local subject pronoun, then the surprisal of *were* should be higher than that of *was*. Similarly, if the model attempts to link *boy* to the copula verb in (9b) despite the local subject pronoun, then the surprisal of *was* should be lower than that of *were*. The presence of the pronoun makes the subject gap impossible, and BERT should be sensitive to that.

We therefore modified the 20 items used in the it-cleft experiment – originally illustrated in (7) – inserted a pronoun in the gap site, while making sure the verb agreed with the fronted phrase, not the pronoun. What we found was a complete reversal of the surprisal values. As Figure 8 shows,

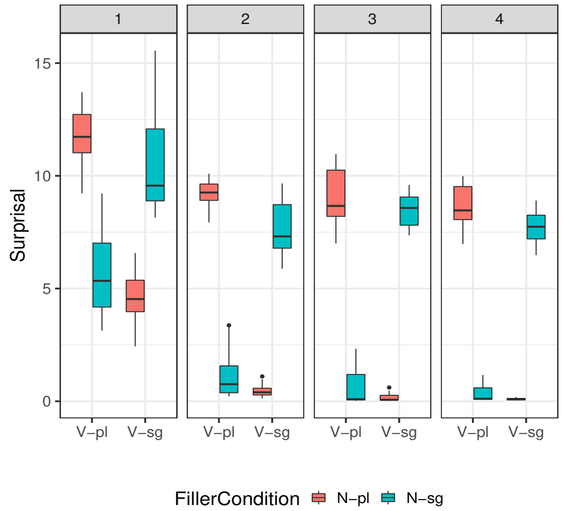

Figure 8: Surprisal of the (dis)agreeing verb in 'it' clefts across 4 levels of embedding (BERT)

BERT suspends the filler-gap linkages in the copula of examples like (9). We conclude that BERTs processing of filler-gap dependencies is not trivially shallow.

As in Experiments 3 and 4 above, we also examined BERT's ability to maintain a filler-gap dependency without the interacting factor of subject-verb agreement. Using the same items as in §3.1 and §3.2, illustrated in (10), we extracted the softmax value of the masked post-gap region item (below, the preposition *at*).

(10) a.*What did we talk about it at the party? [WH-NOGAP, LEVEL1]

b. What did we talk about _ at the party? [WH-GAP, LEVEL1]

c. Did we talk about it at the party? [NOWH-NOGAP, LEVEL1]

d.*Did we talk about _ at the party? [NOWH-GAP, LEVEL1]

As Figure 9 shows, BERT is able to represent the filler gap dependency up to four levels of clausal embedding. Surprisal is highest when there is a gap but no *wh*-phrase, and lower when (i) there is no gap and no *wh*-phrase and (ii) when there is a gap and a *wh*-phrase. The low surprisal obtained for the case where there is no gap and *wh*-phrase is more difficult to interpret, since the model's input has access to information about clausal boundaries. In that sense, the surprisal is lower than one would expect.

BERT faired equally well with the uninverted indirect interrogative counterparts of (5), shown in (11), which were identical to the items used in Experiment 4 above; see §3.2.

(11) a.*People wondered what we talked about it at the party. [WH-NOGAP, LEVEL1]

b. People wondered what we talked about _ at the party. [WH-GAP, LEVEL1]

c. People wondered if we talked about it at the party. [NOWH-NOGAP, LEVEL1]

d.*People wondered if we talked about _ at the party. [NOWH-GAP, LEVEL1]

BERT's masked language objective has an advantage over RNN models in that it has access to input after the masked critical item, e.g. the string *the party* in (5). We therefore ran a 2 × 2 × 4 variant of Experiment 6 in which the masked critical items were adverbs like *yesterday*, *repeatedly*, *again*, and *then*, in sentence-final position:

(12) a.* What did we talk about it yesterday? [WH-NOGAP, LEVEL1]

b. What did we talk about _ yesterday? [WH-GAP, LEVEL1]

c. Did we talk about it yesterday? [NOWH-NOGAP, LEVEL1]

d.* Did we talk about _ yesterday? [NOWH-GAP, LEVEL1]

The results were radically different, as the surprisal was essentially inverted as shown in Figure 10. This pattern remained the same in deeper embedding levels, suggesting that BERT's ability to maintain filler-gap dependencies is brittle.

Finally, we also ran a variant of this experiment where the 20 items were converted into embedded interrogatives, without inversion. Again, the masked critical items were the adverbs in sentence-final position:

(13) a.* People wondered what we talked about it repeatedly.
[WH-NOGAP, LEVEL 1]

b. People wondered what we talked about _ repeatedly.
[WH-GAP, LEVEL 1]

c. People wondered if we talked about it repeatedly.
[NOWH-NOGAP, LEVEL 1]

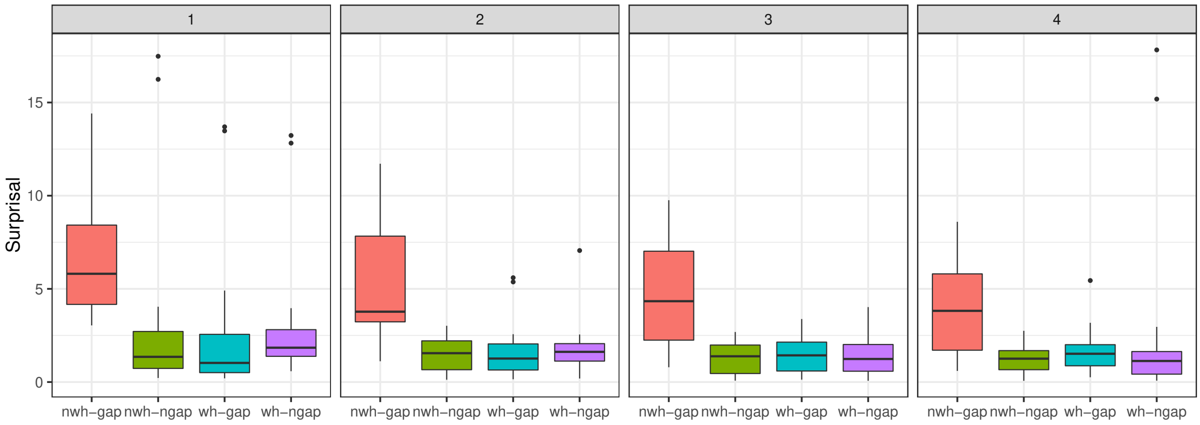

Figure 9: Surprisal of the post-gap region in subject-inversion interrogatives across embedding levels (BERT)

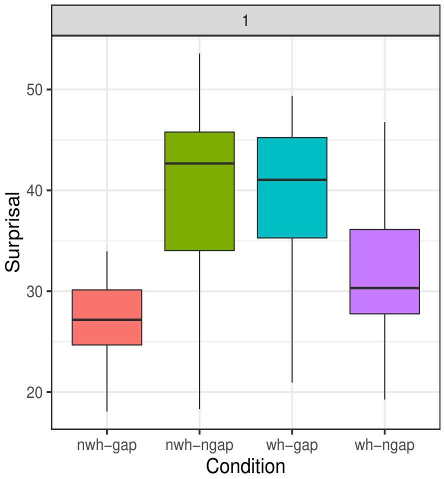

Figure 10: Surprisal of the sentence-final adverb in subject-inversion interrogatives, embedding 1 (BERT)

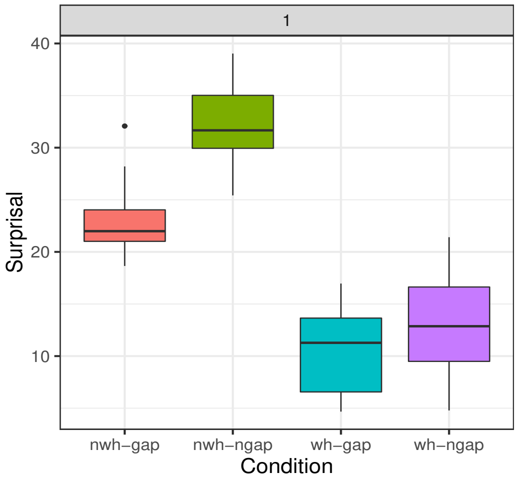

Figure 11: Surprisal of the sentence-final adverb in uninverted indirect interrogatives at embedding 1 (BERT)

d.* People wondered if we talked about _ repeatedly.
[NWH-GAP, LEVEL 1]

Now, the condition with the highest surprisal was nwh-ngap, suggesting that the model does not expect sentence-final adverbs to follow pronouns in the absence of a filler-gap dependency. The first level of embedding is shown in Figure 11. BERT's modelling of filler-gap dependencies is better than all other models surveyed so far but still brittle.

5 XLNet

XLNet (Yang et al., 2019) is like BERT in that it uses a masked model training objective and learns bidirectional contexts. Although XLNet is claimed to achieve better results than BERT in a number of tasks, we found that it performed worse in the same experiments ran on BERT, failing to provide clear evidence that filler-gap dependencies are attended to. For example, XLNet did much worse with clefts items, like those illustrated in (8). As can be seen in Figure 12, there is a significant overlap across subject-verb agreeing and non-agreeing conditions. Had the model learned about agreement in filler-gap dependencies, the surprisal of V_{pl} (*were*) in the N_{pl} condition should be significantly lower than that of V_{pl} in the N_{sg} condition. Similarly, the surprisal of V_{sg} in the N_{pl} condition should be significantly higher than that of V_{sg} (*was*) in the N_{sg} condition.

Similarly poor results were found for the interrogative subject-agreement items, like those in (7), as Figure 13 indicates. As in the case of Transformer-XL, there is little evidence that the model attends to filler-gap dependencies at all.

6 GPT-2

Unlike Google's BERT, the OpenAI GPT-2 model uses the same training objective as LSTM RNNs.

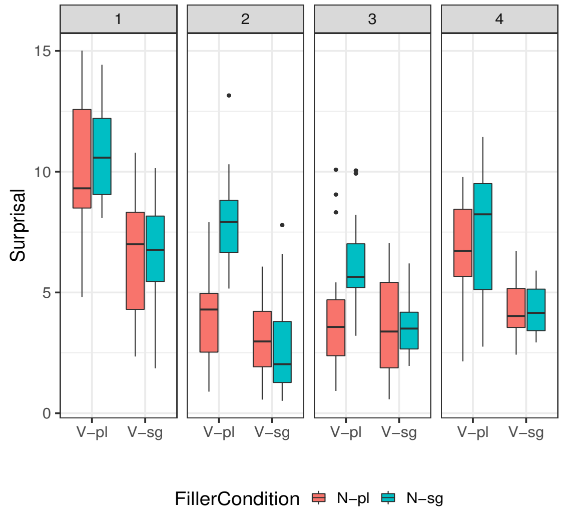

Figure 12: Surprisal of the gap-agreeing verb in ‘it’ clefts across 4 levels of embedding (XL-Net)

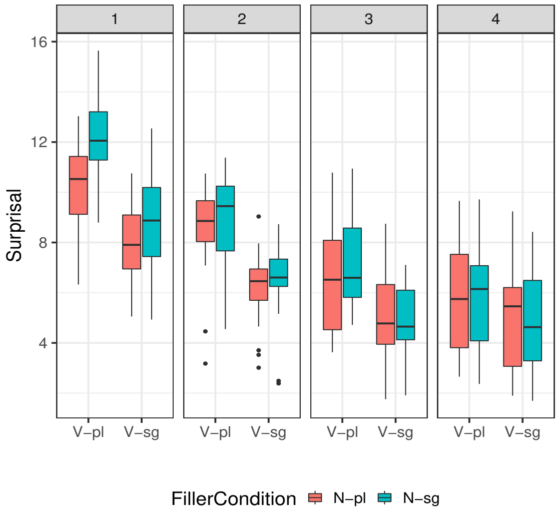

Figure 13: Surprisal of the gap-agreeing verb in ‘which’ questions across embedding levels (XLNet)

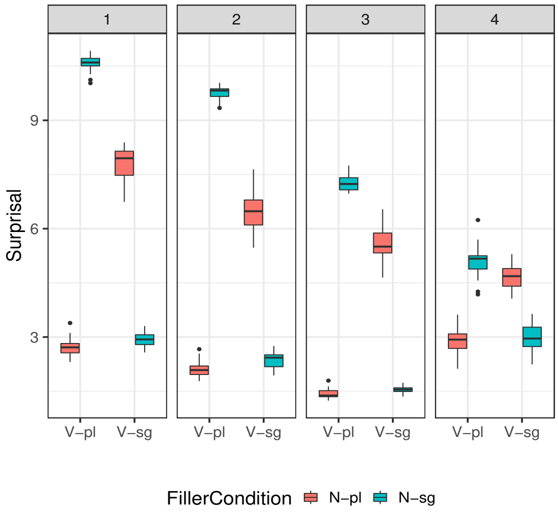

Figure 14: Surprisal of the gap-agreeing verb in ‘it’ clefts across 4 levels of embedding (GPT-2)

It is therefore possible to simply take the softmax activation of the word of interest after the sentence is processed. Preliminary evaluations on subject-verb agreement data by Wolf (2019) indicate that GPT-2 is worse than BERT on the Linzen et al. (2016) dataset but better in the more complex Marvin and Linzen (2018) dataset. In what follows, we report our findings for the more recent 345 million parameter version of GPT-2, hf. **GPT-2 medium**.

We begin with the 20 cleft items from Experiment 1, illustrated in (3), and repeated in (14). As before, we extracted the softmax activation of the words *was* and *were* across all conditions and converted the values to surprisal.

(14) a. It was the lawyer(s) who I think was/were ... [N$_{sg/pl}$, LEVEL1, V$_{sg/pl}$]

b. It was the lawyer(s) who I think you said was/were ...
[N$_{sg/pl}$, LEVEL2, V$_{sg/pl}$]

c. It was the lawyer(s) who I think you said you thought was/were ...
[N$_{sg/pl}$, LEVEL3, V$_{sg/pl}$]

d. It was the lawyer(s) who people believe I think you said you thought was/were ...
[N$_{sg/pl}$, LEVEL4, V$_{sg/pl}$]

The GPT-2 medium results are shown in Figure 14, and are clearly superior to BERT’s. For all levels of embedding, the agreeing conditions received statistically lower surprisal than that of the non-agreeing conditions. Notice how the differential across the conditions tends to diminish with deeper clausal embeddings, suggesting that the dependency is lost in deeper embeddings.

The dataset from Experiment 2 – consisting of *which* embedded interrogative like those in (4), repeated here as (15) – yielded virtually the same results, as shown in Figure 15. This suggests that GPT-2 medium is cross-constructionally robust up four levels of clausal embedding.

(15) a. Someone wondered which lawyer(s) I think was/were ...
[N$_{sg/pl}$, LEVEL1, V$_{sg/pl}$]

b. Someone wondered which lawyer(s) I think you said was/were ...
[N$_{sg/pl}$, LEVEL2, V$_{sg/pl}$]

c. Someone wondered which lawyer(s) I think you said you thought was/were ...

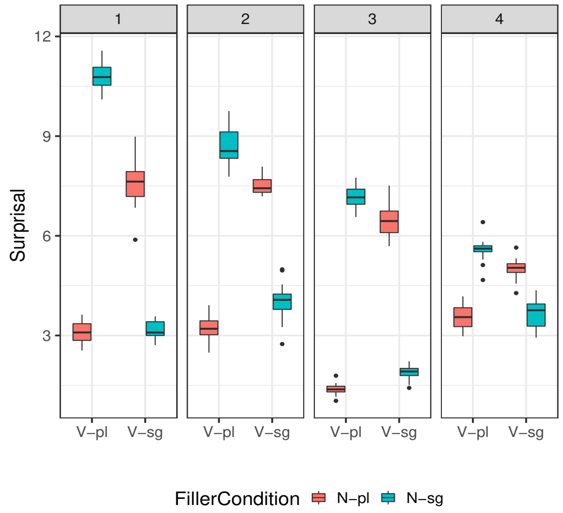

Figure 15: Surprisal of the gap-agreeing verb in ‘which’ questions across levels of embedding (GPT-2)

[N$_{sg/pl}$, LEVEL3, V$_{sg/pl}$]

d. Someone wondered which lawyer(s) who people believe I think you said you thought was/were ...
[N$_{sg/pl}$, LEVEL4, V$_{sg/pl}$]

For completeness, we also examined GPT-2’s ability to maintain a filler-gap dependency without the interacting factor of subject-verb agreement in both clefts and interrogatives, analogously to what was done in Experiments 3 and 4. The same items were used, and as in the LSTM RNN and Transformer-XL cases we extracted the softmax activation of the word at the end of the item, after the prefix string is processed.

As Figure 16 shows, GPT-2 medium performed moderately well for the 20 cleft items (same data as Experiment 3), though the results were not as strong as BERT’s. One major difference is that the surprisal of the wh-gap condition was systematically higher than that of the nwh-ngap condition. Ideally, the two should overlap. The relatively high surprisal of the wh-ngap condition is arguably due to the model maintaining expectations that the gap is further downstream in the sentence. Still, the results overall suggest that the filler-gap dependency is attended do.

The subject inversion counterpart of the 20 items (same data as Experiment 4) led to results closer to BERTs, whereby the surprisal of the wh-gap condition overlapped with that of nwh-ngap condition (all $p > 0.3$), as seen in Figure 17. In both of these experiments, the results were the same in subsequent embeddings.

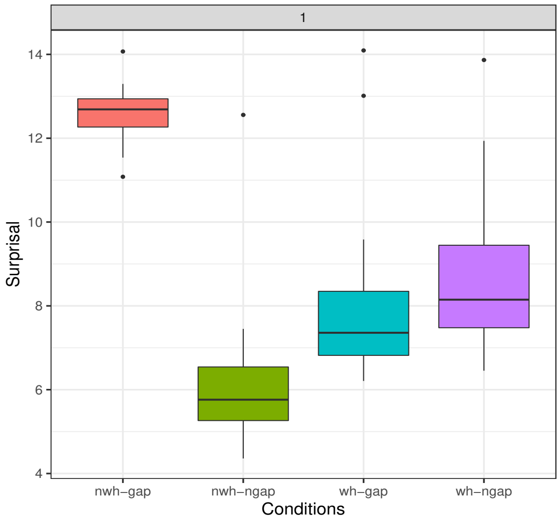

Figure 16: Surprisal of the post-gap region in uninverted indirect interrogatives in embedding 1 (GPT-2)

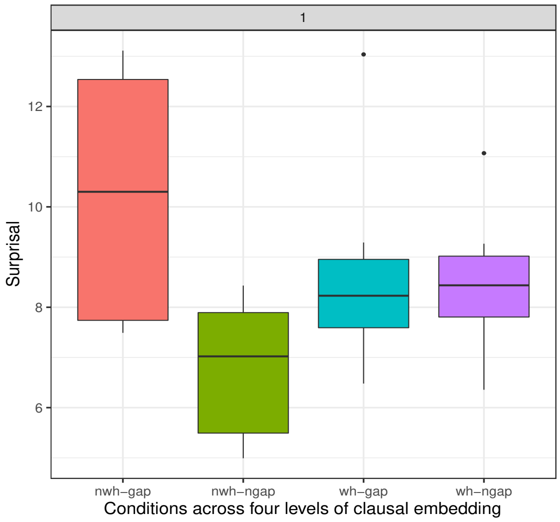

Figure 17: Surprisal of the post-gap region in inverted interrogatives in embedding 1 (GPT-2)

7 Discussion

Filler-gap dependencies still pose challenges for general-purpose large-scale state-of-the-art neural architectures. We show LSTM RNNs fair very poorly, despite the results of Wilcox et al. (2018) and Wilcox et al. (2019b). More modern models like Transformer-XL and XLNet do even worse.

However, BERT and GPT-2 perform rather well, although not without some mixed results. For example, the performance differs significantly across different kinds of filler-gap dependency, which suggests that the models are somewhat brittle even though they are extremely large, and were trained on an enormous amount of data.

References

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*.

Zihang Dai, Zhilin Yang, Yiming Yang, William W. Cohen, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. ArXiv:1901.02860v3, arxiv.org/pdf/1901.02860.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv:1810.04805, http://arxiv.org/abs/1810.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. Unpublished ms. https://arxiv.org/pdf/1901.05287.pdf.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*, pages 1195–1205.

John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001, Pittsburg, PA*, pages 159–166. ACL.

Rafal Jozefowicz, Vinyals Oriol, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 692–697. Cognitive Science Society, Austin, TX.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In Tim Rogers, Marina Rau, Jerry Zhu, and Chuck Kalish, editors, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2600–2605. Cognitive Science Society, Austin, TX.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of Blackbox NLP at ACL*, page pp.10.

Ethan Wilcox, Roger P. Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? In *Proceedings of the Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Ethan Wilcox, Roger P. Levy, Takashi Morita, and Richard Futrell. 2019b. What syntactic structures block dependencies in RNN language models? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci)*.

Thomas Wolf. 2019. Some additional experiments extending the tech report "Assessing BERT's Syntactic Abilities" by Yoav Goldberg. Unpublished ms. huggingface.co/bert-syntax/extending-bert-syntax.pdf.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. arxiv.org/abs/1906.08237.