Can language models capture syntactic associations without surface cues? A case study of reflexive anaphor licensing in English control constructions

Soo-Hwan Lee
Department of Linguistics
New York University
soohwan.lee@nyu.edu

Sebastian Schuster
Department of Linguistics
Center for Data Science
New York University
schuster@nyu.edu

1 Introduction

Recent studies have shown that language models (LMs) have the ability to capture many long-distance dependencies such as filler-gap dependencies (Wilcox et al., 2018) and subject-verb agreement (Linzen et al., 2016) despite only learning from surface strings. However, this ability has primarily been shown for constructions for which the surface strings frequently provide information about dependencies in the form of agreement patterns. For example, if a model has access to sentences with and without a noun phrase intervening between the subject and the main verb (1), it is often able to infer the agreement dependencies from the surface string alone: (Linzen et al., 2016; Marvin and Linzen, 2018; Goldberg, 2019; Gulordava et al., 2018; Hu et al., 2020b). The surface cues are boldfaced in (1):

(1) The **girls** who the boy likes **are** smiling.

Importantly, such agreement patterns are not available for all constructions. Consider, for example, English control constructions with non-finite embedded clauses (2-3). The main verb in the embedded clause cannot be inflected and therefore the clause generally lacks agreement information. The main exception to this is when the embedded clause contains a reflexive anaphor (e.g., *himself*). In such cases, the anaphor refers to either the subject or the object in the higher clause (the *controller*) and thus has to agree with the controller. In (2), the anaphor *himself* is co-referential with the subject under the subject control predicate *promise*. In (3), the anaphor is co-referential with the object under the object control predicate *persuade*.

(2) The **artist** promised the lawyers to make fun of **himself**. (Subject control)

(3) The artists persuaded the **lawyer** to make fun of **himself**. (Object control)

Given the lack of agreement information on the verb, it is difficult to infer whether the controller should be the subject or the object of the matrix clause from the surface string alone, unless the embedded clause contains a reflexive anaphor. Such constructions, however, are almost non-existent in corpora.[1] Hence, LMs trained on naturalistic corpora likely fail to capture this type of dependency.

In this work, we examine a Transformer-based LM, namely Generative Pre-trained Transformer 2 (GPT-2) (Radford et al., 2019), which is trained only on surface strings, to see whether or not the model makes correct predictions about the agreement patterns of reflexive pronouns in subject and object control constructions. Our findings show that GPT-2 struggles with subject control constructions such as (2), but does quite well on object control constructions such as (3). One reason might be that the model tries to associate the anaphor with the closest noun phrase. Moreover, while we find that a model with a larger number of parameters shows higher accuracy on the tasks related to subject control constructions, performance remains below chance and the model does not mimic human behavior.

2 Language model

We evaluated to what extent an LM predicts the correct agreement patterns for subject and object control constructions involving a reflexive anaphor. Given its strong performance on many other syntactic tasks (Warstadt et al., 2020), we used GPT-2 (Radford et al., 2019) through the HuggingFace Transformer library (Wolf et al., 2020). GPT-2 uses a self-attention mechanism that enables it to learn to focus on certain parts of the input that are

[1] For example, the Corpus of Contemporary American English (Davies, 2008), which contains more than 1 billion words, includes exactly one example with *promise* in which a reflexive agrees with the controller.

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 206-211.
Held on-line February 7-9, 2022

		Condition	Example
With object	SUBJECT CONTROL		
	Promise	Baseline	The **lawyer** promised the artist to make fun of **himself**.
		Distractor	The **lawyer** promised the artists to make fun of **himself**.
		Ungrammatical	*The **lawyers** promised the artist to make fun of **himself**.
	Offer	Baseline	The **lawyer** offered the artist to make fun of **himself**.
		Distractor	The **lawyer** offered the artists to make fun of **himself**.
		Ungrammatical	*The **lawyers** offered the artist to make fun of **himself**.
	Guarantee	Baseline	The **lawyer** guaranteed the artist to make fun of **himself**.
		Distractor	The **lawyer** guaranteed the artists to make fun of **himself**.
		Ungrammatical	*The **lawyers** guaranteed the artist to make fun of **himself**.
	OBJECT CONTROL		
	Persuade	Baseline	The lawyer persuaded the **artist** to make fun of **himself**.
		Distractor	The lawyers persuaded the **artist** to make fun of **himself**.
		Ungrammatical	*The lawyer persuaded the **artists** to make fun of **himself**.
	Tell	Baseline	The lawyer told the **artist** to make fun of **himself**.
		Distractor	The lawyers told the **artist** to make fun of **himself**.
		Ungrammatical	*The lawyer told the **artists** to make fun of **himself**.
	Force	Baseline	The lawyer forced the **artist** to make fun of **himself**.
		Distractor	The lawyers forced the **artist** to make fun of **himself**.
		Ungrammatical	*The lawyer forced the **artists** to make fun of **himself**.
No object	SUBJECT CONTROL		
	Promise	Baseline	The **lawyer** promised to make fun of **himself**.
		Ungrammatical	*The **lawyers** promised to make fun of **himself**.
	Offer	Baseline	The **lawyer** offered to make fun of **himself**.
		Ungrammatical	*The **lawyers** offered to make fun of **himself**.
	Guarantee	Baseline	The **lawyer** guaranteed to make fun of **himself**.
		Ungrammatical	*The **lawyers** guaranteed to make fun of **himself**.

Table 1: Associates are **boldfaced**. Baseline, Distractor, Ungrammatical conditions are based on Hu et al. (2020a).

recognized to be more important for predicting the next word than others. The model is pre-trained on the WebText dataset (Radford et al., 2019) which is estimated to contain 8 billion tokens (see Warstadt et al., 2020). The corpus is tokenized into sub-word units using the byte pair encoding compression algorithm (Sennrich et al., 2016). GPT-2 is an autoregressive language model, that is, its pre-training objective is a next-token prediction task in which it aims to maximize the probability of each token given its previous tokens.

To examine whether an increase in the number of parameters affects performance on the agreement task, we evaluated two differently sized pre-trained GPT-2 models: GPT-2 (small) with ∼117 million parameters and GPT-2 XL with ∼1.5 billion parameters. Both models were trained on the same corpus and only differ in their number of parameters.

3 Experimental design

The frequency of each reflexive pronoun in English (e.g., *himself*, *herself*, and *themselves*) differs greatly from one another in many standard corpora (Hu et al., 2020a). In order to minimize this confound, we keep the reflexive word constant in all of our stimuli and vary the preceding context as little as possible. Table 1 shows our example stimuli with the reflexive anaphor, *himself*, embedded in a non-finite clause. We used *himself* instead of *herself*, since *himself* is usually more frequent than *herself* in corpora. We avoided using *themselves* mainly due to its number-neutral usage. Under our experimental design, the anaphor *himself* is associated with either the subject or the object in the matrix clause depending on the matrix predicate (e.g., *promise* or *persuade*). We used 5 noun phrases for subjects and objects, 3 matrix verbs for subject control, 3 matrix verbs for object control, and 5 embedded clauses (see Appendix A).

Adapting Hu et al.'s (2020a) experimental design, we generated grammatical sentences by matching the number of the reflexive anaphor and the controller (the *associates*) while being flexible about the number of the non-associate. The 'Baseline' condition consists of (non-)associates that always match in number. The 'Distractor' condition consists of a non-associate that differs from the associates in number. The associates are boldfaced and the non-associates are underlined in (4-5):

(4) The **lawyer** promised the artist to make fun of **himself**. (Baseline)

(5) The **lawyer** promised the artists to make fun of **himself**. (Distractor)

For the 'Ungrammatical' condition, the number of the associates are mismatched while the number of the anaphor and the non-associate are matched as shown in (6):

(6) *The **lawyers** promised the artist to make fun of **himself**. (Ungrammatical)

As mentioned in the previous section, GPT-2 assigns a probability to every token in a sentence based on its preceding tokens. For minimal pairs such as (4-6), we expect the probability assigned to *himself*, P(himself), in the 'Ungrammatical' condition to be lower than P(himself) in both the 'Baseline' and 'Distractor' conditions. Hence, chance accuracy is 33%. We constructed 100 minimal pairs for each of the matrix verbs shown in Table 1.

Since LM performance on reflexive anaphor licensing has generally been mixed (Marvin and Linzen, 2018; Futrell et al., 2019; Hu et al., 2020a), we also examined whether GPT-2 can make correct associations between the reflexive anaphor and the controller when there is no distracting noun (non-associate) intervening between the two. Hence, we examined simple control cases where the non-associate is absent using subject control constructions (7-8). Note that this is not possible with object control constructions, since neither the subject nor the object can be omitted.

(7) The **lawyer** promised to make fun of **himself**. (Baseline)

(8) *The **lawyers** promised to make fun of **himself**. (Ungrammatical)

We constructed 25 minimal pairs: 25 sentences for the 'Baseline' condition and 25 sentences for the 'Ungrammatical' condition. We expect P(himself) in the 'Ungrammatical' condition to be lower than P(himself) in the 'Baseline' condition. Hence, chance accuracy is 50%.

4 Results

Table 2 shows that GPT-2 (small)'s mean accuracy on subject control constructions with objects (4%) is significantly lower than its mean accuracy on object control constructions (100%). The larger GPT-2 XL shows higher accuracy on subject control constructions used with the matrix verbs *promise* (13% → 47%) and *offer* (0% → 20%). However, GPT-2 XL's accuracy on subject control constructions used with the matrix verb *guarantee* more or less remains the same (0% → 3%). The model's mean accuracy on subject control constructions with objects (23%) is thus still below chance accuracy (33%) and is significantly lower than its mean accuracy on object control constructions (97%). The results from the control experiment in Table 3 show that the poor performance on subject control with objects cannot be attributed to the issues related to reflexive anaphor licensing per se. Both models perform at ceiling on sentences without objects (100%), which suggests that the models are generally able to predict licensing patterns between reflexives and noun phrases based on number.

Taken together, the results suggest that both versions of GPT-2 primarily rely on the heuristic to associate the reflexive anaphor with the object NP. One likely reason for this behavior is that the reflexive anaphor is linearly closer to the object than to the subject. Given that syntactically complex sentences are not commonly represented in corpora (Marvin and Linzen, 2018), it is likely that the model learned to associate reflexives with the linearly closest noun phrase from naturalistic training corpora. Further, that both versions of GPT-2 perform similarly poorly suggests that an increase in the number of parameters does not lead to a considerable increase in accuracy.

	GPT-2 (small)	GPT-2 XL
Promise	0.13	0.47
Offer	0.00	0.20
Guarantee	0.00	0.03
Mean	0.04	0.23
Persuade	1.00	0.95
Tell	1.00	0.95
Force	1.00	1.00
Mean	1.00	0.97

Table 2: GPT-2 performance on transitive subject and object control constructions (with object). Mean accuracy for each type of constructions is included. Chance accuracy is 0.33.

	GPT-2 (small)	GPT-2 XL
Promise	1.00	1.00
Offer	1.00	1.00
Guarantee	1.00	1.00
Mean	1.00	1.00

Table 3: GPT-2 performance on intransitive subject control constructions (no object). Mean accuracy is included. Chance accuracy is 0.50.

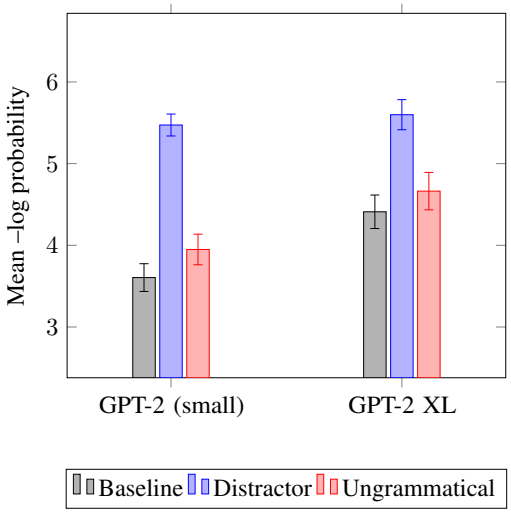

Figure 1: Mean negative log probability at the reflexive anaphor in transitive subject control constructions. Error bars indicate 95% confidence intervals.

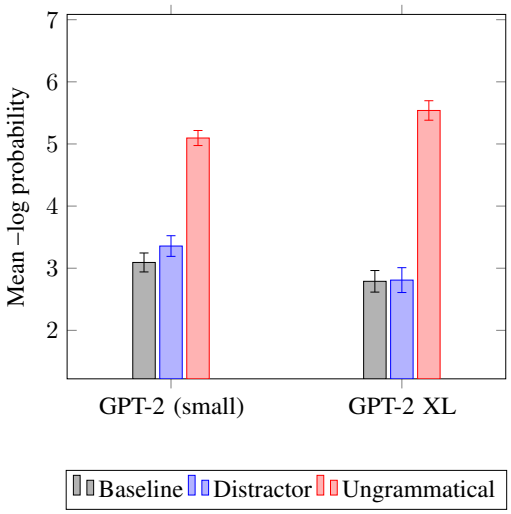

Figure 2: Mean negative log probability at the reflexive anaphor in transitive object control constructions. Error bars indicate 95% confidence intervals.

To further investigate the reason for the low performance on the agreement task for the transitive subject control constructions, we computed the mean surprisal values at the reflexive word *himself* for each of the 3 conditions. Figure 1 shows that, for the subject control constructions, both versions of GPT-2 have higher surprisal values in the 'Distractor' condition than in the 'Ungrammatical' condition, which provides additional evidence that the model adopts the strategy of agreeing with the closest NP. For object control constructions, on the other hand, both versions of GPT-2 show higher surprisal values in the 'Ungrammatical' condition than in the 'Distractor' condition (Figure 2), as already indicated by the near-perfect accuracy on the object control tasks. Moreover, we find that the surprisal of *himself* is almost identical in the conditions in which the object NP is singular ('Baseline' and 'Ungrammatical' for subject control constructions, and 'Baseline' and 'Distractor' for object control constructions), which further suggests that the model bases its predictions primarily on the number of the object NP in both types of constructions.

5 Discussion

The results from our experiments suggest that GPT-2 is unable to correctly distinguish subject control from object control constructions.[2] One potential strategy for increasing model accuracy is to augment the training data with examples of the form that we used for evaluation, which may lead models such as GPT-2 to learn the correct generalizations. However, while such a strategy may solve the problem for these specific constructions, the results that we presented here also highlight important limitations of training models from surface strings present in naturalistic corpora alone. This suggests that successfully mimicking human linguistic behavior may require a model that has access to meaning during training, as recently argued by Bender and Koller (2020), so that for example, it can learn the differences between subject and object control verbs (e.g., *promise* versus *persuade*).

Acknowledgements

We would like to thank Tal Linzen, Alec Marantz, and the anonymous reviewers for their thoughtful feedback. This material is based upon work supported by the National Science Foundation under Grant #2030859 to the Computing Research Association for the CIFellows Project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

[2] Some speakers of English also do not accept transitive subject control constructions (Courtenay, 1998). However, GPT-2 does not behave like this group of speakers either: If it did, it should assign similarly high surprisal values to all items with an object and a subject control verb, which is not what we observed in our experiments (see Figure 1).

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Karen Courtenay. 1998. Summary: Subject control verb PROMISE in English. https://linguistlist.org/issues/9/9-651/.

Mark Davies. 2008. The corpus of contemporary American English: 450 million words, 1990-present.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint*, abs/1901.05287.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020a. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York. Association for Computational Linguistics.

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

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.

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, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

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

A Stimuli

Noun phrases We manually constructed the following list of noun phrases: *the professor*, *the lawyer*, *the artist*, *the student*, and *the child*. The plural versions of the noun phrases were also used to generate grammatical and ungrammatical sentences. Each noun phrase is realized in the subject and object positions equally often in transitive sentences. Each noun phrase is realized with each of their matrix verbs equally often as well.

Matrix verbs The matrix verbs determine whether a given construction is subject or object control. For subject control verbs, we used *promise*, *offer*, and *guarantee*. For object control verbs, we used *persuade*, *tell*, and *force*.

Embedded clauses We manually constructed a list of non-finite embedded clauses hosting the reflexive anaphor *himself*: *to make fun of himself*, *to examine himself*, *to diagnose himself*, *to embarrass himself*, and *to disguise himself*. The embedded anaphor refers back to either the subject or the object depending on the matrix verb.