Andrew McInnerney – University of Michigan

Emergence of island effects in causal vs. masked language model training

Synopsis: We measured island effects in English (in terms of surprisal at "gap-requiring words"; see Wilcox et al. 2023) at different training checkpoints of two large language models with different training regimes. We found that island effects emerge more rapidly in training of RoBERTa (Liu et al. 2019, Warstadt et al. 2020), which uses a *masked* training regime, than in training of Pythia (Biderman et al. 2023), which uses a *causal* training regime.

Background: Filler-gap constructions (e.g. *wh*-fronting in (1)) are unified by their shared behavior with respect to various syntactic dimensions (Chomsky 1977). Of interest to this paper, filler-gap dependencies can be embedded up to an arbitrary depth (1b), yet they are sensitive to syntactic islands (1c). The contrast between island and embedded non-island extraction (e.g. (1b) vs. (1c)) has traditionally been a major motivation for the Argument from the Poverty of the Stimulus (APS) in generative grammar (see Hoekstra & Kooij 1988, Phillips 2013, Schütze et al. 2015, Adger 2020). However, Wilcox, Futrell & Levy (2023, henceforth WFL) recently argue that the behavior of Transformers on a variety of different types of syntactic island violations refutes the APS in this domain. Specifically, they found large "*wh*-effects" (WFL:11) in non-island contexts like (1a), which were significantly reduced in island contexts like (1c). As this contrast is consistent with accurate generalizations regarding islands, WFL conclude that Transformers set a lower bound for island acquisition via domain general mechanisms, i.e., for refutation of the APS based on islands. (Note: WFL also tested RNNs in support of this argument, but Chaves 2020 identifies a number of problems with this idea. We therefore focus on Transformers here.)

Experiments: We aimed to narrow WFL's in-principle lower bound by analyzing Transformer behavior on filler-gap dependencies across training. Because Transformers are trained on vastly more data than is psychologically realistic in the human learning case, refutation of poverty-of-stimulus arguments with Transformers depends on the extent to which behavior like that identified by WFL is contingent on training datasets of this size (Warstadt & Bowman 2023). Additionally, the crucial test of island knowledge comes not from the comparison between matrix and island extraction (e.g. (1a) vs. (1c)), but from the direct comparison of island and embedded non-island extraction (more like (1c) vs. (1b); see Kim & Goodall 2023). We therefore created 24 new item sets each for Adjunct Islands and Relative Clause Islands to directly make this comparison. An Adjunct Island example is given in (2), and an RC island example in (3).

Results: A model that has accurately learned island conditions should show large *wh*-effects in both the matrix and embedded non-island conditions and diminished *wh*-effects in the island conditions – there should be a three-way interaction between embed-type, gap-type, and filler-type. We found that, in the RoBERTa models, *wh*-effects emerged in the matrix conditions relatively early (~10M tokens of training; compare Zhang et al. 2020, Pérez-Mayos et al. 2021), but the critical interaction did not emerge until later (~30B tokens). In the Pythia models, *wh*-effects did not emerge in the matrix conditions until ~30B tokens, and the island vs. embedded non-island contrast never emerged. See plots below (using WFL's plotting conventions) for results with the Adjunct Island items.

Discussion: The RoBERTa models did achieve behavior consistent with accurate island knowledge, which contradicts a strong version of APS regarding islands: island constraints are not out of bounds for domain-general learning (to the extent Transformers qualify as such). However, the contrast between RoBERTa and Pythia are consistent with an APS which is relativized to the human learning environment (a common form of the APS, see e.g. Legate & Yang 2002, Berwick et al. 2011). First, the (more cognitively plausible) causal training regime of the Pythia models might be insufficient to achieve accurate island generalizations even after hundreds of billions of tokens of training. Further, even the comparatively information-rich masked training regime of RoBERTa requires orders of magnitude more training data than is available to humans. This broadly supports theories that attribute island effects to unlearned (syntactic or non-syntactic) mechanisms to some extent.

(1) a. [**Who** did you insult __ yesterday]?
b. [**Who** did you say [you think [… [it's clear [I insulted __ yesterday]] …]]?
c. *[**Who** do you work at [the place where I insulted __ yesterday]]?

(2) **Matrix** That's the man $\left\{\begin{matrix}\textit{to}\\ \textit{from}\end{matrix}\right\}$ whom the kids were $\left\{\begin{matrix}\textit{introduced}\\ \textit{protected}\end{matrix}\right\}$ yesterday.

NonIsland That's the man $\left\{\begin{matrix}\textit{to}\\ \textit{from}\end{matrix}\right\}$ whom you thought before [that the kids were $\left\{\begin{matrix}\textit{introduced}\\ \textit{protected}\end{matrix}\right\}$ yesterday].

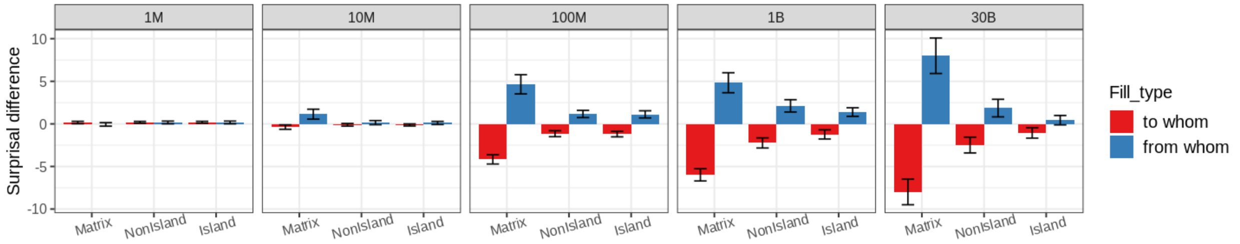
Island That's the man $\left\{\begin{matrix}\textit{to}\\ \textit{from}\end{matrix}\right\}$ whom you thought that [before the kids were $\left\{\begin{matrix}\textit{introduced}\\ \textit{protected}\end{matrix}\right\}$ yesterday].

(3) **Matrix** That's the man $\left\{\begin{matrix}\textit{to}\\ \textit{from}\end{matrix}\right\}$ whom the kids were $\left\{\begin{matrix}\textit{introduced}\\ \textit{protected}\end{matrix}\right\}$ yesterday.

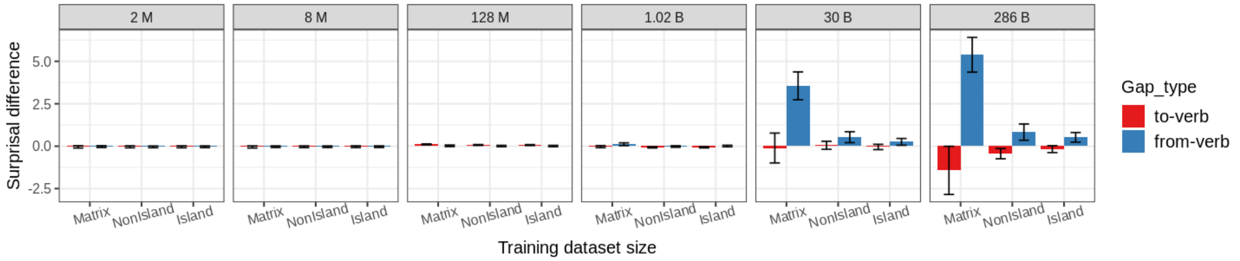
NonIsland That's the man $\left\{\begin{matrix}\textit{to}\\ \textit{from}\end{matrix}\right\}$ whom I told the boy [that the kids were $\left\{\begin{matrix}\textit{introduced}\\ \textit{protected}\end{matrix}\right\}$ yesterday].

Island That's the man $\left\{\begin{matrix}\textit{to}\\ \textit{from}\end{matrix}\right\}$ whom I know the boy [that the kids had $\left\{\begin{matrix}\textit{introduced}\\ \textit{protected}\end{matrix}\right\}$ yesterday].

Adjunct islands: RoBERTa wh-effects (filler region) by tokens of training data



Adjunct islands: Pythia 2.8b wh-effects (gap region) by tokens of training data



Abridged References: Adger 2020. Syntax and the failure of analogical generalisation: A commentary on Ambridge (2020). || Biderman et al. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. || Chomsky 1977. On wh-movement. || Kim & Goodall 2023. The island/non-island distinction in long-distance extraction: Evidence from L2 acceptability. || Liu et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach || Pérez-Mayos et al. 2021. How much pretraining data do language models need to learn syntax? || Phillips 2013. On the nature of island constraints. II: Language learning and innateness. || Warstadt et al. 2020. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). || Warstadt & Bowman 2023. What Artificial Neural Networks Can Tell Us About Human Language Acquisition. || Wilcox et al. 2023. Using Computational Models to Test Syntactic Learnability. || Zhang et al. 2020. When Do You Need Billions of Words of Pretraining Data?