Word-Order Error Detection Helps Data-Efficient Language Models Learn Syntax

Alexander Fung* (alexfung@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139

Chengxu Zhuang* (chengxuz@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139

Steven T. Piantadosi (spiantado@gmail.com)
Departments of Psychology and Neuroscience, UC Berkeley
Berkeley, CA 94720

Jacob Andreas (jda@mit.edu)
Department of Electrical Engineering and Computer Science, MIT
Cambridge, MA 02139

Evelina Fedorenko (evelina9@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139

Neural language models (LMs) require vast amounts of data to master syntax—a set of rules for how word arrangements create complex meanings. In contrast, children learn efficiently from a small amount of linguistic input. Inspired by findings of early sensitivity to word order information in children, we here augment LM training with a novel objective that emphasizes word order, in an attempt to minimize the data efficiency gap between LMs and humans. The new objective requires discriminating between grammatical sentences and sentences with word-order perturbations. After training LMs on developmentally plausible amounts of data, we find that LMs with this augmented training outperform control LMs (trained on standard masked language modeling) on select components of an established benchmark of syntactic knowledge (BLiMP) and a new benchmark we developed that targets word-order error detection. These results suggest that integrating synthetic tasks can effectively reduce the data efficiency gap between neural LMs and human learners.

Keywords: syntax learning; neural language models; word order; masked language modeling; data augmentation.

Introduction

Within the first few years of life, children learn how words go together in a language, which allows them to decode complex meanings from others' productions and express their own ideas through language (MacWhinney 2013). Neural language models (LMs) also acquire syntactic knowledge but require massively more linguistic data to do so (Alex Warstadt et al. 2023; A. Warstadt and Bowman 2022). Of course, children and LMs learn language in different ways. One important difference is that children are exposed to grammatically ill-formed utterances, which get corrected; this includes i) their own early productions, which commonly use incorrect word orders, and which adults often correct (Saxton, Backley, and Gallaway 2005; Clark 2020); and ii) word-order errors in adults' productions, which typically get self-corrected (Levelt 1983; Roelofs 2020; Chouinard and Clark 2003). Aiming to similarly emphasize word-order correctness, we supplement LM training with a learning objective whose goal is to differentiate between grammatical sentences and sentences with word-order errors. We train LMs on developmentally plausible amounts of data and then evaluate them on the standard BLiMP benchmark (Alex Warstadt et al. 2019) and two new benchmarks: word-order error detection and masked function word prediction. The new benchmarks test core syntactic knowledge in a theory-neutral way (cf. BLiMP, which targets specific phenomena, many of which are relevant to a particular theory of grammar) and have the additional advantage of using naturalistic materials constructed from the validation and test sets of the same corpus as the training sentences.

Methods

Model training

The network architecture of our models is an encoder-only 12-layer transformer (Vaswani et al. 2017). We train models on 5M, 15M, and 50M tokens using either masked language modeling (MLM) on fixed length (128 tokens) input ("Baseline models" in Fig. 1) or the combination of MLM and the new word-order emphasizing objective ("Word-Order models" in Fig. 1). Specifically, we prepend a [CLS] token to the 128-token input, then map the 5th hidden state of the model output to a binary correct/incorrect order label with a 3-layer perceptron of hidden size 768. The training objectives use the same dataset, but with independently drawn batches, and loss is summed across the objectives. Our MLM objective follows the training paradigm of RoBERTa (Liu et al. 2019). Our training data is sampled from Smashwords (Alex Warstadt et al. 2020).

Benchmarks

Masked Function Word Prediction In this benchmark, 128-token chunks are sampled from the test set of the training corpus, and the models are asked to predict masked function word tokens, as identified by Stanza (Qi et al. 2020).

Word-Order Error Detection This benchmark involves discriminating between individual sentences from the test set of the training corpus and incorrect word-order versions of each sentence generated by randomly shuffling two consecutive words. For Word-Order models, we use the word-order readout head to perform this evaluation. We also perform a surprisal-based evaluation for both model types: first, we find the

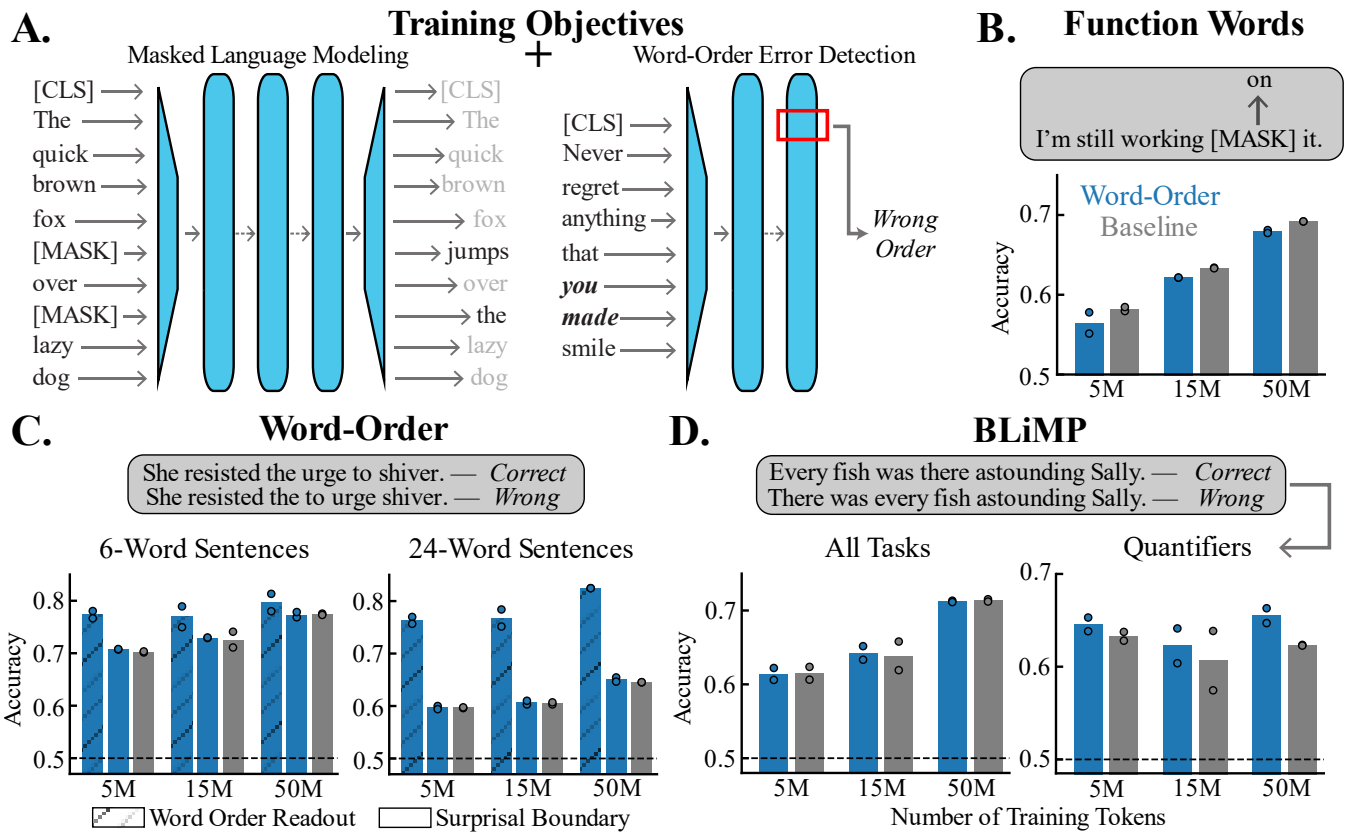

Figure 1. Training on the word-order emphasizing objective improves some but not all evaluation benchmarks. A. Models are trained on both a masked language modeling objective and a word-order objective in alternating batches. **B.** Predicting masked function words. Two models starting from different seeds are trained for each model class and each size of the training dataset. The dots represent their individual performances and the bar height is their average. **C.** Performance on the word-order benchmark for 6-word (left panel) and 24-word sentences (right panel). **D.** Mean performance on all 67 BLiMP tasks (left panel) and the 4 "Quantifiers" tasks (right panel).

best discriminating sum-surprisal threshold for the validation set, then apply this threshold to the test set.

BLiMP The BLiMP benchmark (Alex Warstadt et al. 2019) is a suite of 67 linguistic phenomena across 13 categories, spanning morphology, syntax, and semantics. The benchmark uses the minimal-pair paradigm from linguistics, where each pair consists of a grammatically correct sentence and a minimally different but incorrect one.

Results

Co-training on the word-order-targeting objective improves performance on the word-order benchmark: Word-Order models outperformed Baseline models when using the word-order readout head, while performance approximately matched Baseline models using the sum-surprisal method. On BLiMP, overall performance is similar, but Word-Order models improve on the Quantifiers task suite, for which word-order information is critical (Fig. 1D). Meanwhile, accuracy on the function word benchmark decreased slightly. In summary, our augmented (Word-Order) models improve on the word-order benchmark and some BLiMP tasks, showing the benefit of a mixed objective that emphasizes word-order information. However, Word-Order models' similar or lower performance on other benchmarks suggests that emphasis on word order does not generalize to all aspects of syntax learning, and additional or different inductive biases are needed to fully bridge the data-efficiency gap between LMs and children.

References

Chouinard, Michelle M., and Eve V. Clark. 2003. "Adult Reformulations of Child Errors as Negative Evidence." *Journal of Child Language* 30 (3): 637–69.

Clark, Eve V. 2020. "Conversational Repair and the Acquisition of Language." *Discourse Processes* 57 (5-6): 441–59.

Levelt, W. J. 1983. "Monitoring and Self-Repair in Speech." *Cognition* 14 (1): 41–104.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1907.11692.

MacWhinney, Brian. 2013. *The Emergence of Language*. Psychology Press.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2003.07082.

Roelofs, Ardi. 2020. "Self-Monitoring in Speaking: In Defense of a Comprehension-Based Account." *Journal of Cognition* 3 (1): 18.

Saxton, Matthew, Phillip Backley, and Clare Gallaway. 2005. "Negative Input for Grammatical Errors: Effects after a Lag of 12 Weeks." *Journal of Child Language* 32 (3): 643–72.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/7181-attention-is-all.

Warstadt, A., and S. R. Bowman. 2022. "What Artificial Neural Networks Can Tell Us about Human Language Acquisition." *Algebraic Structures in Natural*. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003205388-2/artificial-neural-networks-tell-us-human-language-acquisition-alex-warstadt-samuel-bowman.

Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, et al. 2023. "Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora." In . https://doi.org/10.18653/v1/2023.conll-babylm.1.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. "BLiMP: The Benchmark of Linguistic Minimal Pairs for English." *Transactions of the Association for Computational Linguistics* 8 (December): 377–92.

Warstadt, Alex, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R. Bowman. 2020. "Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually)." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2010.05358.