Dependency-Based and Constituency-Based Inductive Biases for BabyLM Challenge

GEORGETOWN UNIVERSITY

Zhuoxuan Ju, Lanni Bu, Dagny Whall, Vattana Chan

Introduction

Inspired by Papadimitriou and Jurafsky (2023), which investigates how two types of inductive biases — nested and crossing structures — influence language models, we aim to further explore the impact of constituency structures and dependency structures on language model performance. Constituency and dependency structures are widely used representations in linguistics, respectively representing purely nested structures and a combination of nested and crossing structures. Our research aims to leverage these structures, observed on real human language data, to influence language model behavior and potentially hope to evaluate the pretrained inductive biases by fine-tuning it on the BabyLM dataset and examine how our language model performs using a variety of evaluation metrics.

Methods

The core idea of our research is to inject purely structural data during pretraining to observe whether language models behave differently depending on the type of structure they are exposed to. Specifically, we pretrain models on data containing only structural information, then finetune them on the BabyLM 100M dataset.

Pretraining Data

We utilize five types of pre-training data with each training split containing 1 billion tokens:

- **Constituency structure**: We retain only brackets and replace each matched bracket pair with the same randomly assigned number.

(ROOT (S (NP (PRP She)) (VP (VBD gave) (NP (PRP me)) (NP (DT the) (NN book))) (. .)))

(((( )) (( ) (( )) (( ) ( )) (( ))))

96 14 33 15 15 33 71 56 56 6 82 82 6 40 38 38 370 370 40 71 153 153 14 96

Figure 1: Example of Transforming Constituency Parser Output into Pretraining Data.

- **Dependency structure**: We use paired numbers to represent dependency relations. Nested pairs indicate projective structures, while crossing pairs represent non-projective structures.

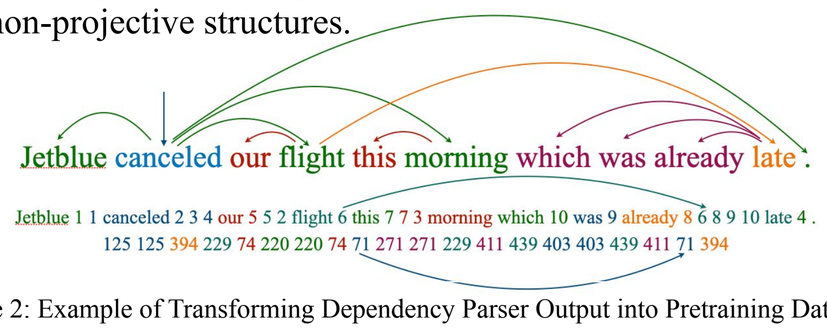

Figure 2: Example of Transforming Dependency Parser Output into Pretraining Data.

- **Nested Data**: A purely nested structure that randomly decides when to open and close brackets.
- **Mixed Data**: Based on the nested structure, we then insert 0.5% crossing pairs to simulate a slight non-projectivity. This insertion rate is chosen based on our estimation that dependency structures in real language data contain approximately over 0.4% crossing dependencies.
- **Random Data**: Contains completely random numbers without any underlying structure.

Finetuning Data

We finetune the pretrained models on the BabyLM 100M dataset. The BabyLM challenge provides a 100M text only dataset, which contains higher proportion child and child-directed speech.

Implementation Details

- **Model**: GPT-2 Small
- **Pretraining**:
  - Vocabulary size: 500 tokens
  - Batch size: 512
  - Training steps: 5,000 (whole dataset is seen approximately 1.5 times)
- **Finetuning**:
  - Batch size: 128
  - Total steps: 3,800 (5 epochs)
  - To accommodate GPT-2 Small's original 50,257-token vocabulary, we initialize the new embedding matrix by randomly sampling (with replacement) rows from the original smaller (500-row) embedding matrix.
- **Evaluation Metrics**:
  We evaluated our results across the five models using perplexity, BLiMP, BLiMP Supplement, a subset of (Super) GLUE tasks, and EWoK on the shared evaluation pipeline provided by the BabyLM Challenge. To clearly observe differences, we select model checkpoints after 1,500 steps of fine-tuning for evaluation.
  - **Perplexity** is a measurement of a language model's uncertainty when predicting the next word in a sequence of words.
  - **BLiMP** is commonly used to evaluate the sensitivity of language models to a variety of different aspects of English phenomena. The aspects range from general to niche (Warstadt *et al.*, 2020).
  - **BLiMP supplement** is used to cover additional aspects of English phenomena that are not used on BLiMP. The phenomena center around dialogue and questions (Warstadt *et al.*, 2023).
  - The subset of (Super) **GLUE** tasks measure natural language understanding in tasks such as question answering and natural language inference (Wang *et al.*, 2019, 2018a).
  - **EWoK** is used to analyze the model's contextual performance by measuring pragmatic, commonsense, and discourse knowledge (Ivanova *et al.*, 2024).

Results

- Models trained on all structured data outperform the the random baseline.
- Models trained on purely nested structures achieve higher perplexity compared to those trained with crossing structures.
- Models trained on dependency structures achieve higher performance compared to those trained on constituency structures.

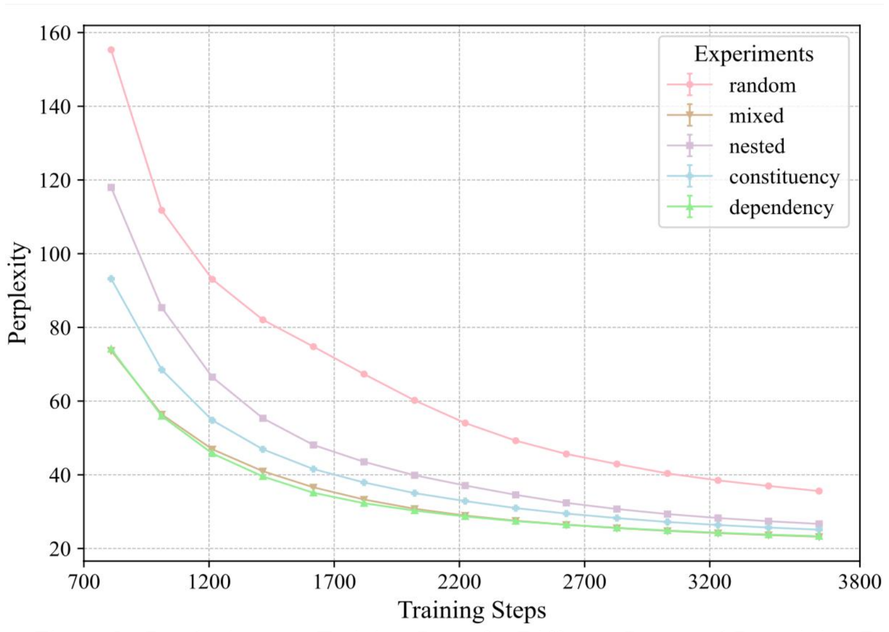

Figure 3 : Perplexity over Training Steps for Different Structural Pretraining Data

Model	BLiMP ↑	BLiMP-S ↑	GLUE ↑	EWOK ↑
Random	63.3	53.5	62.3	49.0
Nested	66.3	55.1	65.7	49.9
Constituency	65.0	52.9	64.5	49.7
Mixed	67.6	**56.8**	65.7	49.5
Dependency	**69.4**	56.62	**66.0**	**50.35**

Table 1 : Performance on Downstream Tasks by Pretraining Dataset

- Nested data outperforms constituency data on all evaluation metrics, even though its perplexity is slightly higher.
- Dependency data outperforms mixed data on three out of four metrics.

Discussion

Although the datasets encode similar biases (dependency vs cross; constituency vs nested), differences in their statistical properties may lead to different training outcomes, eg. tree depth, number of leaf nodes, average branching factor, symmetry between left and right subtrees.

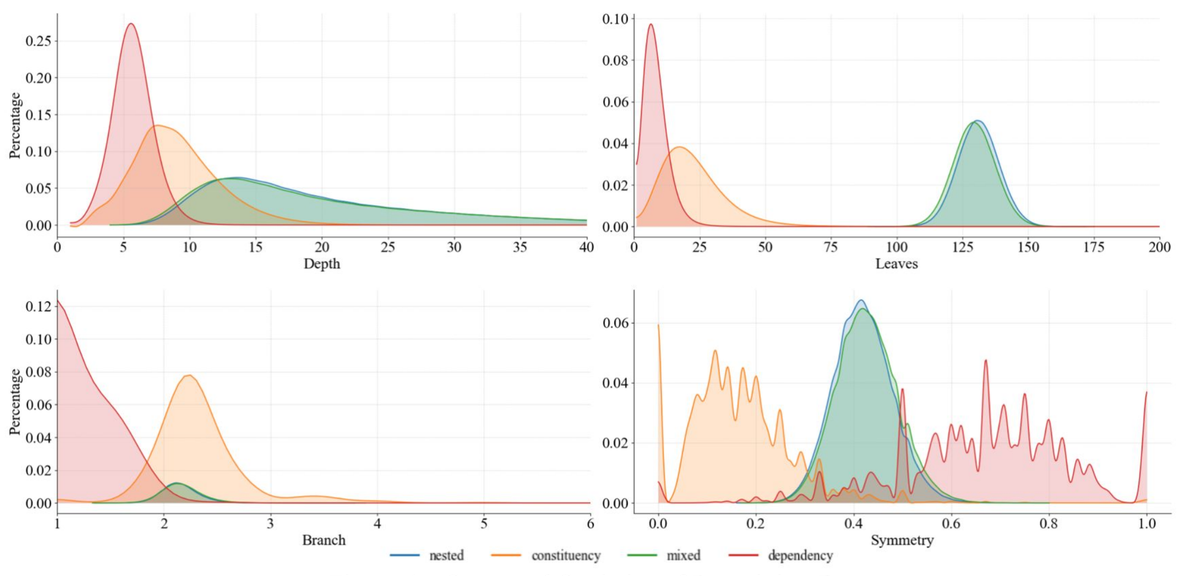

Figure 4 : Structural Statistics of Pretraining Datasets

We can see from these figures that nested vs constituency structures, mixed vs dependency show different structural patterns.

Limitations/Future Work

- We ignore a crucial aspect of dependency trees: their hierarchical structure. Future work should investigate methods to better encode hierarchy in pretraining datasets.
- We did not train long enough to determine whether perplexity fully converges. It would be interesting to finetune models on an even smaller dataset (such as BabyLM 10M) to examine whether convergence patterns remain.
- While we computed several structural statistics to explain the observed performance differences, they may not fully account for all variations. Further experiments and exploration of additional features are necessary.

References