BabyLM Challenge

The Effects of Data Formatting and Structural Biases

Ziling Cheng,[1,2] Rahul Aralikatte,[1,2] Ian Porada,[1,2] Cesare Spinoso-Di Piano,[1,2] Jackie Chi Kit Cheung[1,2,3]

[1]McGill University, [2]Mila – Quebec AI Institute, [3]Canada CIFAR AI Chair

MOTIVATION

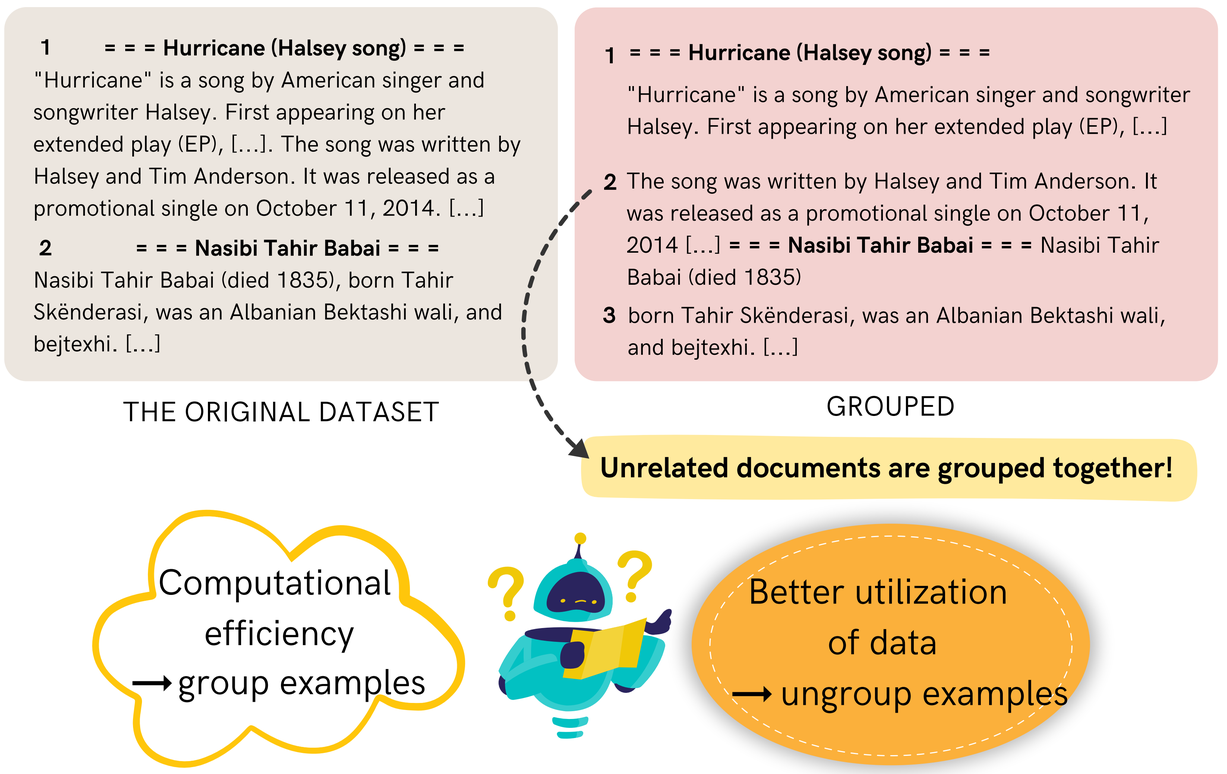
Objective: sample-efficient pretraining: optimize pretraining from scratch with the same amount of linguistic data available to a child (~10M tokens)

- **On the data side:**
  - Current pretraining practice groups unrelated documents into the same training mini-batch.
  - Can we have an improved mini-batch format?
- **On the modelling side:**
  - Currently, only textual input is supplied to the model.
  - Does enriching the training signal by inducing syntactically-informed inductive biases yield benefits?

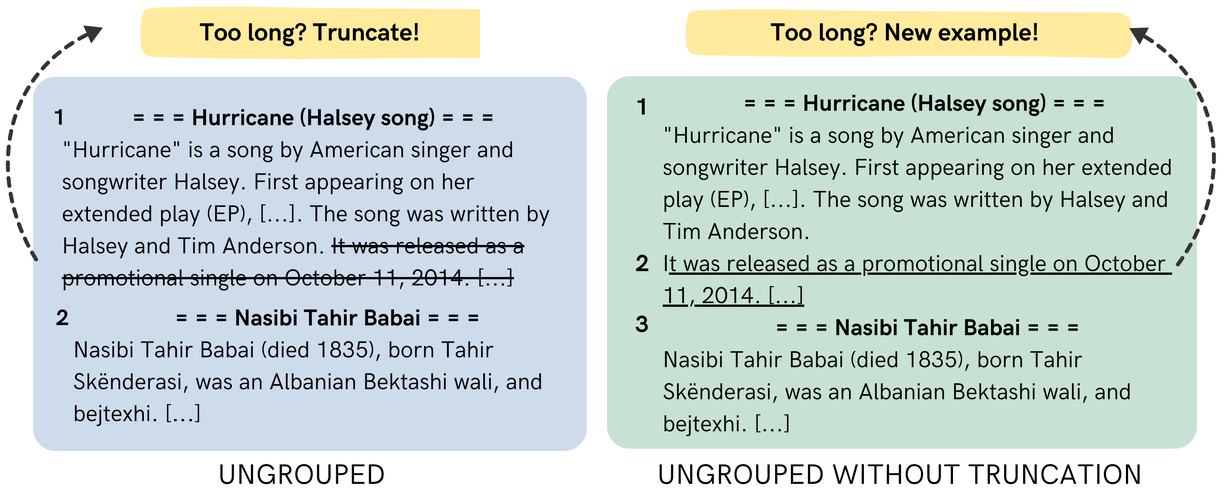
METHODOLOGY

1. Data Formatting

- Current pretraining regime: grouping

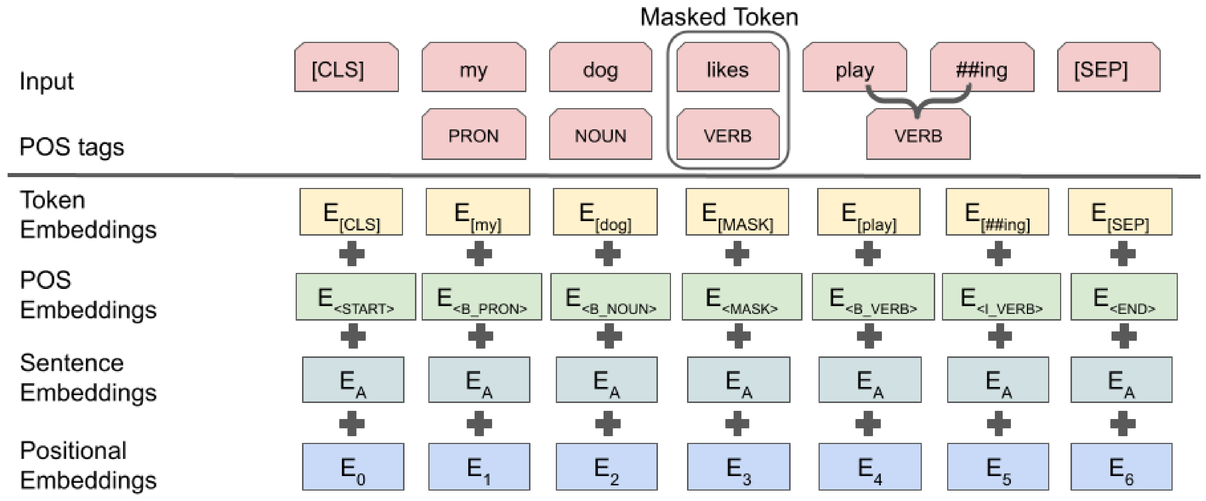


- Proposed method: ungrouping
  - Keep document boundaries



2. Structurally Biased Pretraining

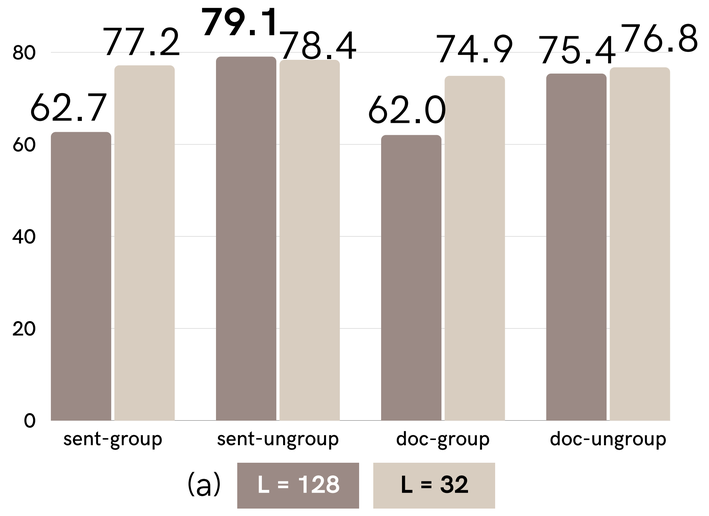
POS Augmentation: augmenting the richness of training signals by inducing syntactically-motivated inductive biases



RESULTS

(a) Effects of Data Formatting:

- The ungrouping strategy demonstrates superior effectiveness on both dataset formats.



Model \ H	768	1152
Vanilla	77.3	77.4
RandPOS	73.9	74.3
POS	**78.0**	**78.9**

(b)

Fig 1: The average performance of BERT base models on BLiMP tasks (accuracy). L refers to the maximum sequence length.

(b) Effects of POS Augmentation:

- Augmenting the training signal with POS does not yield significant benefits.

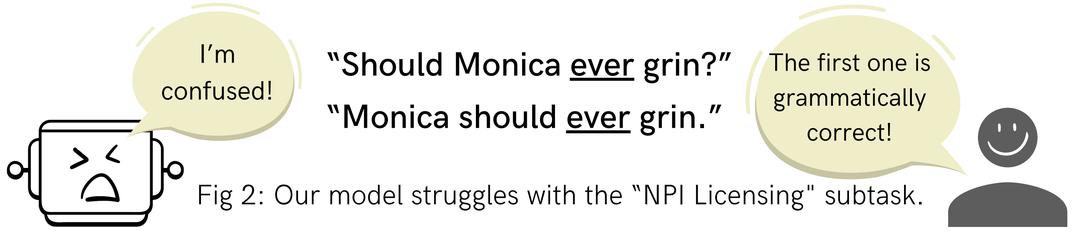


Fig 2: Our model struggles with the "NPI Licensing" subtask.

CONCLUSION

- The formatting of inputs can significantly impact downstream performance.
- Inducing structural biases in the model through part-of-speech trees yields modest benefits.