

BabyLM Challenge



The Effects of Data Formatting and Structural Biases

Ziling Cheng^{1,2}, Rahul Aralikatte, Ian Porada, Cesare Spinoso-Di Piano, Jackie Chi Kit Cheung McGill University, ²Mila – Quebec AI Institute, ³Canada 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

- 1 = = = Hurricane (Halsey song) = = = "Hurricane" is a song by American singer and songwriter Halsey. First appearing on her extended play (EP), [...]. The song was written by Halsey and Tim Anderson. It was released as a promotional single on October 11, 2014. [...]
- = = = Nasibi Tahir Babai = = = Nasibi Tahir Babai (died 1835), born Tahir Skënderasi, was an Albanian Bektashi wali, and bejtexhi. [...]

THE ORIGINAL DATASET

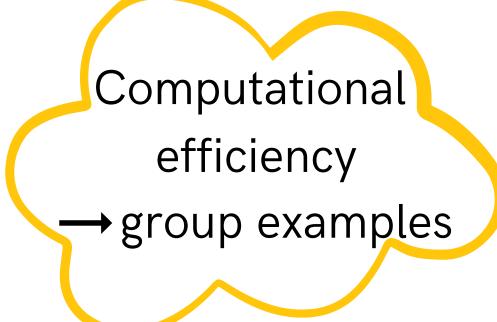
1 = = = Hurricane (Halsey song) = = =

"Hurricane" is a song by American singer and songwriter Halsey. First appearing on her extended play (EP), [...]

- 2 The song was written by Halsey and Tim Anderson. It was released as a promotional single on October 11, 2014 [...] = = = Nasibi Tahir Babai = = = Nasibi Tahir Babai (died 1835)
- 3 born Tahir Skënderasi, was an Albanian Bektashi wali, and bejtexhi. [...]

GROUPED

Unrelated documents are grouped together!





Better utilization of data ungroup examples

- Proposed method: ungrouping
 - Keep document boundaries

Too long? Truncate!

= = = Hurricane (Halsey song) = = = "Hurricane" is a song by American singer and songwriter Halsey. First appearing on her extended play (EP), [...]. The song was written by Halsey and Tim Anderson. It was released as a promotional single on October 11, 2014. [...]

= = = Nasibi Tahir Babai = = = Nasibi Tahir Babai (died 1835), born Tahir Skënderasi, was an Albanian Bektashi wali, and bejtexhi. [...]

UNGROUPED

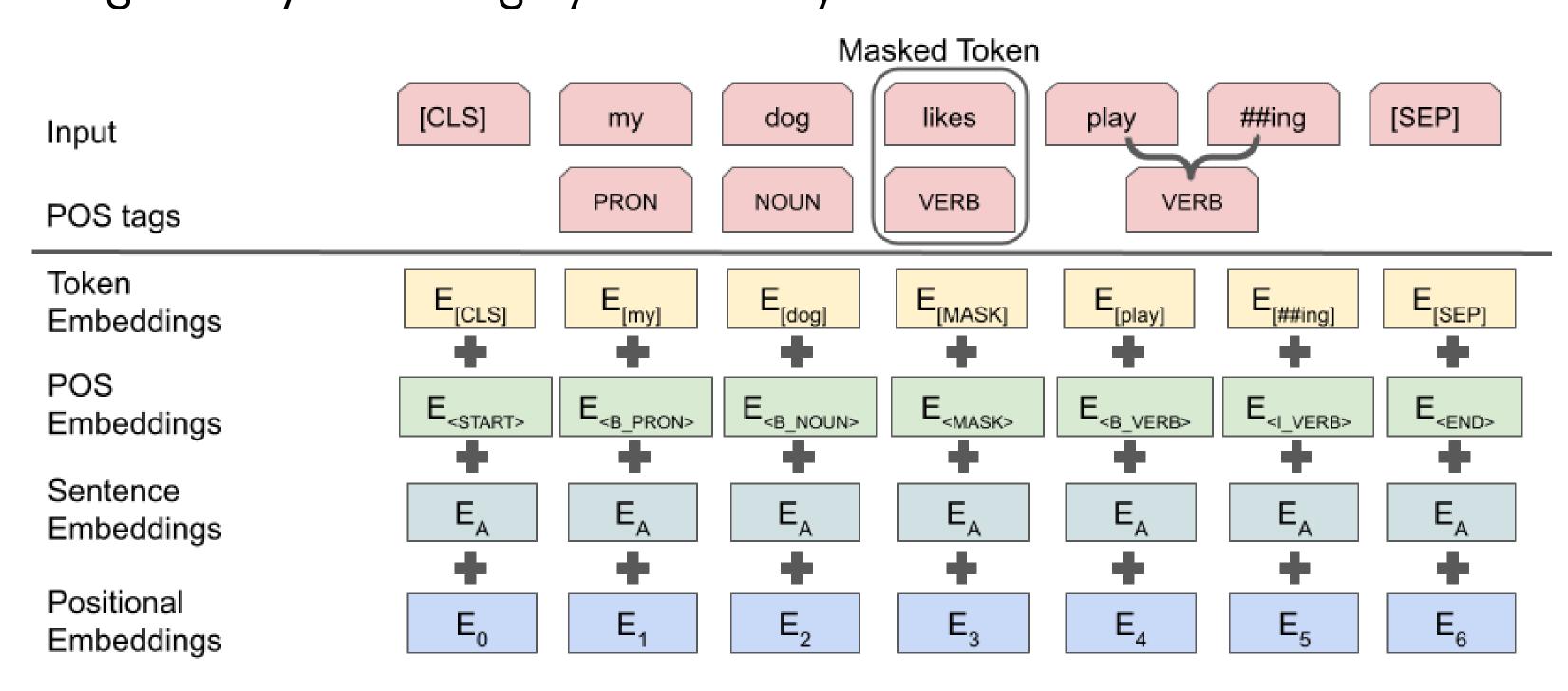
Too long? New example!

- = = = Hurricane (Halsey song) = = = "Hurricane" is a song by American singer and songwriter Halsey. First appearing on her extended play (EP), [...]. The song was written by Halsey and Tim Anderson.
- 2 It was released as a promotional single on October 11, 2014. [...]
- = = = Nasibi Tahir Babai = = = Nasibi Tahir Babai (died 1835), born Tahir Skënderasi, was an Albanian Bektashi wali, and bejtexhi. [...]

UNGROUPED WITHOUT TRUNCATION

2. Structurally Biased Pretraining

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



(a) Effects of Data Formatting:

 The ungrouping strategy demonstrates superior effectiveness on both dataset formats.

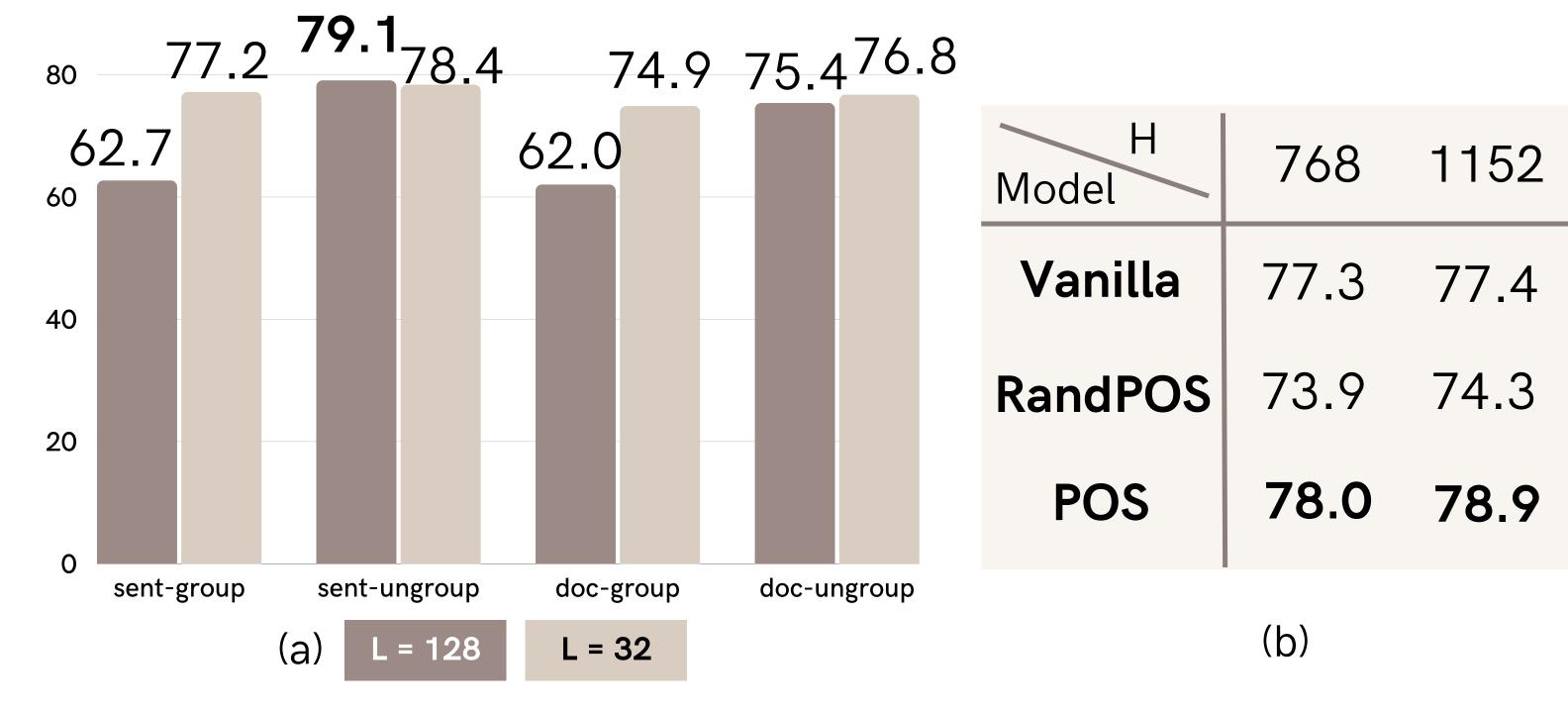
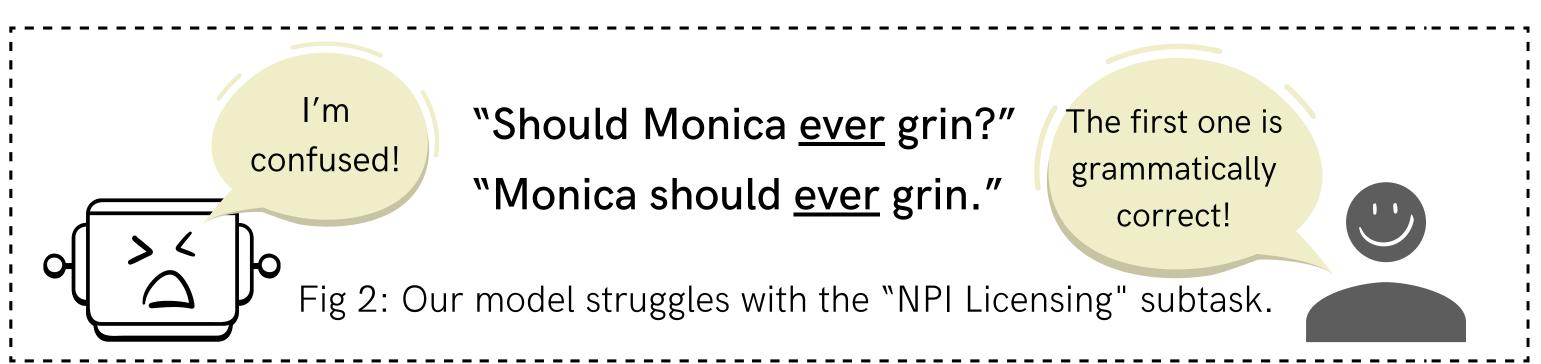


Fig 1: The average performance of BERT base models on BLiMP tasks (accuracy). L refers to the maximum sequence length.

(b) Effects of POS Augmentation:

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



CONCLUSION

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



