Subject-verb agreement with Seq2Seq transformers
Bigger is better, but still not best

Michael Wilson, Zhenghao Zhou & Robert Frank

Yale

Background

- Neural network language models show mixed behavior on subject-verb agreement (Linzen et al. 2016; Goldberg, 2019; Newman et al., 2021). In some but not all cases, errors reflect human performance (Arehalli & Linzen, 2020, 2022).
- Our task: tense reinflection (McCoy et al., 2020).
  - Source: The professor liked the dean. PRES:
  - Target: The professor likes the dean.
- Non-pre-trained (recurrent and transformer) models do poorly on this task (McCoy et al., 2020; Petty & Frank, 2021).
- However, pre-training transformer Seq2Seq (T5) models on unstructured text helps considerably with passivization and question formation (Mueller et al., 2022). Could it help with tense reinflection/agreement?
- How do model and dataset size affect agreement performance, measured by accuracy and similarity to human performance?

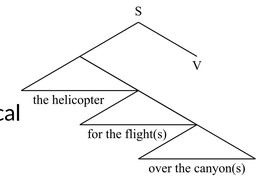
Materials

- Fine-tuning dataset: 1098 examples constructed from English Wikipedia by taking a {past, present} sentence and converting to {present, past}.
- Test dataset: balanced synthetic dataset constructed using a PCFG

Pre-verb noun(s)	Structures	Number
S, P	–	64 ea.
SS, SP; PP, PS	PP; RC	256 ea.
SSS, SSP, SPS, SPP; PPP, PPS, PSP, PSS	PP+PP, PP+RC, RC+PP, RC+RC	256 ea.

Evaluation

- Accuracy: does the model produce the correct form?
- Attraction effects (Bock & Cutting, 1992; Franck et al., 2002) Does the model...
  - ... show more errors in SP than in PS?
  - ... show more errors in PP than in RC?
  - ... show more errors with structurally local (SPS) than linearly local (SSP) distractors?



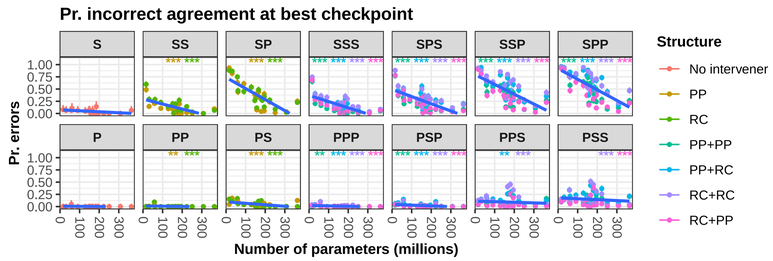
Summary of Results

- **Bigger models show higher accuracy**, especially in singular-subject conditions.
- **More training data generally produces higher accuracy**, but **more simple data yields worse performance** in some singular-subject conditions.
- Most models **replicate the singular-plural asymmetry**.
- Almost all models **fail to replicate the PP-RC asymmetry**, and **none replicate the structural-linear distractor asymmetries**.

Accuracy

Effects of model size

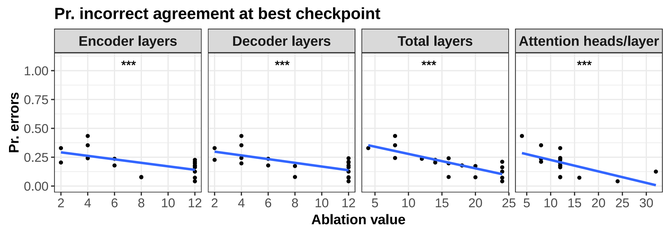
- Test on T5 Efficient ablations (Tay et al. 2021), differing in number of parameters (number of encoder and decoder layers, number of attention heads).



- More parameters yield better performance.

Effects of model architecture

- Test on T5 Efficient models, broken down by location of ablation.

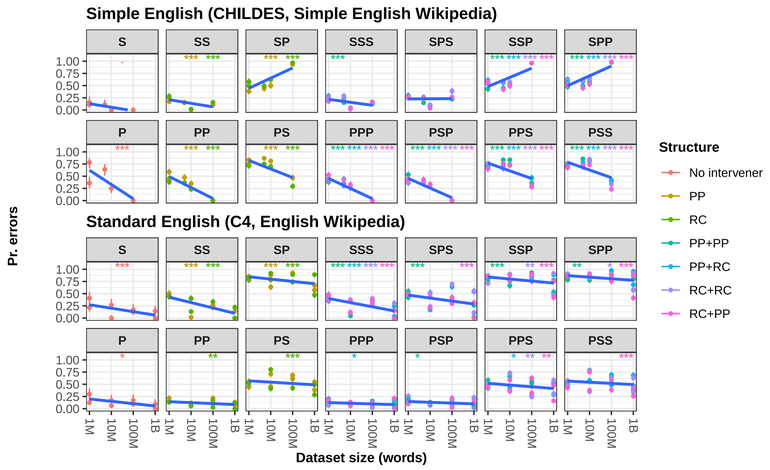


- All aspects of the architecture affect performance, though encoder layers have more impact than decoder layers:

$$\text{EL} - \text{DL}\ \beta = -0.00593,\ z = -2.87,\ p = 0.00415$$

Effects of dataset type and size

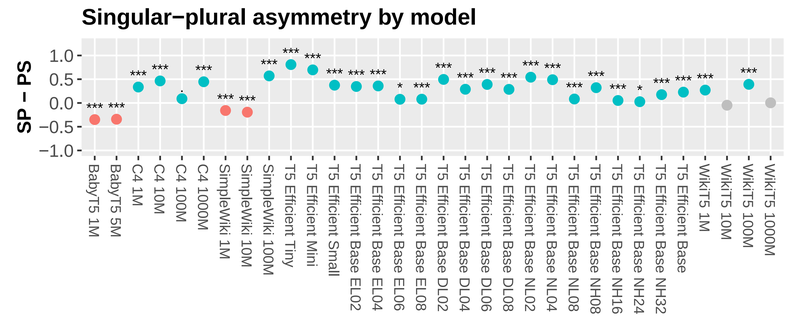
- Test on T5 models trained on CHILDES (1M, 5M), Simple English Wikipedia (1M, 10M, 100M), English Wikipedia (1M, 10M, 100M, 1B), and C4 (1M, 10M, 100M, 1B) (A. Mueller, p.c.)



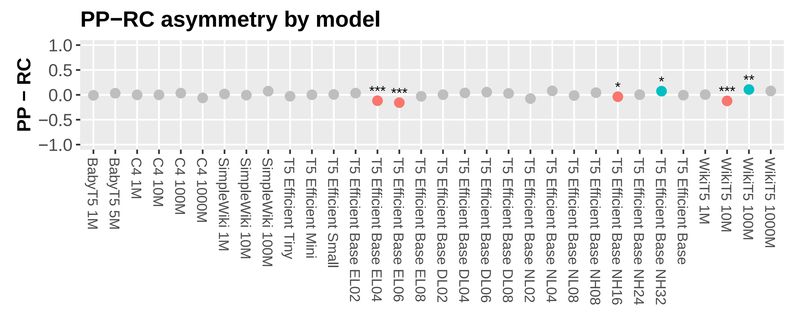
- More data yields better performance overall; for simpler datasets it yields worse performance in some singular-subject conditions.

Attraction effects

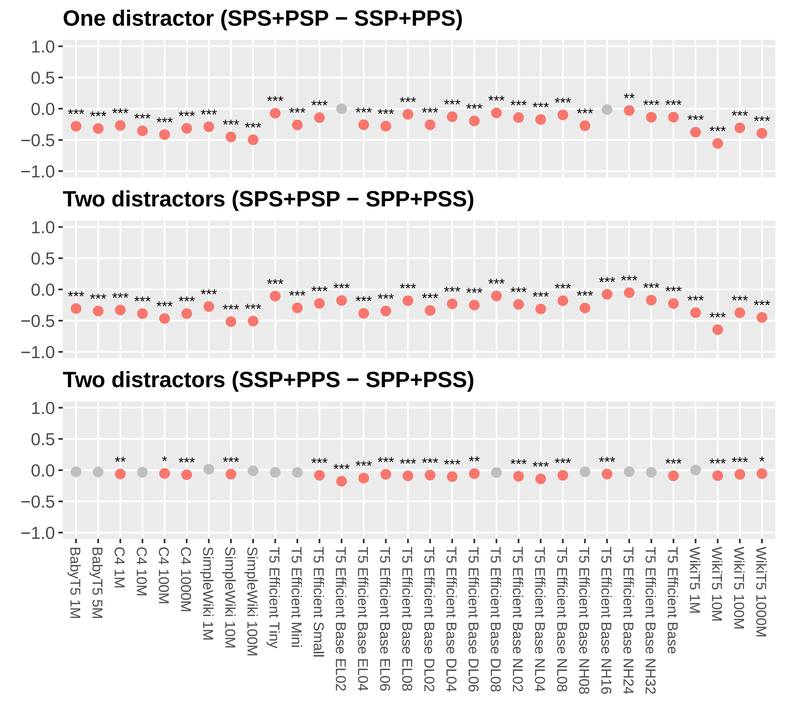
Singular-plural asymmetry ✓



PP-RC asymmetry ✗



Structural vs. linear effects ✗



This work was made possible by support from National Science Foundation grant BCS-1919321.

{michael.a.wilson, herbert.zhou, robert.frank}@yale.edu