Rethinking the Role of Text Complexity in Language Model Pretraining

Dan John Velasco * and Matthew Theodore Roque *

Samsung R&D Institute Philippines
{dj.velasco,roque.mt}@samsung.com

* Equal Contribution

Compare these texts:

(A) As the sunset cast its warm orange glow over Manila Bay, people relaxed on the sideline benches, enjoying the peaceful view of the sunset.

(B) The sunset gave Manila Bay a warm, orange light. People sat on the benches and enjoyed the view of the sunset.

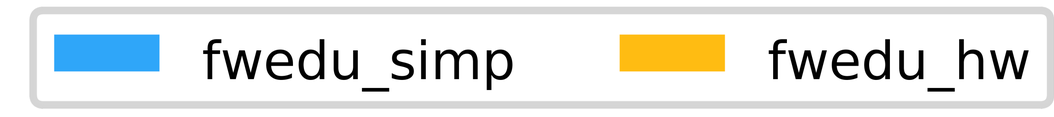
Same meaning, different text complexity.

What if our corpus is more like (B)? Can we still learn useful representations by training solely on simplified text with a simpler vocabulary and sentence structure?

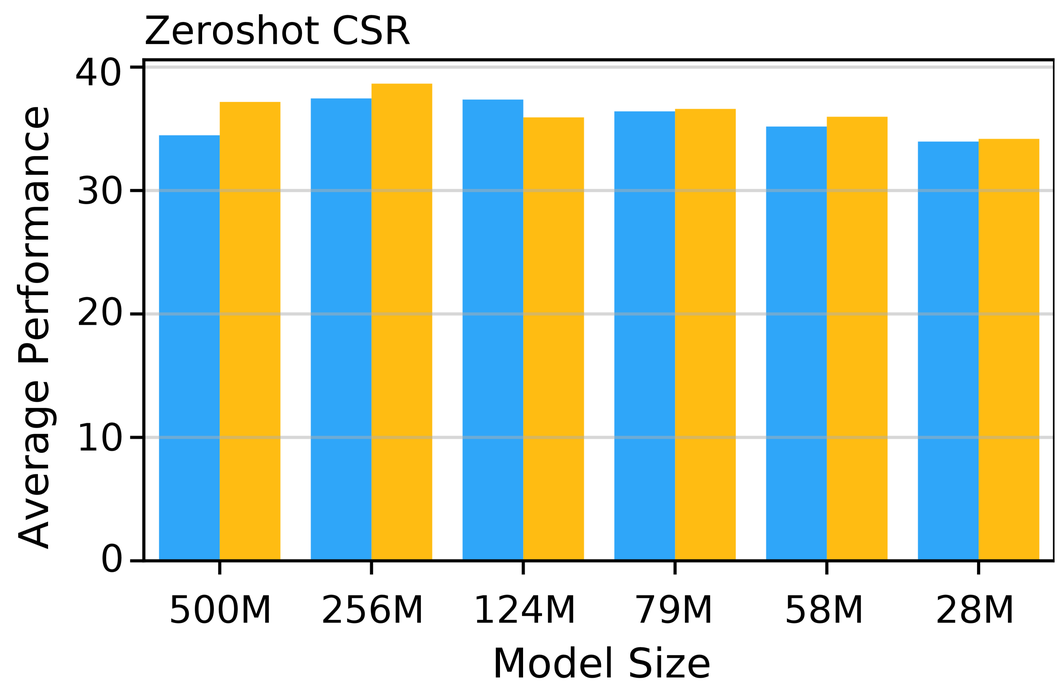
TLDR: Yes!

Results #1: fwedu_simp (simplified) performs better on blimp-supp. fwedu_hw (complex) is better on ewok (a bit) and entity tracking (a lot)

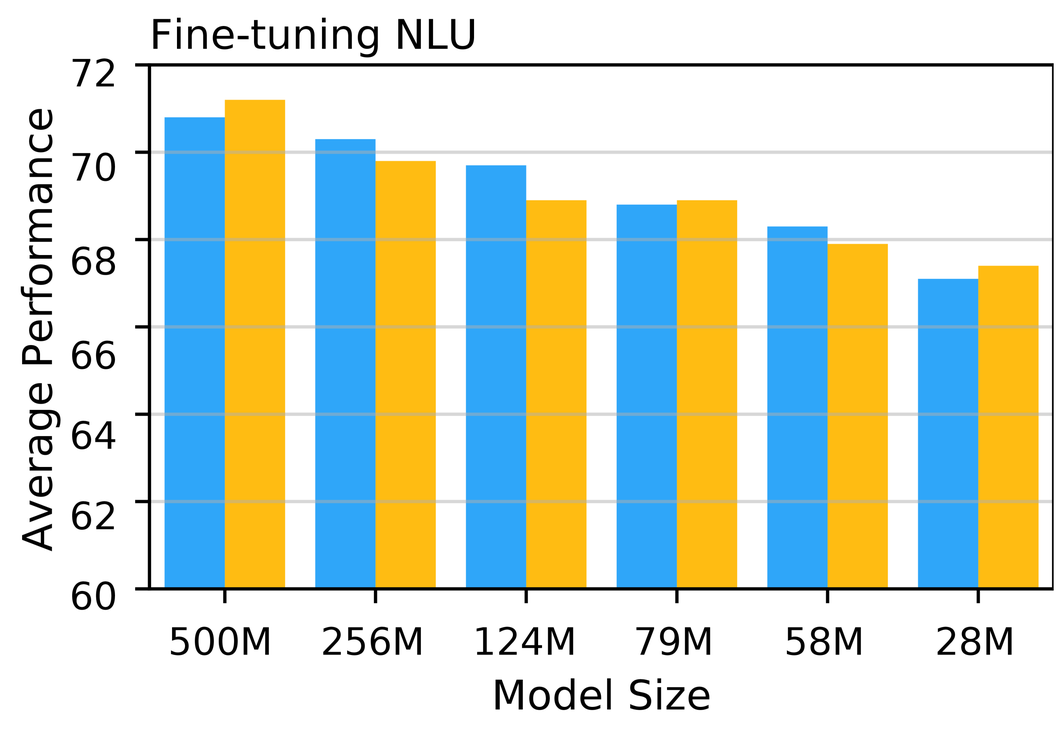
Model	blimp	blimp-supp	ewok	entity
28M				
fwedu_hw	**67.83**	55.90	**52.79**	16.15
fwedu_simp	66.90	**57.47**	52.67	**18.78**
58M				
fwedu_hw	69.69	59.03	52.69	**26.35**
fwedu_simp	**70.73**	**62.15**	**53.81**	18.36
79M				
fwedu_hw	**70.67**	60.47	**54.09**	**28.89**
fwedu_simp	70.44	**61.55**	53.09	20.73
124M				
fwedu_hw	**71.64**	62.61	**54.07**	**25.09**
fwedu_simp	71.30	**63.27**	54.01	21.79
256M				
fwedu_hw	72.37	62.61	**56.18**	**29.71**
fwedu_simp	**72.53**	**63.65**	55.09	22.58
500M				
fwedu_hw	72.23	61.85	**56.72**	20.81
fwedu_simp	**72.60**	**63.79**	54.85	**22.05**



Results #2: fwedu_hw (complex) has slight advantage on zero-shot commonsense reasoning tasks



Results #3: text complexity has little effect on fine-tuning



For those curious about the data: