# ACL Paper Summary

## CS4395.001 Human Language Technologies

**Brenden Healey**
**BMH180001**

**Quy Giang**
**QTG190000**

## Introduction

The research paper *When Do You Need Billions of Words of Pretraining Data?*, authored by Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman, details how Natural Language Processing is dominated by pretrained language models that are trained on billions of words. The average human learns language and grammar using only 10 to 100 million words, so how does a language model benefit from being trained on 10 to 100 times that amount of words?

## Prior Work

Probing neural network models has been an active area of research for the past several years. Probing of large transformer language models, such as BERT and its derivatives, has been used to locate linguistic features with great results. Studies from Van Schijndel et al. (2019) found large improvements in knowledge of subject-verb agreement and reflexive binding up to 10M words, though there was little improvement between 10M and 80M words. Similarly, Hu et al. (2020) found that GPT-2 had seriously diminished returns on performance when trained on a dataset of 4 billion words as opposed to 42 million words. Raffel et al. (2020) also found that a SuperGLUE model improved in performance when trained on datasets from 8 million to 34 billion words, though models trained on only 500 million tokens performed similarly to those trained on 34 billion.

## Unique Contributions

Zhang et al.'s research differed from existing research in that they used models that were pretrained with early stopping, as opposed to pretraining for a fixed number of iterations as in earlier studies. Earlier studies also used an encoder-decoder architecture, which meant their models had about double the parameters of the largest of the MiniBERTas tested.

**Evaluation Method**

Zhang et al. used five styles of evaluation in their study, including: classifier probing, information-theoretic probing, unsupervised relative acceptability judgments, unsupervised language model knowledge probing, and fine-tuning on NLU tasks. The models evaluated were a group of RoBERTa models pretrained on 1M, 10M, 100M and 1B words. Learning curves were then created to track the growth of each of these measures with respect to each of the data volumes they trained on. In total, 16 RoBERTa variants were pretrained. 12 were trained on 1M, 10M, 100M, and 1B word datasets. 1 variant was trained on a 30B word dataset, and 3 more were initialized with random parameters.

In each experiment, all 16 models were tested on each task equally. To show trends in the model's improvement, non-linear least squares were used to fit the learning curve to the results. Results of each task were plotted in a figure where the y-axis is the score and the x-axis is the volume of pretraining data. Some plots were normalized to adjust the results into the range [0, 1].

**Importance**

The research presented in this report demonstrates that the ability of language models reaches a point of greatly diminishing returns at a volume of only 10M to 100M tokens of pretraining data. This is significant because training a model on extremely vast volumes of data comes with non-trivial environmental impacts. These large scale models are also prohibitively expensive for smaller research groups or scientists with access to fewer resources.

# Works Cited

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When Do You Need Billions of Words of Pretraining Data?. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1112–1125, Online. Association for Computational Linguistics.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1725–1744, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-totext transformer. Journal of Machine Learning Research, 21(140):1–67.