Quantity doesn't buy quality syntax with neural language models

Cornell University

mv443@cornell.edu
https://vansky.github.io

Marten van Schijndel[1], Aaron Mueller[2], and Tal Linzen[3]

[1]Department of Linguistics, Cornell University
[2]Center for Language and Speech Processing, Johns Hopkins University
[3]Department of Cognitive Science, Johns Hopkins University

Johns Hopkins University

Abstract

Neural language models (LMs) can predict upcoming words remarkably well on average, but they often assign unexpectedly high probability to ungrammatical words. In this work we investigate to what extent these shortcomings can be mitigated by increasing the size of the network and increasing the amount of training data.

Evaluation data

Marvin and Linzen (2018) Syntactic Challenge Corpus [4]

Grammatical sentence should be more likely than ungrammatical one

P(The author laughs) > P(*The author laugh)

Models

2-layer LSTM LMs (5 random initializations each) trained ...

$$\text{with}\left\{\begin{array}{l}\text{100 hidden units}\\ \text{200 hidden units}\\ \text{400 hidden units}\\ \text{800 hidden units}\\ \text{1600 hidden units}\end{array}\right.\text{on}\left\{\begin{array}{l}\text{2M tokens}\\ \text{10M tokens}\\ \text{20M tokens}\\ \text{40M tokens}\\ \text{80M tokens}\end{array}\right. = \text{125 models}$$

Baseline Models

- Gulordava LSTM LM [3]
  Unidirectional
  2-layer, 650 hidden units (39M parameters)
  80M tokens
- GPT Transformer [5]
  Unidirectional
  12-layer, 110M parameters
  1B (1000 M) tokens
- BERT (Base) Transformer [2]
  Bidirectional (w/ future context removed [6])
  12-layer, 110M parameters
  3.3B (3300M) tokens
- Human grammaticality judgments [4]
  84 humans
  ≈10 judgments / pair

Aggregate results

Corpus size		Layer size	
2M → 10M	**5508.8**	100 → 200	**768.5**
10M → 20M	0.1	200 → 400	**63.5**
20M → 40M	**12.9**	400 → 800	0.2
40M → 80M	0.2	800 → 1600	0.1

Table 1: Strength of evidence for improvements in agreement prediction accuracy as a result of increasing corpus size averaging across layer size (left) or layer size averaging across corpus size (right), as quantified by Bayes factors. Boldfaced Bayes factors indicate strong evidence of improvement.

Agreement accuracy by construction

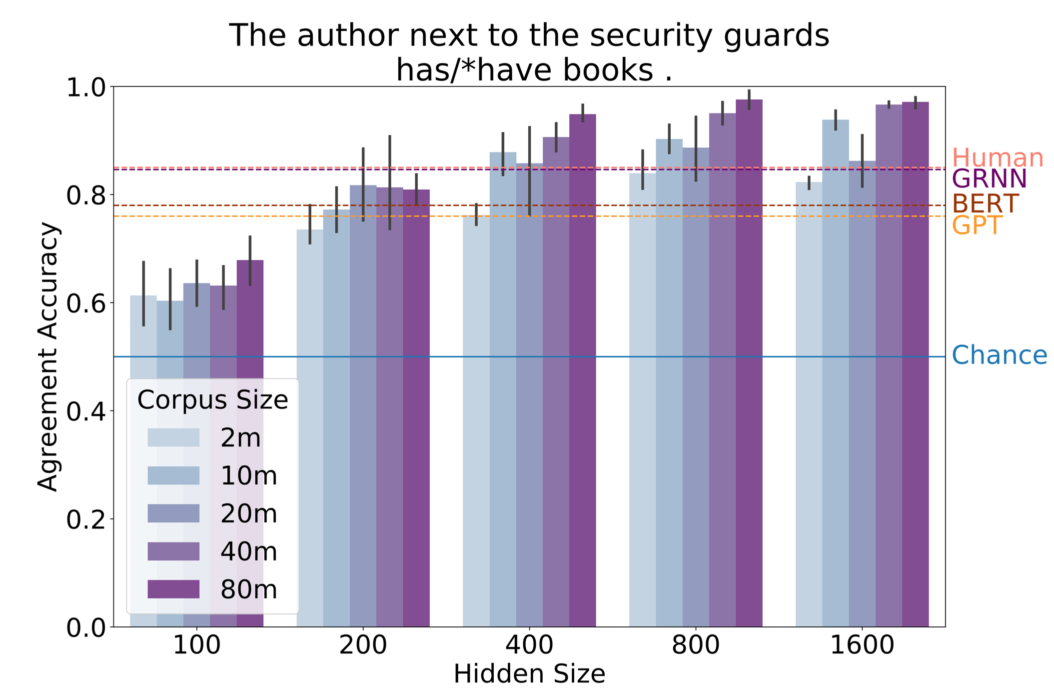

Figure 1: Agreement across a prepositional phrase

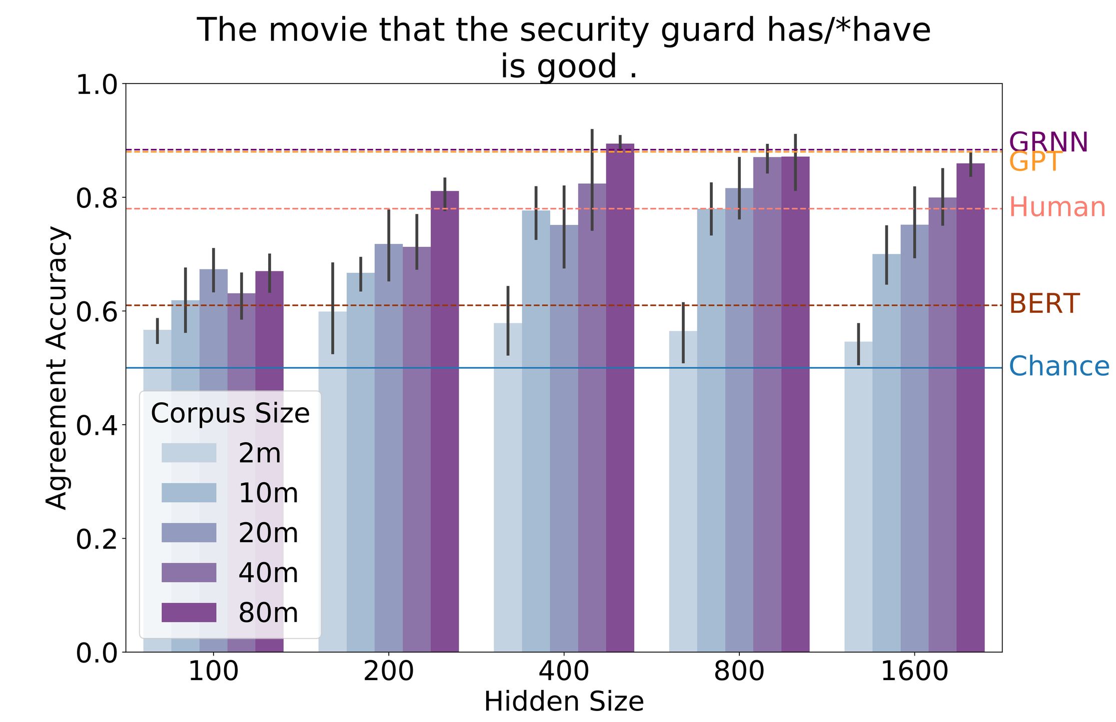

Figure 2: Agreement in an object relative clause

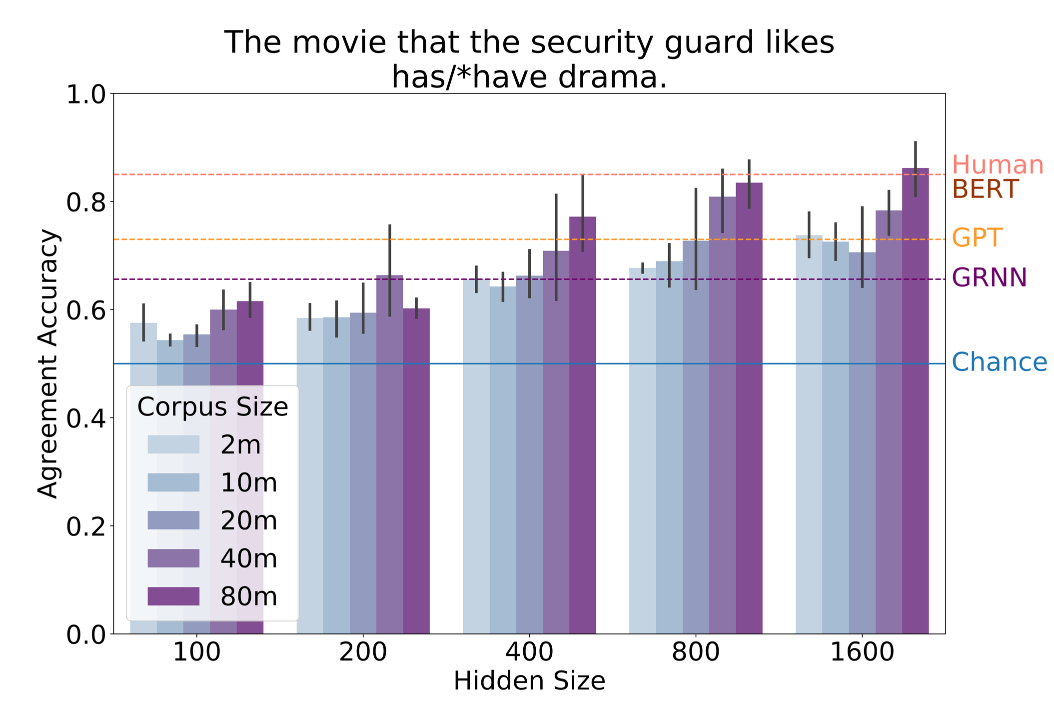

Figure 3: Agreement across an object relative clause (ORC)

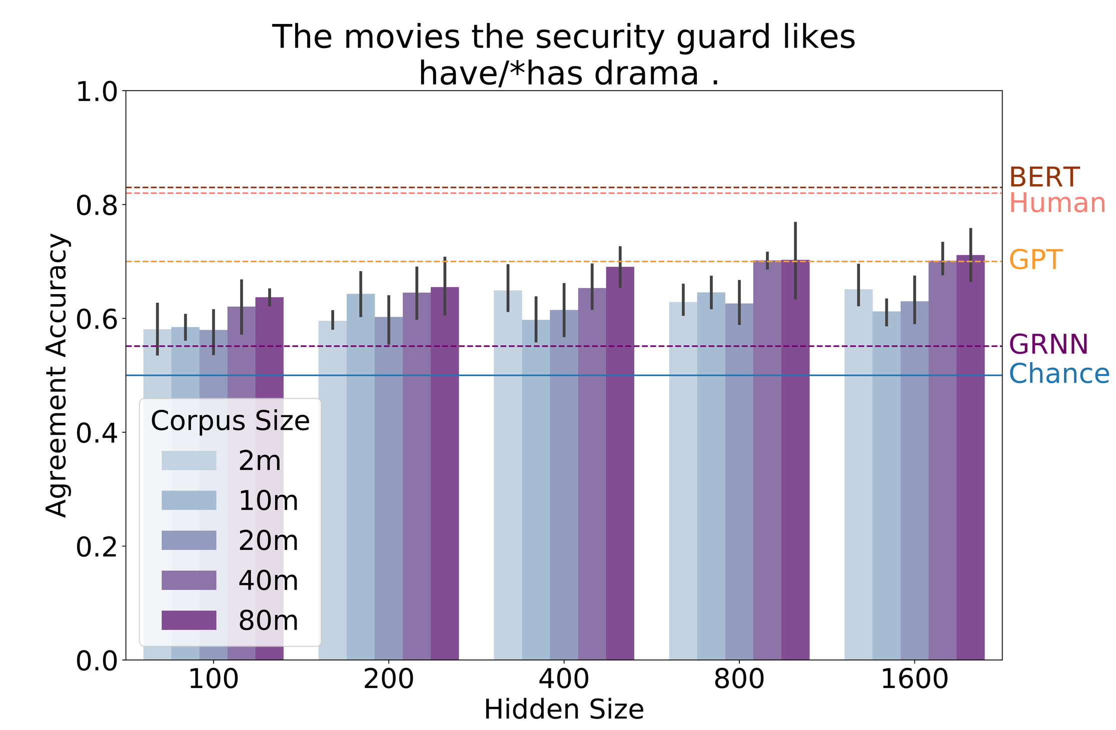

Figure 4: Agreement across an ORC (no *that*)

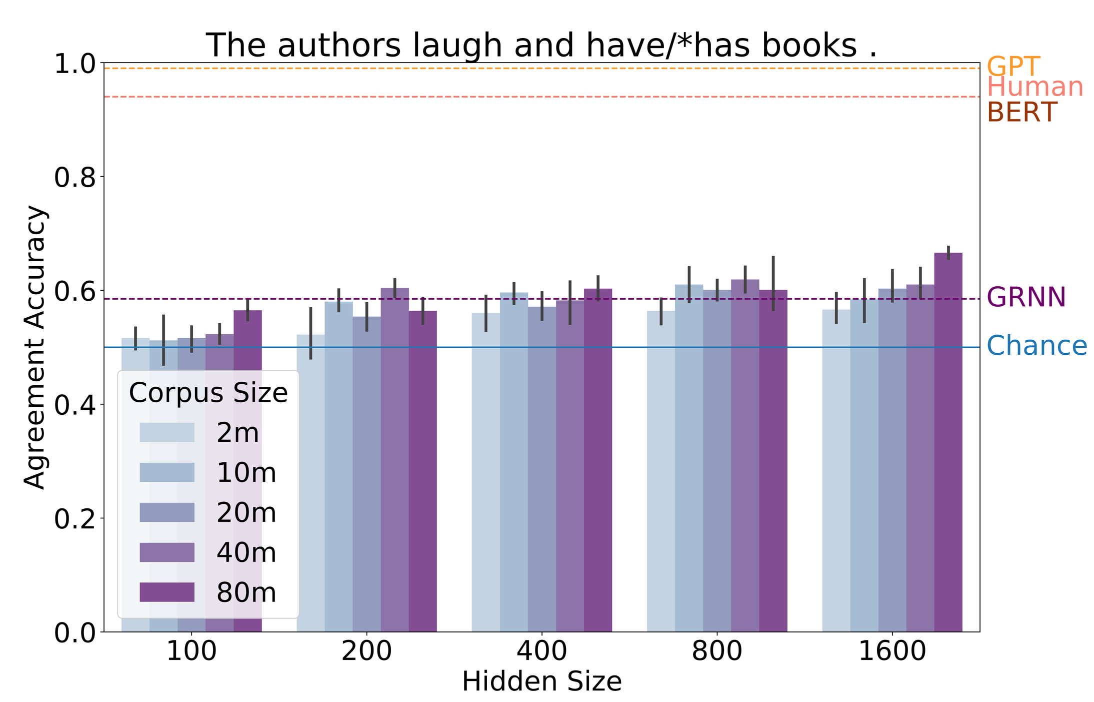

Figure 5: Agreement in a short coordinated verb phrase

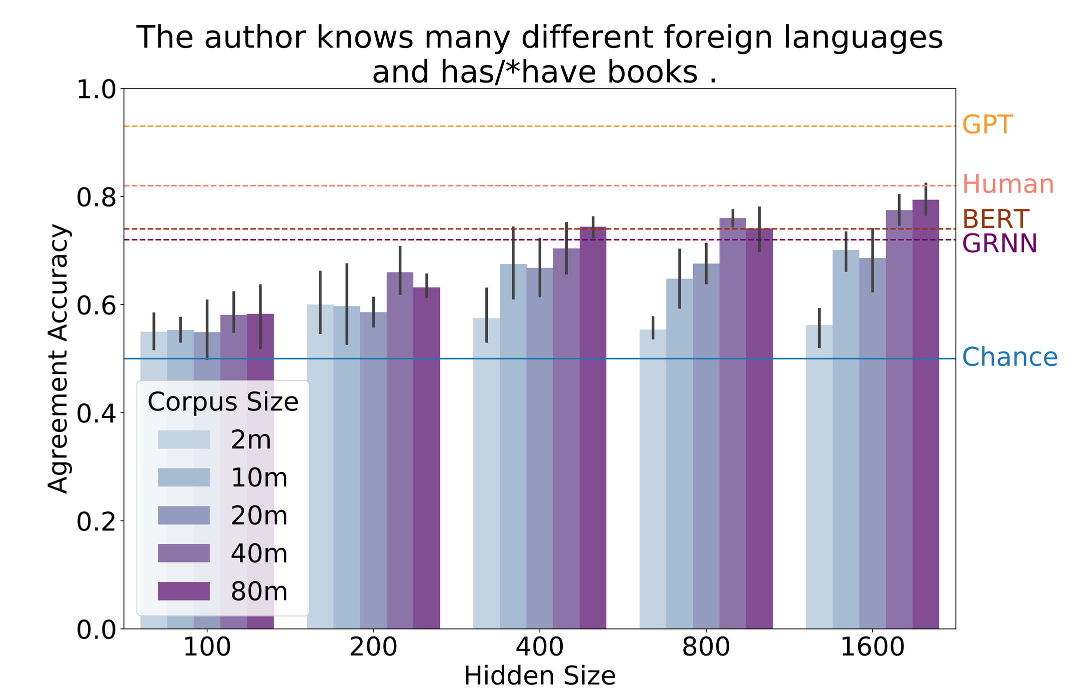

Figure 6: Agreement in a long coordinated verb phrase

How much data is enough?

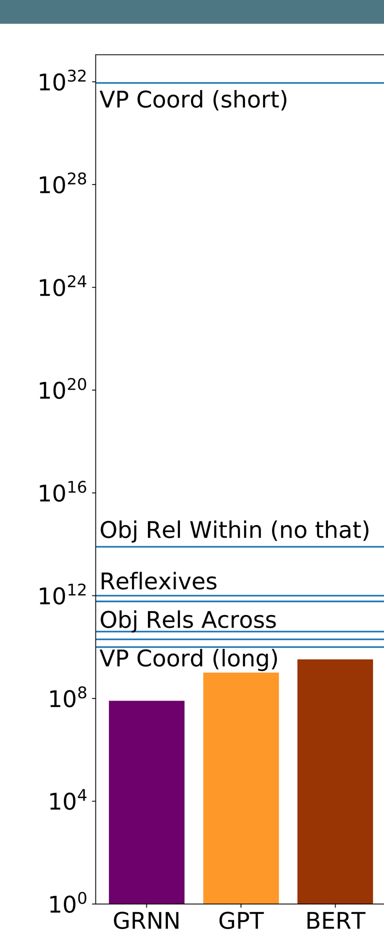

Figure 7: Human

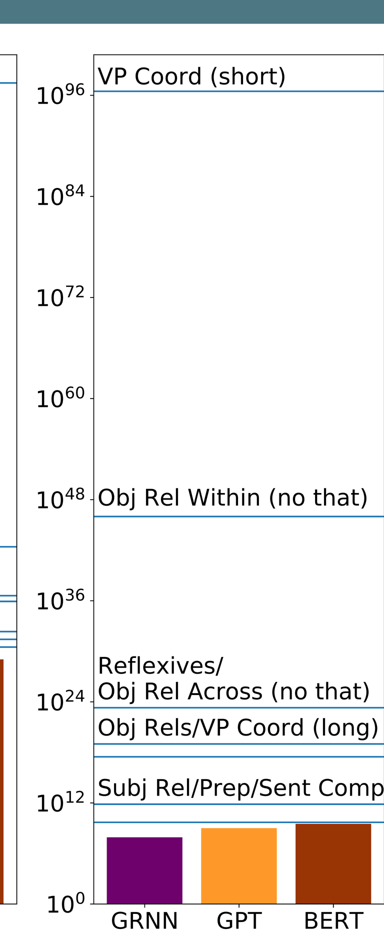

Figure 8: 99.99%

Number of training tokens required to reach human and near perfect accuracy in each construction, *assuming 20M→40M rate of improvement for every doubling of data*

Conclusions

- Layer size improves syntactic performance to a point.
- More training data helps sporadically
  But even with consistent improvement, LMs require an unreasonable amount of data to solve such a simple task.

We should likely focus on syntactically structured architectures or explicit syntactic supervision.

Related Finding

Pre-training a BERT-like LM on more data produces tiny downstream improvements [1].
562M → 18G ($5.6e^{8} \rightarrow 1.8e^{10}$) tokens improved NLI accuracy from 81.7%→82.3%.

References

[1] Alexei Baevski, Sergei Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. Technical report, Facebook AI Research, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[3] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018.

[4] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics, 2018.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[6] Thomas Wolf. Some additional experiments extending the tech report "assessing BERT's syntactic abilities" by Yoav Goldberg. Technical report, Huggingface Inc, 2019.