The human unlikeness of neural language models in next-word prediction

Cassandra L. Jacobs
University of Wisconsin - Madison
cjacobs2@wisc.edu

Arya D. McCarthy
Johns Hopkins University
arya@jhu.edu

Abstract

The training objective of unidirectional language models (LMs) is similar to a psycholinguistic benchmark known as the cloze task, which measures a word's predictability in context. However, LMs lack the rich set of experiences that people do, and humans can be highly creative. To assess *human parity* in these models' training objective, we compare the predictions of three neural language models to those of human participants in a freely available behavioral dataset (Luke & Christianson, 2016). Our results show that while neural models show a close correspondence to human productions, they nevertheless assign insufficient probability to how often speakers guess upcoming words, especially for open-class content words.

1 Introduction

The statistical regularities of language allow for people and statistical natural language processing models to easily learn the dependencies between words, phrases, semantic propositions, and syntactic structures. Typically, predictability in psycholinguistics is studied in part using a cloze task (Taylor, 1953), in which participants attempt to complete a sentence given a preamble, or context. This procedure has shown considerable success in NLP as well, with many prominent pre-trained language models in use today being trained on a cloze-like objective. In these tasks, the prediction task is more general: the model must predict the next word using the prior context for all of the words in the sentence. However, despite the successes of neural language models, whether these models achieve parity to humans on this specific training objective is unclear.

In this study, we tested how predictions by American English–speaking human participants in a word-by-word cloze task (next-word prediction) differ from three pretrained English neural language models (LMs) using a dataset of cloze predictions for naturalistic sentences (Luke & Christianson, 2016). Given a preamble (e.g. *You can glance at...*), participants' (and thus models') goal is to predict the next word (i.e. *the*). We use the Luke and Christianson (2016) dataset, which contains approximately 40 separate predictions of each next word in 120 distinct sentences, with 2687 to-be-predicted words in total. This rich dataset, containing over 41,000 unique human predictions, allows us to reliably estimate how well participants predict the next word. We can then compare different computational models against human performance.

2 Data and analysis

We assessed the performance of three neural LMs that differ in their architectures and the datasets they were trained on: `wmt19`, a `fairseq` model trained on the English portion of a machine translation dataset (Ng et al., 2019); `GBW`, an adaptive model trained on the Google Billion Words dataset (Chelba et al., 2014); and another adaptive model trained on the WikiText-103 benchmark (`wiki103`; Baevski & Auli, 2018; Merity et al., 2016). These three models are trained on the unidirectional language modeling objective (predicting a word given its predecessors), rather than the masked language modeling objective which has become common in recent years but is inappropriate for our task. Further, these models are

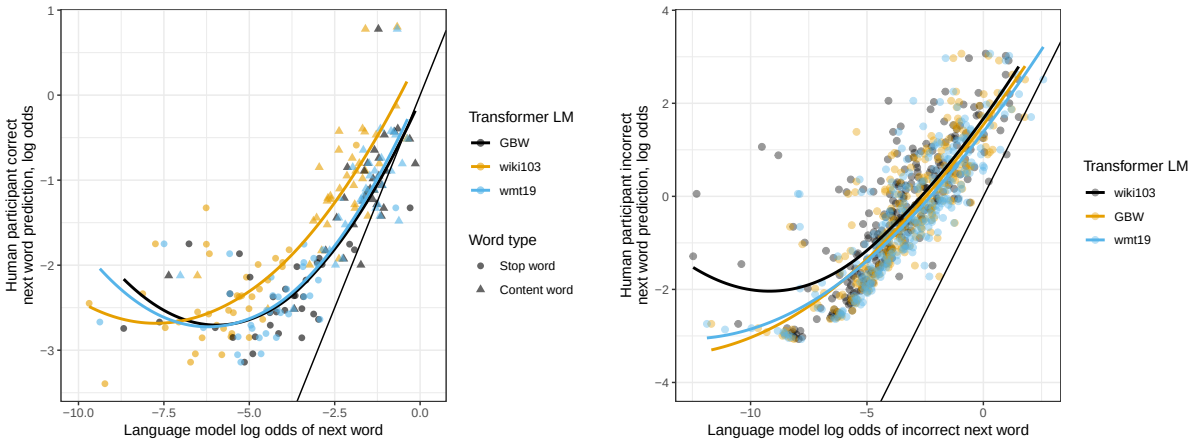

Figure 1: Correlation between log-odds of human and model predictions. There is greater correspondence between stop word predictions and content words, but the models consistently under-predict the odds of both the correct (left) and incorrect (right) guesses made by human participants.

able to predict entire completions of sentences, which lends to exciting avenues for future work. Finally, the three are implemented in a common toolkit, simplifying code reuse.

In prior work, Smith and Levy (2013) proposed that a successful language model will show a one-to-one relationship between human and model predictions in log odds space. Specifically, we calculate the odds that participants predicted a particular next word and the activation of that same word in neural language models. Following their proposal, we concern ourselves with the log odds of a completion, defined as $\log(p(w \mid c)/(1-p(w \mid c)))$ where $p(\cdot \mid c)$ for the human data is the proportion of participants selecting a particular response; for the neural models, $p(\cdot \mid c)$ is the activation of that same word given the preceding sentence context.

3 Results

Human guesses can be correct or incorrect about the identity of the upcoming word. Consequently, we present odds data for correct and incorrect predictions separately in Figure 1. At all levels, the neural LMs underestimate the frequencies of guesses made by human participants. For both correct (Figure 1, left) and incorrect (right) guesses made by participants, there is a strong relationship between human predictions and the activation of that prediction in the neural language models. A linear regression predicting the log odds of human predictions for correct answers using model log odds scores shows a close correspondence between the two ($\beta_0 = 0.88$, $\beta_1 = 2.27$; $p < .05$). For incorrect answers, this relationship is weaker ($\beta_0 = -2.87$, $\beta_1 = 1.39$, $p < .05$).

The error between the model and human log-odds (sum of squared residuals, *SSR*) shows greater prediction error for less predictable words, such as content words (*SSR* $= 415.0$) relative to stop words (*SSR* $= 94.1$). Some models perform worse overall, potentially due to training data set size (`wiki103` provided the worst fit, with ≈100M tokens for training). The disparities between these models are interesting on two fronts. First, the factors that influence fit to human behavioral data are still somewhat poorly understood (Goodkind & Bicknell, 2018), with greater language model training data not necessarily translating to better fit of human language processing (van Schijndel, Mueller, & Linzen, 2019). Second, researchers in psycholinguistics often want to use the models with the closest correspondence to human data so they can generate better predictors for statistical controls in their experiments. These data also show the potential value of using human behavioral data to evaluate model fit that is independent of perplexity on a corpus. Nevertheless, that the behavior is consistent across the three neural language models shows that this is not a quirk of any individual architecture or training data.

One possible explanation for the disparity between human and neural language models is that human participants may be *suboptimal* predictors in cloze tasks, though this seems unlikely because there are broader differences in accuracy. For example, human participants' most frequent predictions were more

likely to be correct (32%) than the top predictions of the strongest model (`wmt19`, with 26%), for both stop words (46% vs. 40%) and content words (20% vs. 15%).

4 Conclusions

These results show that neural language models of English sometimes fail to capture human creativity. Specifically, neural language models consistently underestimate the odds of a given completion in a cloze task, potentially because they know less about the world and the context to make predictions about upcoming words. Future work to understand the performance of the language models themselves should probe the specific factors that keep neural LMs from correctly estimating the predictions that people make about the linguistic future. Similarly, the rich dataset of Luke and Christianson (2016) provides potential future opportunities to investigate the relationship between reading times and eye movements as they relate to language model predictions (commonly referred to as "surprisal"; Hale, 2001), a topic that has gained considerable interest over the last decade.

References

Alexei Baevski, Michael Auli. 2018. Adaptive input representations for neural language modeling. In *ICLR 2019*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, Tony Robinson. 2014. One Billion Word benchmark for measuring progress in statistical language modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL 2019* (pp. 2978–2988).

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

Steven G. Luke, Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Stephen Merity, Caiming Xiong, James Bradbury, Richard Socher. Pointer sentinel mixture models. In *ICLR 2017*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, Sergey Edunov. 2019. Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation* (Volume 2: Shared Task Papers, Day 1), (pp. 314—319).

Nathaniel J. Smith, Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

Marten van Schijndel, Aaron Mueller, 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) (pp. 5835-5841).

Wilson L. Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415-–433.