Graded grammatical expectations in transformer models

Pablo Contreras Kallens
Cornell University, Ithaca, New York, United States

Rick Dale
UCLA, Los Angeles, California, United States

Morten Christiansen
Cornell University, Ithaca, New York, United States

Abstract

Large language models (LLMs) can be reasonably thought of as models of idealized statistical learners. Thus, the extent to which they grasp the grammar of the language they are trained on suggests how much of it can be learned from memorization, abstraction, and generalization of linguistic input. However, the knowledge of LMMs' grammar has largely been gleaned from examples of their outputs or datasets not designed to assess how native-like its knowledge is. In this study, we probed the knowledge of an LLM, GPT-3, with a graded grammatical acceptability task previously normed on humans. GPT-3's ratings were correlated with human ratings, even with minimal examples. Moreover, GPT-3's deviation from the human norms was predicted by the between-subject variation for each item, and these deviations were rarely outside of the range of human ratings. Follow-up analyses tested the extent to which local probabilistic structure drives these judgments using n-gram models.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*