Non-linear relational composition in large language models

Michael B McCoy
UC Irvine, Irvine, California, United States

Taylor Webb
Microsoft Research, New York, New York, United States

Anna Leshinskaya
UC Irvine, Irvine, California, United States

Abstract

The longstanding question of how neural networks could implement relational composition has been buoyed by recent success showing relational abstraction in transformer-based large language models (LLMs). We address recent findings showing some, but imperfect, generalizability in linear composition during knowledge retrieval of attributive triplets [Her- nandez, E. et al, (2024). Linearity of relation decoding in transformer language models arXiv:2308.09124)]. We report that limitations to relational generalization are explainable by two systematic factors. First, relational combinations that are more accurately retrieved generalize better than uncertain or inaccurate ones. Second, relational generalization scales with the semantic similarity of the entities being bound across triplets, showing that it is in fact non-linearly dependent on component meanings rather than being purely invariant. This aligns with longstanding findings that human judgments of adjectival combinations are likewise non-linearly interactive.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*