# Evaluating locality in NMT models

**Itay Itzhak**
Tel Aviv University, Tel Aviv, Israel, Israel

**Koustuv Sinha**
McGill, Montreal, Quebec, Canada

**Brenden Lake**
NYU, New York, New York, United States

**Adina Williams**
Facebook AI Research, New York, New York, United States

**Dieuwke Hupkes**
Facebook AI Research, Paris, France

**Abstract**

With a series of theoretically-informed tests, Dankers, Bruni, and Hupkes (2021) investigated how compositional the behavior of neural networks that are trained on fully natural data is. Focusing on neural machine translation (NMT), one of their key findings is that models appear to be modulating poorly between local and global behavior, where local changes in the input often affect the output in an unwanted manner. While their study is based exclusively on the behavior of the models, we take one step further and investigate how this non-locality manifests itself within the model. We develop metrics to quantify internal locality on the encoder side of the model, focusing on the attention mechanism. We find strikingly different patterns in models trained on different amounts of data that go beyond what could be observed behaviourally and present a range of experiments showing how local and global behavior is modulated within different setups.



In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. 