The Learnability of Model-Theoretic Interpretation Functions in Artificial Neural Networks

Adrian Brasoveanu (UC Santa Cruz), Jakub Dotlačil (Utrecht University)

A key issue in cognitive science going back to at least Fodor and Pylyshyn (1988) is explaining the systematicity of natural language interpretation: our capacity to interpret a potentially infinite number of novel expressions by systematically combining familiar elements; that is, the extrapolating generalizability of our learned natural language interpretation ability to out-of-training-sample (OOTS) novel combinations of familiar expressions. Do we need symbolic representations to explain this ability, or can it be captured by artificial neural networks (ANNs) that do not covertly implement a symbolic system? Frank et al. (2009) (see Venhuizen et al 2022 for another example of this approach) investigate this issue by examining whether a simple recurrent network (SRN) architecture can learn to map input sentences in a microlanguage to the correct truth conditions encoded as dense distributed representations with fuzzy truth values in the closed interval $[0, 1]$. The distributed representations capture the underlying binary model-theoretic information of whether the input sentence is true (1) or false (0) in a large representative sample of $250\,000$ models. To reformulate this in formal semantics terms, we investigate the *learnability of model-theoretic interpretation functions*, and crucially *evaluate the systematicity of the learned interpretation function by examining its ability to generalize to novel test sentences*, which come in 4 groups (word, sentence, complex and basic event) in increasing order of systematicity-related difficulty.

Our contribution. We extend Frank et al. (2009) in four ways. **(I)** We improve the sampling methodology and use MCMC, which samples from the full joint probability distribution over truth-conditional target vector components (in contrast to the independent sampling of Frank et al 2009 and Venhuizen et al 2022). **(II)** We modify the semantic representations available to ANNs. In the original study, target vectors encode only truth-conditional information. However, entities are a fundamental formal semantics theoretical primitive alongside truth values, essential for phenomena such as quantification and discourse anaphora (Kamp 1981, Heim 1982 a.o.). We test whether explicitly encoding entity information helps ANNs learn compositional structure, leading to better OOTS generalization. **(III)** We evaluate modern architectures: LSTMs and Attention / Transformer models. **(IV)** We extend the set of sentences in each of the 4 systematicity tests used to evaluate the models.

Results. Following Frank et al, we evaluate systematicity by examining comprehension scores (see Eq. 1), which range from -1 to $+1$, with higher values indicating closer match between model predictions z and target meaning vectors a. Systematicity is measured as the gap between comprehension scores for described (left panels in Fig. 1) and competing / distractor (right panels) events across 4 tests, with Basic Event the most challenging systematicity test. The rows in Fig. 1 show results for ATTN, LSTM, and SRN: comprehension score gaps between the left and right panels in each row decrease across the 4 tests, with the smallest gap for Basic Event. Scores are shown both with and without entity vectors (solid vs. hatched bars). When capacity-matched at ≈66k trainable parameters for no-entity models, all architectures achieve comparable systematicity (hatched bars). Total parameter match is achieved with different hidden dimensions: 48 for ATTN (which have 2 layers, each with 4 attention heads + MLP), 80 for LSTM, and 178 for SRN. **Entity vectors** (which double the dimension of the target vectors from 150 to 300) **improve all architectures**. SRN shows the largest absolute benefit, but SRN+entity also has the highest number of additional parameters (≈93.5k parameters) due to its very high hidden dim, compared to LSTM+entity (≈79k) and ATTN+entity (≈73k). ATTN+entity achieves comparable performance to SRN+entity with 27% fewer parameters. Thus, **architectural sophistication leads to better parameter efficiency when leveraging richer, theoretically-motivated semantic representations.**

$$\text{Comprehension}(a|z) = \begin{cases} \frac{\mathbb{P}(a|z)-\mathbb{P}(a)}{1-\mathbb{P}(a)} & \text{if } \mathbb{P}(a|z) > \mathbb{P}(a) \\ \frac{\mathbb{P}(a|z)-\mathbb{P}(a)}{\mathbb{P}(a)} & \text{otherwise.} \end{cases} \quad (1)$$

The prior probability $\mathbb{P}(a)$ that the target meaning vector a holds is the averaged L1 norm of meaning vector a. The conditional probability $\mathbb{P}(a|z)$ is computed as the dot product between the target meaning vector a and the model output z, divided by z's L1 norm.

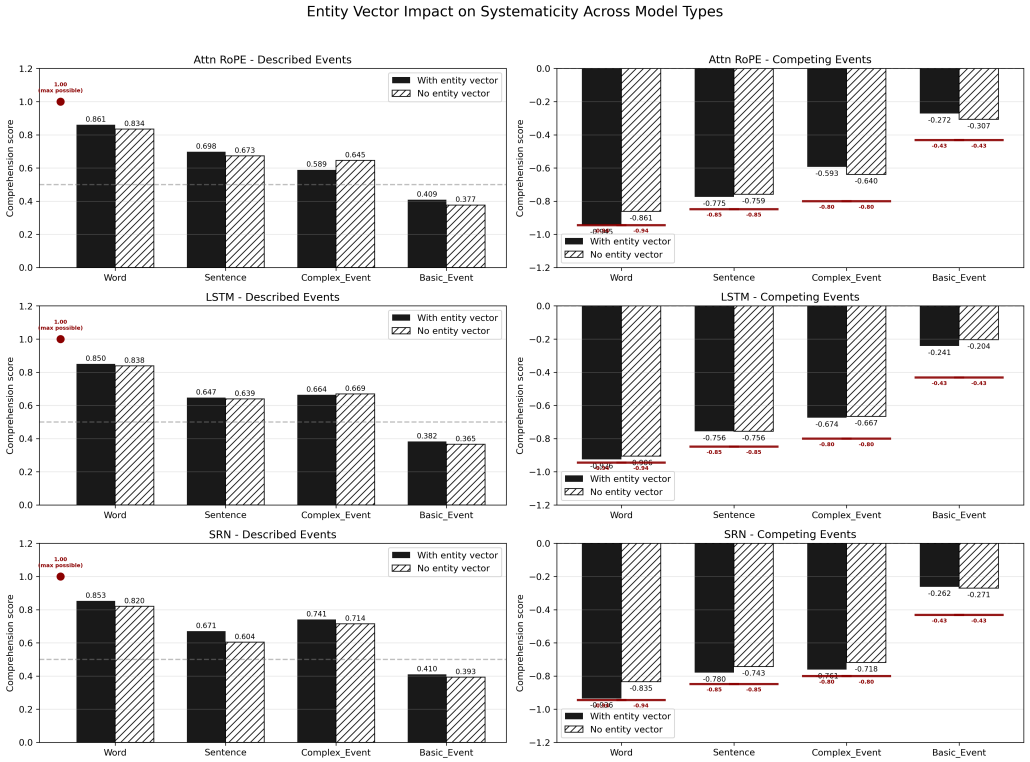

Fig. 1. Average comprehension scores for target described events (left) vs. competing distractor events (right). Hatched bars: no entity vectors. Solid bars: with entity vectors. Competing events are similar in form but semantically distinct. With capacity-matched models (no entity), all architectures perform comparably (3.6% difference). With entity vectors, attention models achieve comparable performance to Simple RNN using 27% fewer parameters, demonstrating parameter efficiency as the key manifestation of architectural sophistication. Dark red lines indicate comprehension scores that would be achieved by an ideal model that would learn to perfectly reproduce the target vector of the described event.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture. *Cognition*.

Frank, S. L., Haselager, W. F., & van Rooij, I. (2009). Connectionist Semantic Systematicity. *Cognition*.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD Dissertation, UMAss Amherst.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. Republished in Portner & Partee (2008).

Venhuizen, N. J., Hendriks, P., Crocker, M. W., & Brouwer, H. (2022). Distributional Formal Semantics. *I & C*.