Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning

Forrest Davis and Marten van Schijndel
Cornell University | fd252@cornell.edu

Details

Motivation
- Prior work has shown that English models learn, at least some, aspects of implicit causality (IC; Davis and van Schijndel, 2020; Upadhye et al., 2020).
- IC is attested cross-linguistically for humans (see Hartshorne et al., 2013; Ngo and Kaiser, 2020).
- Do neural models exhibit IC behavior in languages besides English?

Method
- Investigated 4 languages: English, Chinese, Spanish, and Italian
- English and Chinese models behave in accordance with IC
- Spanish and Italian models do not
- Spanish and Italian have a competing process, ProDrop, which influences pronouns
- Can fine-tuning demote this competing process and trigger more human-like behavior?

ACL 2021 | August 2-4, 2021

Takeaway
- Models learn systems of competing constraints which crucially differ across languages
- Fine-tuning on very little data which align with constraints found cross-linguistically can make model behavior more human-like

The base Spanish and Italian models (Figure 1) show little, or no, IC conditioned behavior: pronouns referring to objects, in light purple, are not more likely after object-biased IC verbs and vice versa for pronouns referring to subjects, dark purple. After fine-tuning on data demoting ProDrop (data with overt pronouns; Figure 2), models show IC conditioned behavior: IC verb bias influences the likelihood of pronouns.