Appendix: Construction Identification and Disambiguation Using BERT: A Case Study of NPN

Anonymous ACL submission

A   Limitations

This work is limited in several ways. Due to natural relative frequencies of various constructions, the dataset used for N*to*N is unbalanced between the N*to*N *construction* and *pattern*. This means that the training set for the classifier was quite small, because we ensured that training was balanced between the different classes. While the probing classifiers do achieve high accuracy, it is unclear how much accuracy is being capped by the limited data available. However, this fact, alongside our experiments with reduced training set sizes, indicate that the probes can learn with relatively little training signal.

This is experiment is also limited in only considering N*to*N, as opposed to the broader NPN construction. This is an intentional choice, as "to" has the most semantic subtypes of NPN associated with it. Future work is needed to see if the results here are robust to the inclusion of additional NPN examples with other lemmas into the dataset. We also only consider the English NPN construction, though the construction has been observed in a range of languages, including Dutch, English, French, German, Norwegian, Japanese, Mandarin, Polish, and Spanish (Weissweiler et al., 2024).

Finally, this work utilizes the probing classifier methodology, which has been criticized for providing indirect/correlational evidence of linguistic information in LM representations (Belinkov, 2022). Future work is needed to broaden the analysis to include other more direct, causal probing methodologies (e.g. AlterRep, Ravfogel et al. 2021; MaPP, Karidi et al. 2021; Reconstruction Probing, Kim et al. 2022).

B   Annotation Information

The entire dataset (6599 instances) were annotated by a team of two annotators. One annotator was a co-author of the paper. The other annotator was an undergraduate research assistant who received course credit for their annotation work. During annotation, roughly 400 examples were excluded from the final dataset for including potentially offensive content. The two annotators met to discuss any disagreements in their annotations and to jointly decide on the gold label. All instances which are used for training and testing are double annotated and adjudicated in this manner. IAA between the two annotators before adjudication was 84%.

References

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend, and Vivek Srikumar. 2021. Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 10300–10313, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Najoung Kim, Jatin Khilnani, Alex Warstadt, and Abed Qaddoumi. 2022. Reconstruction Probing. (arXiv:2212.10792). ArXiv:2212.10792 [cs].

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, page 194–209, Online. Association for Computational Linguistics.

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-*

COLING 2024), pages 16919–16932, Torino, Italia. ELRA and ICCL.