A principled approach to feature selection in models of sentence processing

Garrett Smith & Shravan Vasishth (University of Potsdam)
gasmith@uni-potsdam.de

Cue-based memory retrieval has proven to be a useful framework for understanding when and how processing difficulty arises in the resolution of long-distance dependencies. Most previous work in this area has assumed that discrete, morphosyntactic retrieval cues like [+subject] or [+singular] do the work of identifying (and sometimes misidentifying) a retrieval target in order to establish a dependency between words. However, recent work by Cunnings and Sturt on semantic interference suggests that hand-picked retrieval cues like these may not be enough to explain illusions of plausibility (Cunnings & Sturt, 2018), which can arise in sentences like *Sue remembered the letter that the butler with the plate/tie shattered*. Consistent with cue-based retrieval's predictions for such implausible sentences (Lewis & Vasishth, 2005; Vasishth et al., 2019), participants read the verb *shattered* faster when the distractor (*plate*) was a good direct object compared to when the distractor was a poor one (e.g., *tie*). Capturing such retrieval interference effects requires lexically specific features and retrieval cues, e.g., [±shatterable], but hand-picking the features is hard to do in a principled way and greatly increases modeler degrees of freedom. Moreover, competing models (e.g., self-organization Smith et al., 2018) often make different choices about which features to include, making direct comparisons problematic. To remedy these issues, we derive distributed numerical vectors for lexical features and retrieval cues using well-established methods from computational linguistics.

Method: We first parsed the British National Corpus using the automatic dependency parser from Qi et al. (2018), which produced dependency relation-governor-dependent triples like obj(*shattered*, *plate*). We then constructed a positive point-wise mutual information (PPMI) matrix (Church & Hanks, 1990) from the co-occurrence counts of dependent words (e.g., *letter* or *plate*) with particular syntactic attachment sites of governor words (e.g., *obj-shattered*) instead of window-based co-occurance. PPMI is a measure of the strength of association between the dependent and the governing attachment site. To create lexical feature and retrieval cue vectors, we applied truncated singular value decomposition to the PPMI matrix, keeping 300 dimensions (Deerwester et al., 1990). The cosine of the angle between these vectors is a measure of feature match, quantifying the plausibility of, e.g., *tie* or *plate* as direct objects of the verb *shattered*.

Results: To evaluate the resulting plausibility measure, we used `brms` (Bürkner, 2017) to fit Bayesian mixed effects models to the log-transformed total reading times at the verb from Cunnings and Sturt's two eye-tracking experiments. We defined the continuous "distractor advantage" predictor to be the scaled and standardized difference in our plausibility measure between the distractor noun (*plate*) and the correct retrieval target (*letter*). In implausible sentences, we found that the more plausible the distractor was compared to the target, the faster reading times were (-54ms, 95% credible interval [-89, -23]). In plausible sentences, the results were inconclusive: the distractor advantage effect was -12ms (95% credible interval [-41, 18]). These results are consistent with Cunnings and Sturt's original analysis.

Discussion: This work demonstrates that a corpus-derived plausibility measure based on distributed feature and cue vectors can predict illusions of plausibility. Our method allows us to derive these features quasi-automatically, greatly restricting modeler degrees of freedom. Embeddings of this sort are often used for semantic tasks like word analogy (Mikolov et al., 2013), but future work will determine whether this method is capable of handling purely morphosyntactic similarity-based interference (e.g., subject-verb number agreement or reflexive binding Dillon et al., 2013; Jäger et al., 2019). More broadly, our distributed feature vectors can be readily plugged into existing parsing models by swapping out the discrete, hand-selected ones, putting very different models (e.g., cue-based retrievaland self-organized sentence processing)on more equal footing and facilitating future quantitative comparisons.

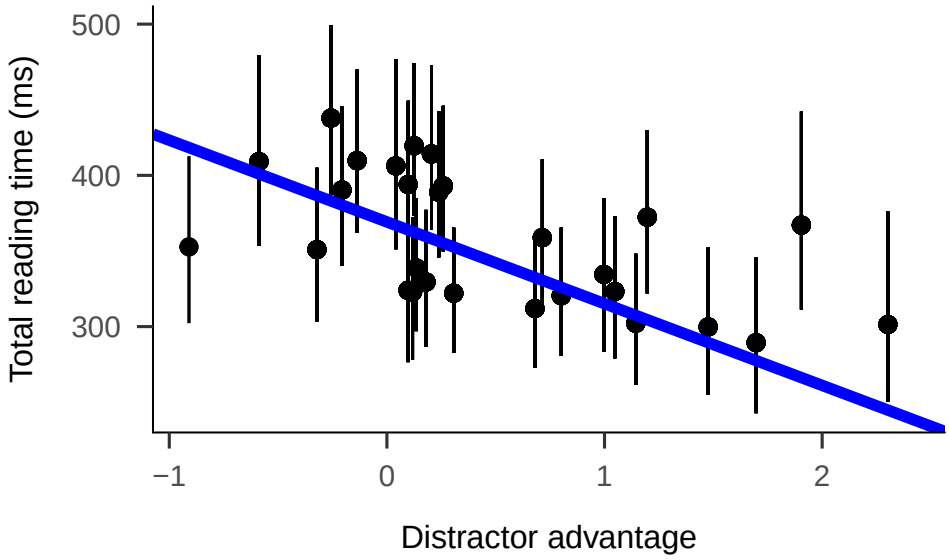

Figure 1: Total reading times at the verb in implausible sentences as a function of distractor advantage. The points show each item's posterior mean reading time and 95% credible interval. The blue line shows the population-level effect of distractor advantage: Reading times decrease as the distractor becomes a better fit to the verb's retrieval cues compared to the target.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cunnings, I., & Sturt, P. (2018). Retrieval interference and semantic interpretation. *Journal of Memory and Language*, *102*, 16–27. doi: j.jml.2018.05.001

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*, 85–103.

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2019). *Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study.*

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018, October). Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 160–170). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://nlp.stanford.edu/pubs/qi2018universal.pdf

Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject-verb number agreement. *Cognitive Science*, *42*(S4), 1043–1074. doi: 10.1111/cogs.12591

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, *23*(11), P968–982. doi: 10.1016/j.tics.2019.09.003