How far does probability take us when measuring psycholinguistic fit?

Sathvik Nair[1], Shohini Bhattasali[2], Philip S. Resnik[1], Colin Phillips[1]
[1]University of Maryland, [2]University of Toronto

During sentence comprehension, humans easily notice most, but not all lexical substitutions. Our work investigates the semantic illusions that occur, for example, when participants respond to the question *What is the name of the raised bumps on paper which enables deaf people to read?* with “Braille.” They may not recognize the statement violates their world knowledge since *deaf* was substituted in the place of *blind* [2]. Some accounts suggest the *semantic similarity* between the correct and substituted items could affect participants' sensitivity to an illusion [11, 1]. Alternatively, the substituted word’s fit within the sentence’s *context* could also explain the effect[4, 9]. In the current work, we apply semantic similarity [5] and probability estimates [6] from language models(LMs) to behavioral data on semantic illusions in English from [8]. We find that the fit of the substituted word in the context is more important to explaining illusions than semantic similarity of the correct and substituted words. However, our metric of goodness of fit, the substituted word’s probability estimated by a language model given the context, only weakly predicts behavioral results. Follow-up analysis of data from a speeded cloze production task [10] sheds light on why LM-based metrics are a weak fit to our psycholinguistic data.

We use data from [8], where 100 participants judged whether 108 illusion sentences in declarative form (example a) were true or false. The primary measure for each sentence was the *illusion rate*, operationalized as the proportion of participants who said the illusion sentence was true. All participants were independently checked for relevant world knowledge. Trials were excluded if participants lacked the relevant factual knowledge.

In a previous study analyzing a smaller set of illusions in question format [9], Muller et al. used a vector-space distributional model of word similarity, word2vec [7], to compare the semantic similarity of the correct and substituted item and found a small but insignificant relation. Models like word2vec are not context-sensitive and do not consider multiple meanings of a word. These issues are addressed by more recent distributional measures, derived from representations under neural language models. We applied a state-of-the-art, contextualized model of semantic similarity[5], which is highly correlated with human judgments. We found much higher degree of similarity between the correct and substituted items than word2vec alone (figs.1 & 2), reflecting what the experimenters had intended to do in their materials. Nonetheless, this metric still showed no significant correlation with illusion rates.

We also considered the fit of the substituted word in context by measuring its estimated probability under a language model. Since human participants had access to both left and right context during the judgment, we predict the probability of the substituted token given the rest of the sentence bidirectionally with RoBERTa [6]. When comparing contextual probability to illusion rates (fig. 3), substitutions with higher estimated probability are more likely to yield illusions. However, correlation with the illusion rate is still weak *($r = -0.21, p < 0.05$)*, so estimated probability appears to be a poor measure of contextual fit. To clarify why, we evaluated RoBERTa’s probability estimates against data from a speeded cloze task (40 participants, example b) [10]. One major conclusion of [10] was that production times more accurately reflect the activation of alternate completions than cloze probabilities. We computed the entropy of completions under RoBERTa, considering only responses produced by three or more participants (269 sentences). Compared to the cloze entropy (fig 4), we found a strong correlation, in line with results from [3], but we did not find a significant effect when we compared entropy under RoBERTa with items’ averaged production latencies (fig. 5). Thus, the fit of the substituted item within the sentence’s context is important to determine illusion rates, but its probability given the context is only one part of the explanation. If RTs better reflect the activation of the possible completions, it is possible that an activation-based account, specifically based on shared semantic features between the correct and substituted item in context[8](not explicitly represented in LMs), could be a next step in modeling illusion rates.

Illusions

(a) The name of the raised dots on paper that enable **deaf** (blind) people to read is Braille.

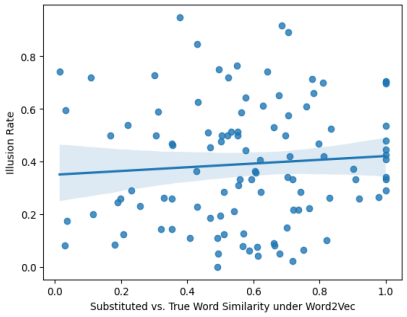

Figure 1: Correlation between illusion rate and similarity of the substituted word with the correct word under word2vec (r = 0.07, p > 0.05)

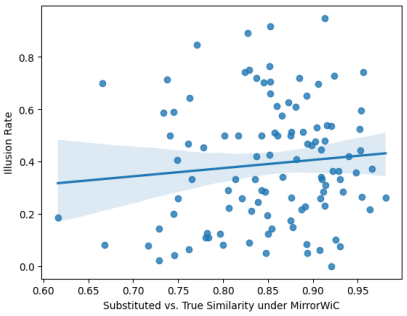

Figure 2: Correlation between each item's illusion rate and similarity of the substituted word with the correct word under contextualized embeddings from MirrorWiC (r = 0.09, p > 0.05)

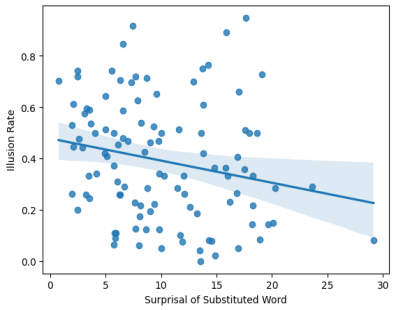

Figure 3: Correlation between each item's illusion rate and surprisal (negative log probability) under RoBERTa (r = -0.21, p < 0.05)

Speeded Cloze

(b) The distant alarm signaled the **fire**

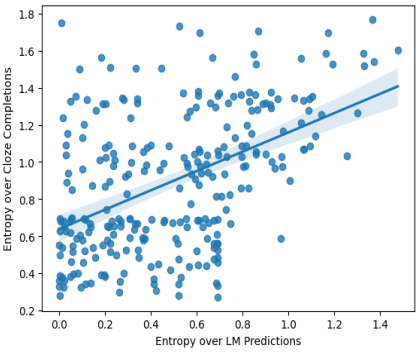

Figure 4: Correlation between entropy of speeded cloze completions under RoBERTa and cloze entropy (r = 0.4, p < 0.001)

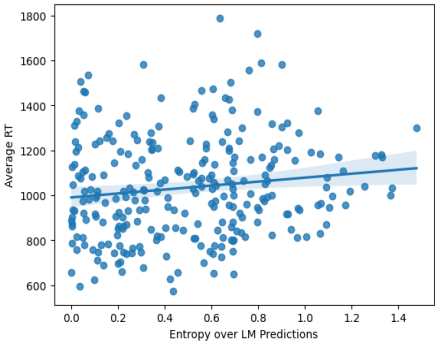

Figure 5: Correlation between entropy of speeded cloze completions under RoBERTa and average production time (RT) for each stimulus (r = 0.12, p < 0.05)

References

[1] Cook, A. E., Walsh, E. K., Bills, M. A., Kircher, J. C., & O'Brien, E. J. (2018). Validation of semantic illusions independent of anomaly detection: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*. **[2]** Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*. **[3]** Eisape, T., Zaslavsky, N., & Levy, R. (2020). Cloze distillation: Improving neural language models with human next-word prediction. Association for Computational Linguistics (ACL). **[4]** Hannon, B., & Daneman, M. (2001). Susceptibility to semantic illusions: An individual-differences perspective. *Memory & cognition*. **[5]** Liu, Q., Liu, F., Collier, N., Korhonen, A., & Vulić, I. (2021, November). MirrorWiC: On Eliciting Word-in-Context Representations from Pretrained Language Models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. **[6]** Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. **[7]** Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*. **[8]** Muller, H. E. (2022). *What Could Go Wrong? Linguistic Illusions and Incremental Interpretation* (Doctoral dissertation, University of Maryland, College Park). **[9]** Muller, H., Resnik, P., & Phillips, C. (2020). Explaining item-wise variability in Moses illusions. In *33rd Annual CUNY Conference on Human Sentence Processing-Amherst, Massachusetts*. **[10]** Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*. **[11]** Van Oostendorp, H., & De Mul, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta psychologica.*