Modeling agreement attraction effects in vector space

Eva Neu[1], Maayan Keshev[2] and Brian Dillon[1]
[1]University of Massachusetts Amherst, [2]The Hebrew University of Jerusalem

Background. Agreement attraction effects (Bock and Miller, 1991; Wagers et al., 2009) are sensitive to syntactic positions of target and distractor. E.g., in (1), speakers are more likely to be distracted by the noun 'flight' than by the more deeply embedded 'canyon' (Franck et al., 2002). Classic feature percolation/spreading models can capture such hierarchical effects, but don't generalize to all attraction environments (Wagers et al., 2009; Bhatia & Dillon, 2022); classic cue-based retrieval does not explain why embedding depth should modulate misretrievals, or why there is almost no attraction from the distractor that is linearly proximate to the verb (Lewis & Vasishth, 2005); and neural language models wrongly predict that the distractor linearly closer to the verb should cause more, not less interference (Arehalli & Linzen, 2024). Here we test the hypothesis that distributed vector representations of hierarchical structure can capture this effect. We build on work by Keshev et al. (2024a) who argue that sentences are encoded in working memory by binding distributed vector representations of lexical items to distributed vector representations of syntactic positions. All item-position bindings are superimposed on the same connection matrix. Therefore, recovering an item is prone to interference from items bound to similar positions, predicting interference between syntactic position vectors that are similar to each other. Keshev et al. demonstrate this effect by directly manipulating the cosine similarities of randomly generated vectors, but their position vectors do not represent hierarchical structure in a principled way. In this work, we demonstrate how to derive vector representations of syntactic positions such that higher cosine similarity between position vectors corresponds to higher rates of interference, and we show that this captures the hierarchical depth effect on agreement attraction.

Computing vectors. We set up a constituency parse for English input sentences such as (1) using the Berkeley Neural Parser and enforce binary branching. Each node in the tree is assigned an orthogonal base vector depending on its syntactic category. E.g., all N/NP nodes receive the same base vector. Using the update rule from the Temporal Context Model of position-coding in working memory (Howard and Kahana, 2002), we compute the position vector of a node as the weighted sum of its own base vector and the position vector of its mother node (if present), using a free parameter alpha (2). Since the function in (2) applies recursively, the position vector of said mother node in turn contains the position vector of its own mother, etc. As a result, the position vector of a node contains the base vectors (the category information) of all dominating nodes. Base vectors of more distant nodes make up a smaller part of its representation. The position vectors assigned to sentences such as (1) correctly result in higher cosine similarity for the hierarchically closer distractor for all alpha (4). Since the contribution of dominating nodes to a node's position vector is diluted across projections, the position vector of the NP immediately dominating the target makes up a larger portion of the position vector of 'flight' than of 'canyon.'

Simulations. To assess how well the relative cosine similarities from our model align with empirical findings on agreement attraction, we first fitted our model to experimental data reported in Keshev et al. (2024b), who tested the subject's final interpretation in single-depth embedding attraction sentences, using a 4-alternative forced choice including singular and plural versions of the target and the distractor (3). E.g., Distractor-Veridical responses correspond to the distractor lexical item with the veridical number marking of the target. Using Maximum Likelihood Estimation, we scaled our cosine similarities (for alpha = 0.3) by a parameter to match Keshev et al.'s data (5)–(6). We then use this fitted parameter to generate predictions for held-out data from Keung and Staub (2018) who presented their subjects with a 2-alternative forced choice between singular and plural agreement for a singular subject followed by either a singular or plural distractor. Crucially, they used distractors in different syntactic positions, i.e, singly and doubly embedded PPs.

Results. As summarized in (7), the algorithm correctly predicts higher interference rates from syntactically similar distractors. For distractors in single-embedded PPs such as 'flight' in (1), interference rates for plural distractors are almost twice as high as for singular distractors. For distractors in double-embedded PPs such as 'canyon' in (1), on the other hand, interference rates are equally low for singular and plural distractors. This is in line with Keung and Staub's (2018) finding that more distant distractors only marginally increase agreement attraction errors and RT (8).

Discussion. Our distributed vector representations succeeds in generating higher rates of interference from distractors in hierarchically closer positions. Besides matching previous data, the vector representation derived by our algorithm can make concrete predictions for interference rates in other syntactic constellations that can be tested empirically. The method we develop for representing syntax in vector space is simple, based on syntactic categories and hierarchical structure only, and lends itself to a wide range of other modifications and implementations.

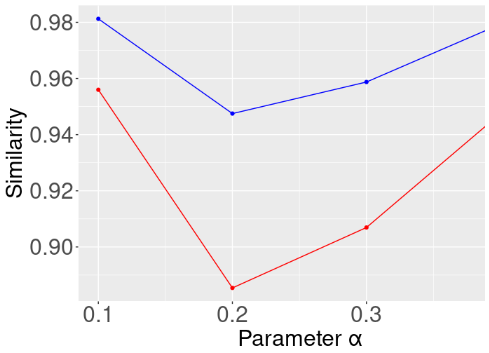
Data

(1) The [$_{\text{target}}$ helicopter] for the [$_{\text{distractor}}$ flight] over the [$_{\text{distractor}}$ canyon] is rusty.

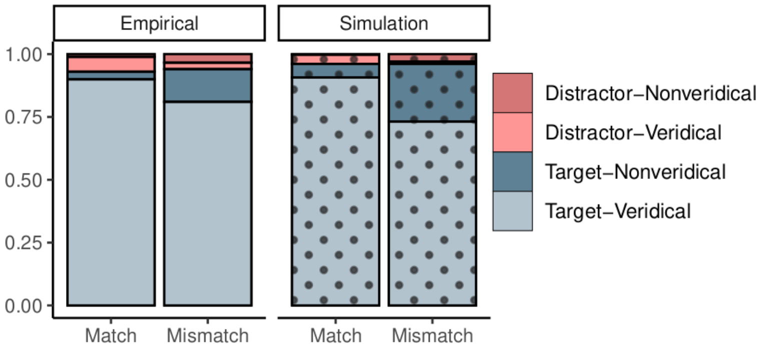
(2) PosVec(X) = alpha × BaseVec(X) + (1 − alpha) × PosVec(MotherOfX)

(3) The apprentice of the chef/chefs worked diligently.
Who worked diligently?
The apprentice / the apprentices / the chef / the chefs

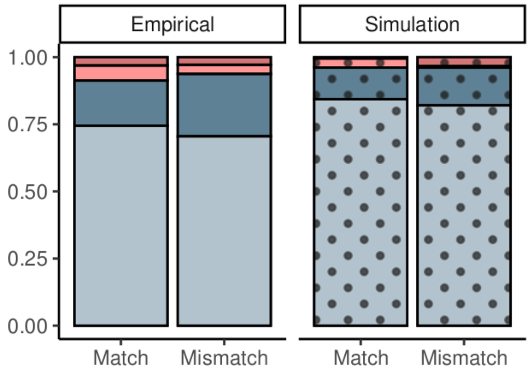
(4) Cosine similarity of 'flight' (blue) & 'canyon' (red) to 'helicopter'



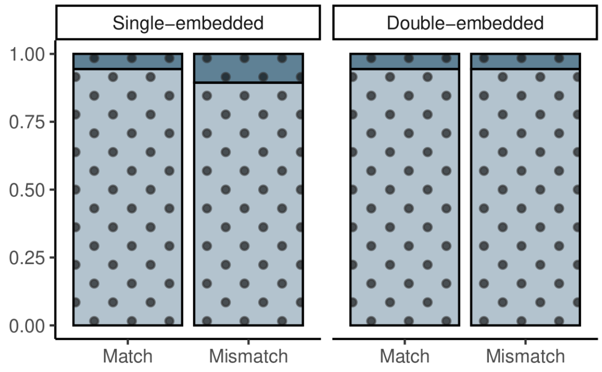
(5) Singular target



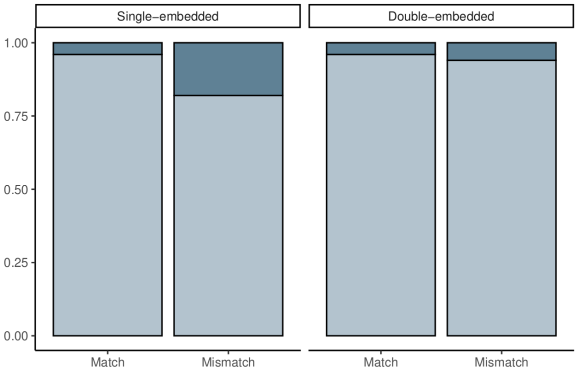
(6) Plural target



(7) Effect of syntactic position



(8) Empirical data from Keung and Staub



References

Bock & Miller. (1991). Broken agreement. Franck, Vigliocco & Nicol. (2002). Attraction in sentence production: The role of syntactic structure. Keshev, Cartner, Meltzer-Asscher & Dillon. (2024a). A working memory model of sentence processing as binding morphemes to syntactic positions. Keshev, Koesterich, Meltzer-Asscher & Dillon. (2024b). Feature distortion and memory updating: Experimental and modeling evidence. Keung & Staub. (2018). Variable agreement with coordinate subjects is not a form of agreement attraction.