Grokking Wug Vectors

Jacob A. Matthews and Marten van Schijndel

Cornell University

jam963@cornell.edu

Abstract

- Pseudowords (*nonce words*) are often used to construct psycholinguistic stimuli.
- Computational linguists commonly use *word vectors,* with *cosine* used to measure semantic similarity.
- Pretrained word vectors may not be appropriate for use with nonce paradigms.

Contextual vs Non-Contextual

Contextual models (like BERT [1] and GPT-2 [2]) output different word vectors given other context words in an input sequence.

Non-contextual models (like fastText [3]) output identical word vectors regardless of their input context.

Bleached vs Authentic Context

For contextual models BERT and GPT-2, we generate word vectors with both *bleached* and *authentic* contexts.

Bleached input: "{word} is a word"

Auth. context input: "… c_{w-1} {word} c_{w+1} c_{w+2} …"

where c_i is a word from an authentic sentence containing the relevant {word}.

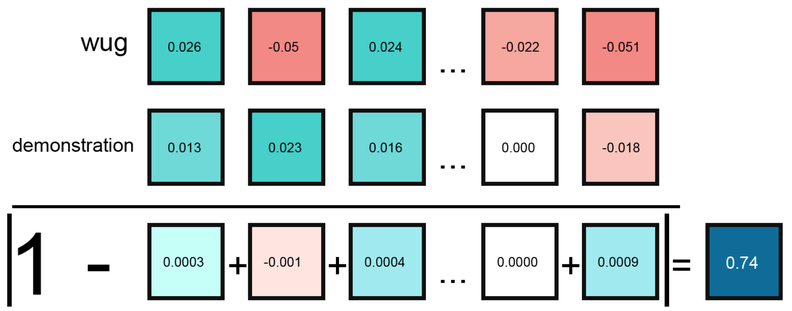
References

[1] Devlin et al. 2019 [2] Radford et al. 2018 [3] Bojanowski et al. 2016

Is wug vector "semantic similarity" actually "orthographic similarity"?

COS

Cosine distance (COS) measures the similarity of two vectors as the normalized sum of their elementwise product (absolute values closer to zero are more similar).



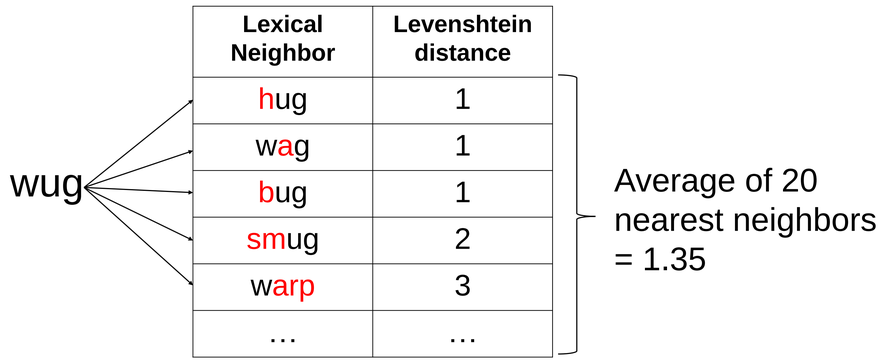
Different models give inconsistent distance measurements when comparing common pseudowords with familiar lexical items.

	fastText	BERT	GPT-2
wug	demonstration (0.74)	sing (0.76)	wicked (0.79)
heaf	any (0.76)	surely (0.75)	deaf (0.46)
glack	lack (0.70)	bake (0.76)	lack (0.35)
stup	stupid (0.58)	commission (0.78)	pupil (0.71)
plad	shirt (0.60)	previous (0.68)	tradition (0.42)

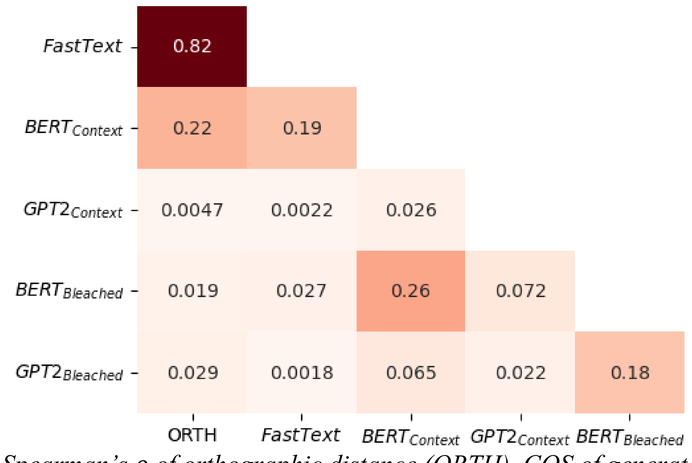
Nearest common English word to selected pseudoword (COS).

ORTH

We use the average Levenshtein orthographic distance of the 20 nearest lexical neighbors to each pseudoword (ORTH).



We find that ORTH is strongly correlated with COS for non-contextual word vectors, but not for contextual word vectors.



Spearman's ρ of orthographic distance (ORTH), COS of generated pseudoword to corresponding lexical item used for generation, by model.

Discussion

Using word vectors with nonce paradigms introduces model-specific confounds. Non-contextual vectors (fastText) are mostly explained by orthographic differences between pseudo- and real words. Contextual representations with authentic contexts (like those generated with BERT or GPT-2) are inconsistent across models.

Takeaways

- ✓ Different models treat pseudowords differently.
- ✓ For pseudoword vectors that encode orthographic features, use fastText.
- ✓ To avoid orthographic confounds, use a contextual model with bleached contexts.