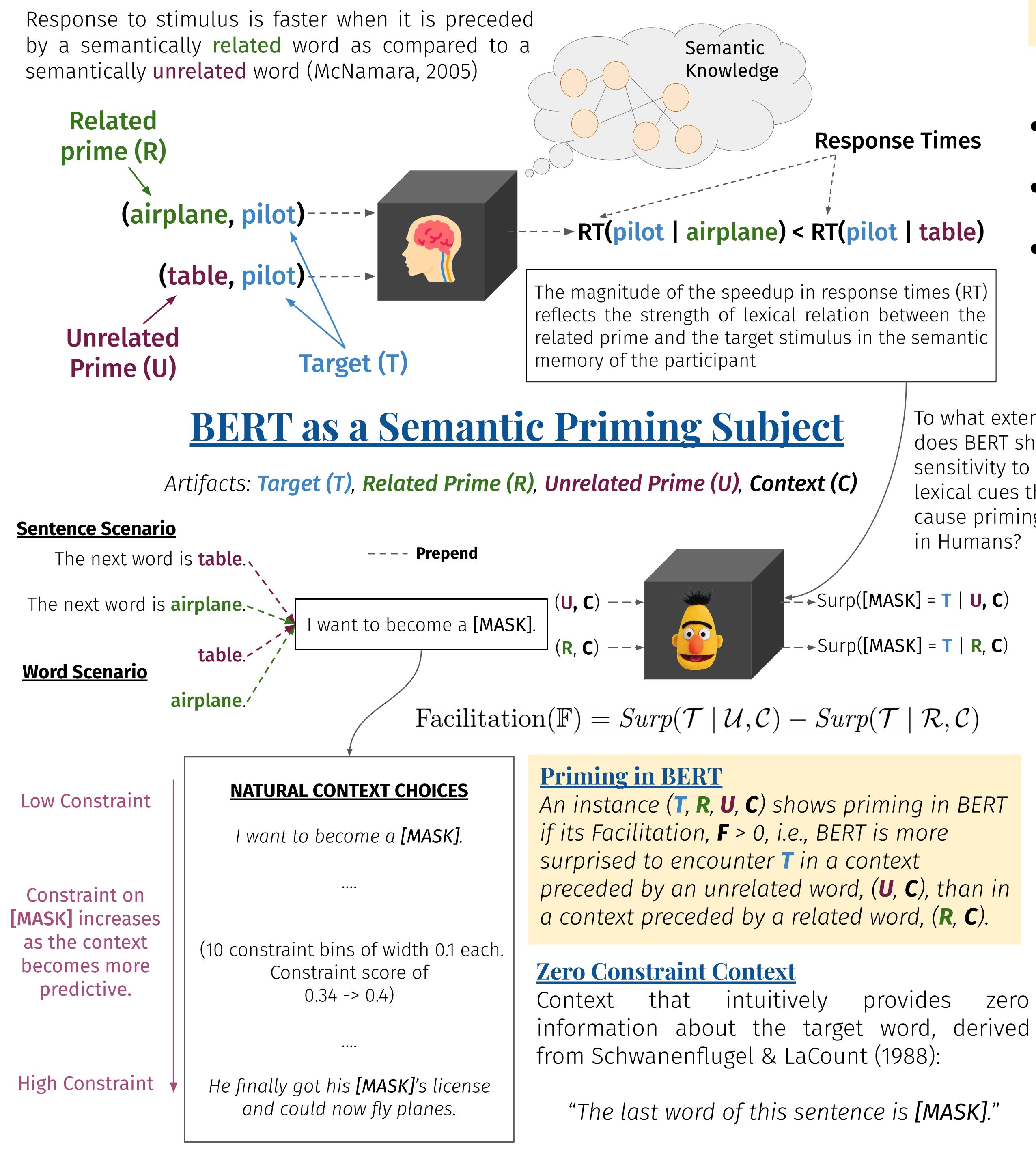
Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming <u>Kanishka Misra</u>¹, Allyson Ettinger², Julia Taylor Rayz¹ Contact: kmisra@purdue.edu, @kanishkamisra Purdue University¹, University of Chicago² Paper: https://arxiv.org/abs/2010.03010

Semantic Priming



Contextual Constraints

- We study BERT's priming behavior under varying levels of predictive constraint on [MASK].
- Our measure of constraint is grounded in Psycholinguistic studies that leverage Cloze Tasks (Taylor, 1953) and formulate the level of constraint of a sentence to be the probability of the best completion (Schwanenflugel & LaCount, 1988; Federmeier & Kutas, 1999).
- Constraint score is computed as average probability of the top-predicted word according to BERT-base and BERT-large:

 $constraint(\mathcal{C})$

$$) = \frac{1}{2} \sum_{m \in \{b,l\}} \max_{x \in \mathcal{V}} P_m$$

predict **[MASK]**.

Guiding Ouestion

How does BERT use lexical cues in context to inform word probabilities?

Experimental Dataset

- T, R, U triples extracted from the Semantic Priming Project (SPP) (Hutchison et al., 2013). The SPP dataset contains 16 unique lexical relations (measured between Target, T, and Related Prime, R).
- Contexts (C) containing target words sampled from the ROCstories corpus (Mostafazadeh et al.,
- **Total instances:** 23232, with 2112 unique triples

Effect of contextual constraint on Priming in BERT

How does BERT's facilitation vary with the predictive properties of the context?

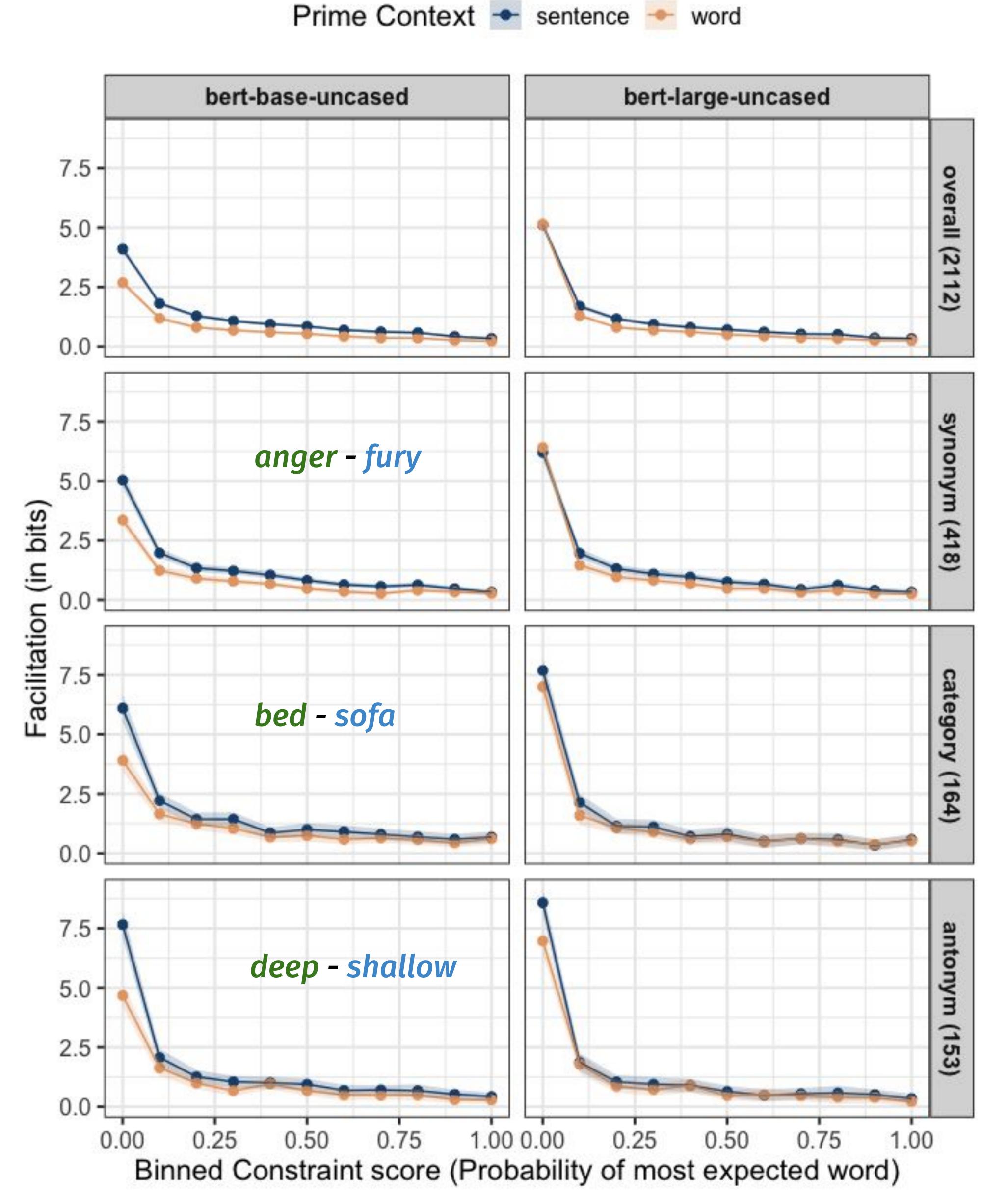


Figure 1: Average Facilitation for BERT-base and BERT-large vs. Binned Constraint Score. The first row shows results on the entire dataset while the next three show results for relations between the related prime and the target where facilitation in low constraint items is greatest. Refer to paper for full results.

/ To what extent does BERT show sensitivity to lexical cues that cause priming

Surp([MASK] = T | U, C)

- --►Surp(**[MASK]** = **T** | **R**, **C**)

provides zero

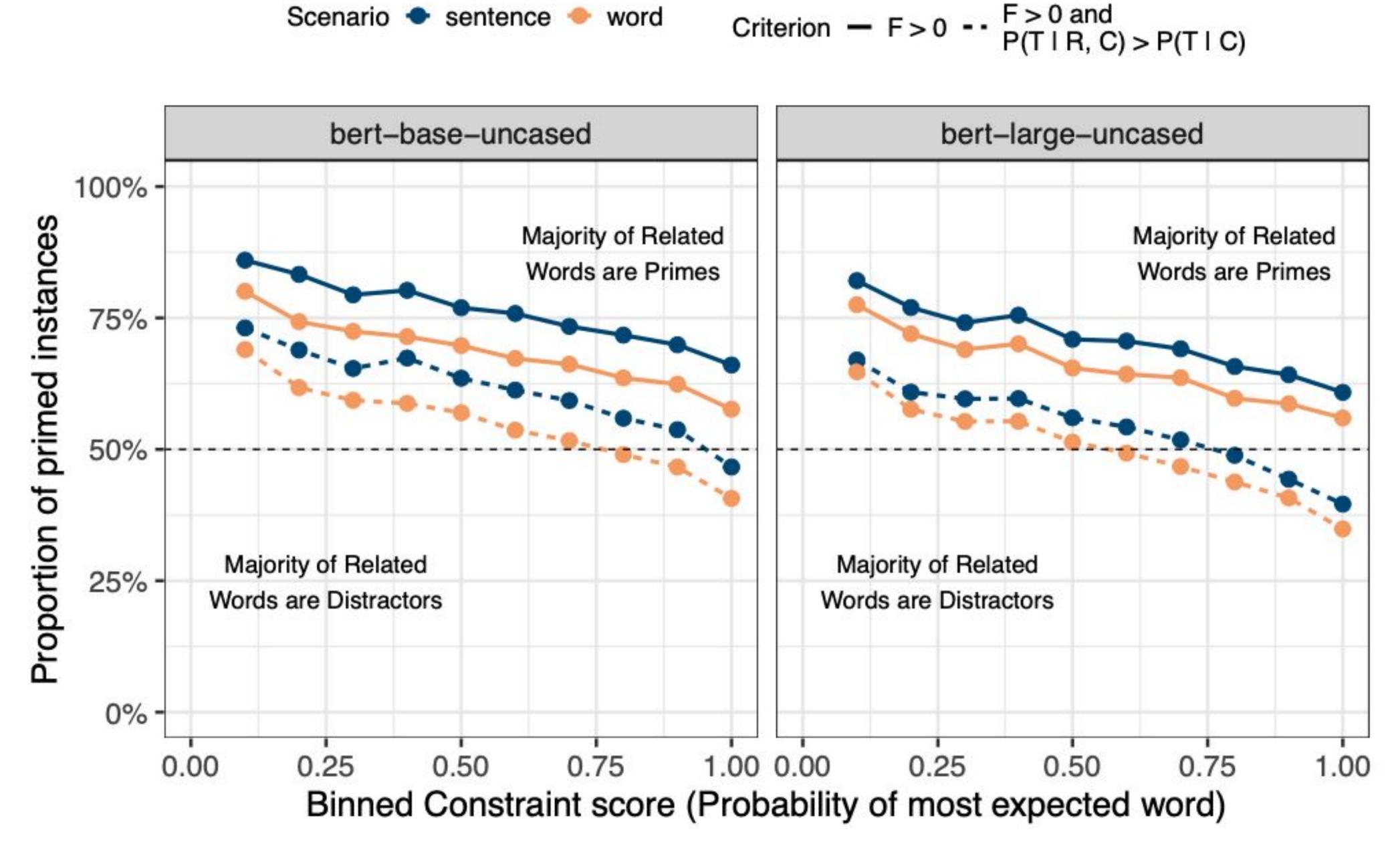
BERT has to rely only on the prime word to

$$[\mathrm{MASK}] = x \mid \mathcal{C}),$$

2016). 11 contexts per triple (10 constraint bins and a "zero-constraint" context)



- presence of an unrelated word.



- decreases.
- other relations.



Distraction Effects

• As the constraint imposed by the context increases, we see more instances in which the probability of the target word in presence of the related word is in fact less than that in

• We observe that in highly constraining cases the probability of the target word in presence of the related word is in fact also lower than that in an un-primed context. • In such cases, the related word acts like a "distractor" rather than a prime.

• We make our criterion for priming more stringent and count an instance as "primed" if: F > 0 and P(T | R, C) > P(T | C).

Figure 2: Proportion of instances that show priming under more (dashed) and less (solid) stringent priming criteria.

Takeaways

• **BERT shows priming:** BERT is reliably sensitive to single word lexical cues, but this effect is localized to minimally constraining contexts.

• Relationship with Constraint: As the amount of constraint posed on masked token by the context increases, the information provided to BERT by individual lexical cues

• Priming across Lexical Relations: In highly unconstraining contexts, BERT shows robust priming behavior for the lexical relations of synonymy, category, and antonymy, than

• **Priming Distraction:** In strongly constraining contexts, BERT is increasingly distracted by related primes, actively demoting the expectation of the target words.

Selected References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. Behavior Research Methods

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., ... & Allen, J. (2016, June). A corpus and cloze evaluation for deeper understanding of commonsense stories. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. Journal of Experimental Psychology:

Timothy P McNamara. 2005. Semantic priming: Perspectives from memory and word recognition. Psychology Press. 45(4), 1099-1114.

Technologies (pp. 839-849).

Learning, Memory, and Cognition, 14(2), 344.