Exploring Lexical Relations in BERT using Semantic Priming

Kanishka Misra[1], Allyson Ettinger[2], Julia Taylor Rayz[1]
[1]Purdue University, [2]University of Chicago

Watch the video: cutt.ly/bert-priming /

AKRaNLU

PURDUE UNIVERSITY

THE UNIVERSITY OF CHICAGO

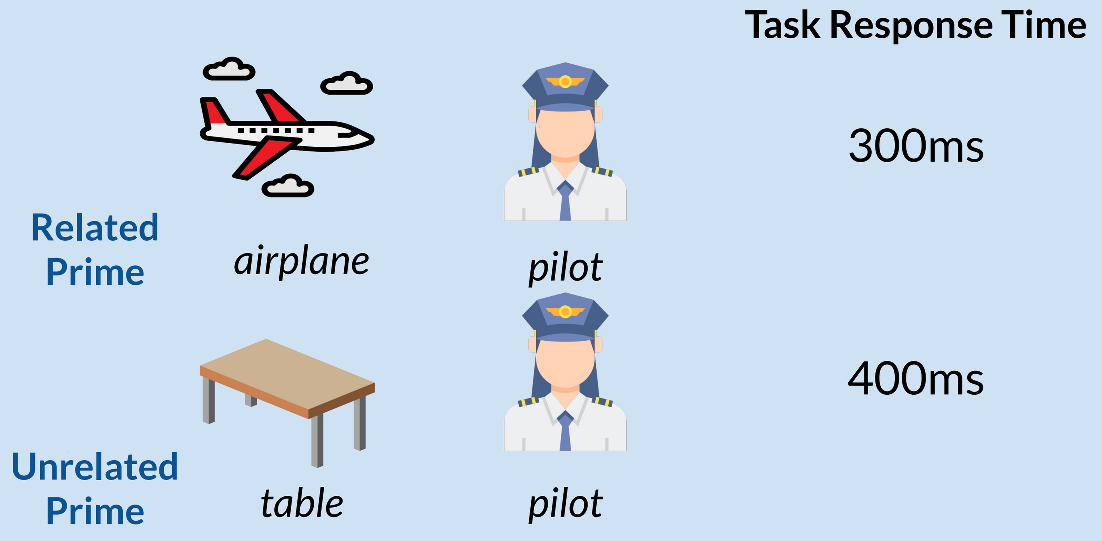
INTRODUCTION

- Pretrained language processing models that estimate word probabilities in context have become ubiquitous in natural language processing (NLP)
- How do these models use **lexical cues in context** in order to inform their word probabilities?
- We present a case study by analyzing BERT (Devlin et al., 2019), a recent pre-trained model, using English lexical items that show **semantic priming** in humans.

BACKGROUND

Semantic Priming

- Response to stimulus is faster when it is preceded by a semantically related word as compared to a semantically unrelated word.



Bidirectional Encoder Representations from Transformers (BERT)

- Deep bi-directional transformer (Vaswani et al., 2017) trained on pairs of sentences from Wikipedia and BookCorpus, using:
  - **Masked Language Model objective:** *predict masked words in sentences.*
    *Oh, I love coffee! I take coffee with **[MASK]** and sugar.*

    Top-5 predictions (with probability): ***cream*** *(0.66), milk (0.15), cinnamon (0.06), sugar (0.02), honey (0.01)*
  - **Next Sentence Prediction objective:** *predict whether the second sentence follows the first sentence.*
- We use two models, differing in number of parameters - BERT-base (110M) and BERT-large (340M)

Contact: kmisra@purdue.edu, twitter: @kanishkamisra

METHOD - BERT AS A PRIMING SUBJECT

● Unrelated Prime (U) ● Related Prime (R) ● Context (C)

negotiation. We made an [MASK].
fancy. We made an [MASK].

P([MASK] = ***agreement***) = ?

Target (T)

Measuring Priming in BERT = Facilitation (F)

F = *surprisal*(T | U, C) - *surprisal*(T | R, C)

surprisal = -log P(w | C)

*Measures how "surprised" is BERT in encountering a word, **w** in context, **C***

*An instance, (**T**, **R**, **U**, **C**) shows priming in BERT **if F > 0**, i.e., BERT is **more surprised** to encounter **T** in a context containing an unrelated word, (**U**, **C**), than in a context containing a related word, (**R**, **C**).*

Contextual Constraints

We Investigate patterns of priming in BERT under **differing predictive constraints**, motivated by sentence priming study by Schwanenflugel and LaCount (1988; see video for details)

[MASK] = *key*

● Low Constraint ● High Constraint

*He lost the **[MASK]** yesterday.*
*She opened the door using the **[MASK]**.*

Continuous Measure of Constraint $\max_{x \in \mathcal{V}} \mathrm{P}_{\mathrm{BERT}}([\mathrm{MASK}] = x)$

Averaged over both BERT models

Also test on a ***neutral context*** *: the last word of this sentence is **[MASK]**.*

- Derived word-to-word from Schwanenflugel and LaCount(1988).
- **Expected to show minimum constraint**

Expectation: *When context is highly constraining, the addition of a related word will not provide sufficient information beyond what is already provided, i.e., **BERT will show less priming in high constraint contexts as compared to that in low-constraint ones.***

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 containing target words, C, sampled from the **ROCstories corpus** (Mostafazadeh et al., 2016). **Processing:** target words replaced with the [MASK] token.
- Calculate constraint scores for all contexts, grouped into 10 equal bins of width 0.1. **Example:** a constraint score of 0.34 will be in the 4th bin. Keep one context per constraint bin per triple.
- Total instances: **23232**, with 2112 unique triples, 10 constraint bins and an additional "neutral" context.

RESULTS

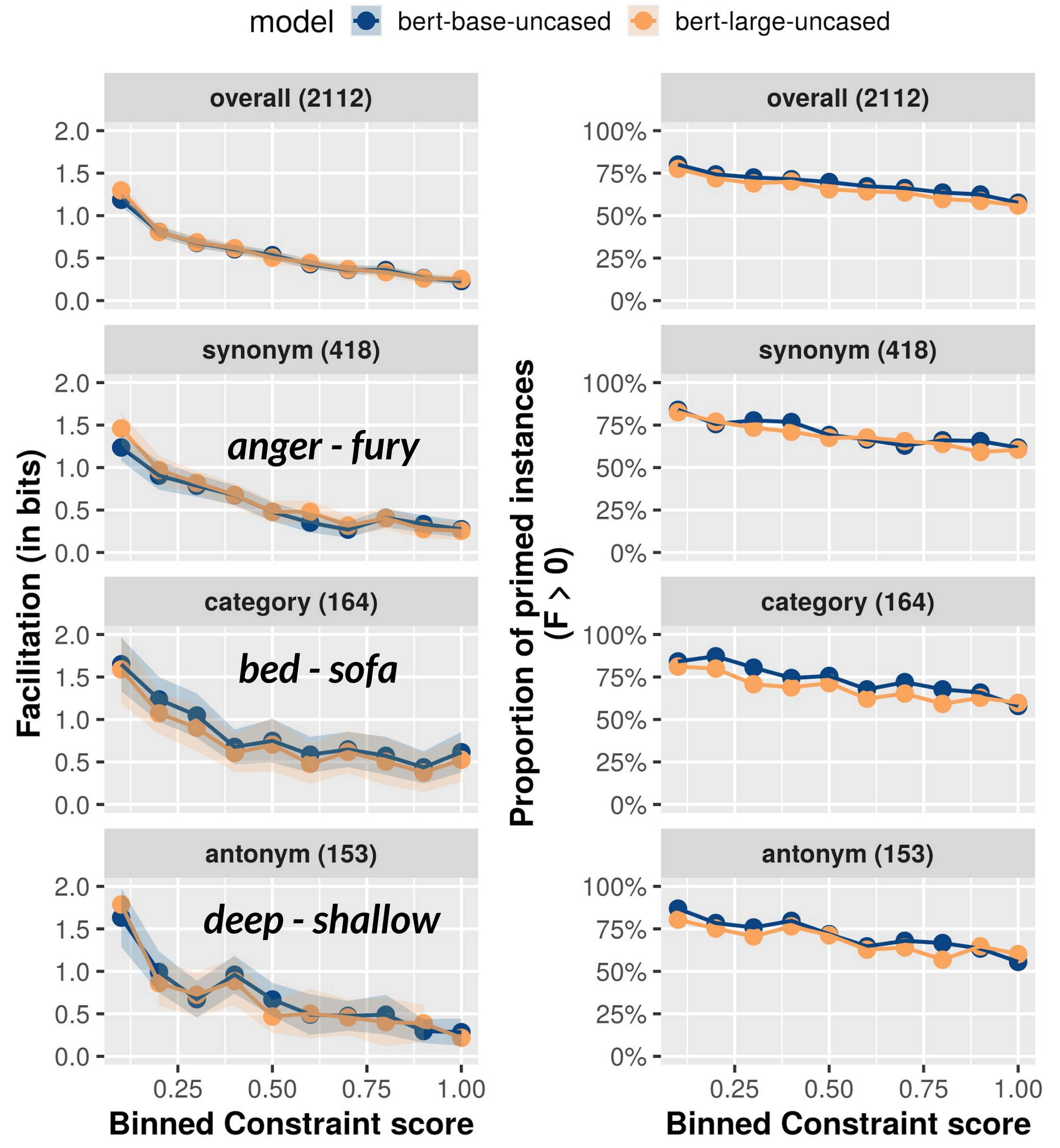


Figure 1: Facilitations (left) and proportion of primed instances (right) across the top-3 lexical relations, along with overall results (first row). The x-axis denotes binned constraint scores (0.1-1.0)

RESULTS (CONTD.)

Table 1: Facilitation and proportion of primed instances for **neutral contexts** (minimal constraint), results with valid constraint bins shown in Figure 1.

Relation / Dataset	N	BERT-base		BERT-large	
		Facilitation	Primed Instances	Facilitation	Primed Instances
overall	2112	2.69 ± 0.11	85.20%	5.14 ± 0.16	91.30%
synonym	418	3.36 ± 0.27	90.20%	6.41 ± 0.36	95.90%
category	164	3.90 ± 0.47	92.70%	7.01 ± 0.54	97.60%
antonym	153	4.68 ± 0.47	93.50%	6.97 ± 0.57	98.00%

NOTE: For detailed priming results on more lexical relations, please refer to the supplemental materials (attached on the "poster stand")

DISCUSSION AND TAKEAWAYS

- **BERT shows priming:** *BERT is reliably sensitive to single word lexical cues, but this effect is localized to minimally constraining contexts (neutral and low constraint contexts show largest facilitation values and most primed instances.)*
- **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 decreases.*
- **Priming in Lexical Relations:** *In highly unconstraining contexts, BERT shows greater priming behavior for the lexical relations of synonymy, category, and antonymy, than other relations (see suppl. materials for full results)*

REFERENCES

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008).

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, 45*(4), 1099-1114.

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 Technologies* (pp. 839-849).

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(2), 344.