Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming

Kanishka Misra[1] **Allyson Ettinger**[2] **Julia Taylor Rayz**[1]
[1]Department of Computer and Information Technology, Purdue University
[2]Department of Linguistics, University of Chicago
[1]{kmisra, jtaylor1}@purdue.edu, [2]aettinger@uchicago.edu

Abstract

Models trained to estimate word probabilities in context have become ubiquitous in natural language processing. How do these models use lexical cues in context to inform their word probabilities? To answer this question, we present a case study analyzing the pre-trained BERT model with tests informed by semantic priming. Using English lexical stimuli that show priming in humans, we find that BERT too shows "priming," predicting a word with greater probability when the context includes a related word versus an unrelated one. This effect decreases as the amount of information provided by the context increases. Follow-up analysis shows BERT to be increasingly distracted by related prime words as context becomes more informative, assigning *lower* probabilities to related words. Our findings highlight the importance of considering contextual constraint effects when studying word prediction in these models, and highlight possible parallels with human processing.

1 Introduction

The field of natural language processing (NLP) has recently seen a dramatic shift toward the use of language model (LM)-based pre-training (Howard and Ruder, 2018; Peters et al., 2018)—training based on estimating word probabilities in context—as a foundation for learning of a wide range of tasks. Leading this charge was the BERT model (Devlin et al., 2019), which is optimized in part to use context information to predict masked words. Because of the impressively strong performance of BERT and its successors (Yang et al., 2019; Liu et al., 2019; Clark et al., 2020), there has been increasing need for understanding how these types of models work, and what linguistic properties LM-based pre-training confers upon them.

In this paper, we focus on the question of how BERT uses individual lexical relations to inform word probabilities in context. For example, if a word like *airplane* is prepended to (1a), to what extent does this increase the model's probability for the word *pilot* in the blank position in (1b)?

(1) a. I want to become a ____.
b. *airplane.* I want to become a ____.

This question is particularly relevant because human brains show a robust phenomenon of *semantic priming* (McNamara, 2005), in which the presence of a word such as "airplane" will give rise to faster reactions to a related word like "pilot". We explore whether the same lexical relations that show priming in humans will also be utilized by BERT to influence word predictions in context.

Our analysis includes three experiments. First, we test BERT's sensitivity to single-word lexical cues for word prediction in context, using word pairs that show priming in humans, and testing for influence of contextual constraint. We find clear priming in BERT, but this effect is primarily localized to contexts that are relatively unconstraining. Next, we examine how BERT's use of these lexical cues varies depending on the type of lexical relation. We find that certain relations—particularly antonymy, synonymy, and category relations—evoke more sensitivity in BERT than others. Finally, we take a closer look at lexical cue dynamics in cases of high-constraint contexts, and we find that in such contexts we often see a phenomenon of "distraction" rather than priming, such that related words actively demote probabilities of counterpart target words.

Our paper has two main contributions. First, we introduce a methodology for fine-grained exploration of lexical cue sensitivity in predictive models, grounded in lexical relation phenomena observed in humans. Second, we apply these methods to shed light on word prediction dynamics of the BERT model. We discuss implications of these

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4625–4635
November 16 - 20, 2020.

findings for considerations of contextual constraint, and for parallels with human processing. We release our datasets and code for further testing.[1]

2 BERT as a Semantic Priming Subject

2.1 Semantic Priming

To study BERT's sensitivity to single-word cues in context, we draw on data from *semantic priming* observed in humans. Semantic priming is an experimental phenomenon widely studied in psycholinguistics, in which participants show a speedup in response to a word stimulus during language tasks when the response is preceded by a semantically related word as opposed to an unrelated one (McNamara, 2005). Participants perform tasks like pronouncing a word out loud ("naming") or deciding whether a given string is a word or not ("lexical decision"). The word to which the response is made is referred to as the *target* and the preceding stimuli are called *primes* (either related or unrelated). Levels of priming are evaluated based on participants' response times (RT). The magnitude of the speedup in RT provides information about the strength of the lexical relation in the context of the participants' cognitive system. The stimuli used in semantic priming experiments elicit responses caused by implicit processing within humans, which makes them an ideal intrinsic testing ground for studying models' quantification of word relations. Leveraging this fact, we take word pairs that show priming in humans, and use them to test BERT's sensitivity to lexical cues that have various types of relations.

2.2 Extending Semantic Priming to BERT

In humans, semantic priming occurs due to the presence of a lexical associate that affects the speed of response to a stimulus. Analogously, we are interested in learning how BERT's behavior (defined as a change in word probability) is affected by a lexical cue present in its input context. We define semantic priming in BERT as an increase in the model's expectation for a target word (or a lack thereof) in a given context in the presence of a semantically related word as compared to an unrelated one. Consider the following example:

(2) a. I want to become a ___.
b. *airplane.* I want to become a ___.
c. *table.* I want to become a ___.

If the probability of the target word, *pilot* is greater in (2b) as compared to that in (2c), then we interpret that the related word (*airplane*) primes BERT more than the unrelated word (*table*) does, for the target *pilot* in the context (2a). Such a test ensures that the only difference in BERT's output for the blank position in both cases is due to the swapping of the primes, allowing us to infer the degree to which BERT relies on single word cues to inform its probability for the target word. Importantly, our work here is not trying to simulate human semantic priming experiments directly—the structure of our tests is adapted for BERT's conventional usage by placing words in context, and thus deviates from standard word-level priming structure.

2.3 Predictive Constraints of Target Contexts

We test how BERT's sensitivity to individual prime words varies based on contextual constraints. Consider the following example for target word *key*:

(3) a. He lost the ___yesterday.
b. She opened the door using a ___.

In (3a), the blank position can be any word that satisfies the semantic role THEME-OF for the event LOSE. The blank position is far more constrained in (3b), which requires a word that satisfies the semantic role INSTRUMENT-OF for the event UNLOCK-DOOR—a set limited to items such as *key, lock-pick,* or perhaps *screwdriver*. As a result, the sentence in (3b) is highly constraining towards predicting a word denoting these concepts or their relatives.

Focusing on how the constraint imposed by the context affects our notion of priming allows us to explore how much more information about the target word, *key*, prepending a related word like *lock* can provide in a high-constraint context such as (3b), beyond words such as "open" and "door". We can then compare priming behavior when *lock* is prepended to (3a), which imposes fewer constraints on the blank position.

Our focus on contextual constraints is in part motivated by studies that use sentence contexts of varying constraint to study priming in humans. In particular, Schwanenflugel and LaCount (1988) found low-constraint contexts to show wider scope of facilitation in lexical decision tasks, as compared to high-constraint ones, which only showed facilitations for the best completions (highest cloze probability). That is, low-constraint contexts produced enhanced facilitation effects in cases when

[1]Data and code available at https://github.com/kanishkamisra/emnlp-bert-priming

the target word has low probability in the context. Taking this into account, when the context is highly constrained towards a particular completion, we expect BERT to show less sensitivity to the presence of an additional lexical cue, which may not provide significant information over and above that of the already constraining context. We hypothesize that in low-constraint contexts, because every word (including the target) is a low probability completion, BERT will be more sensitive to the addition of a single word in the context, thus showing greater priming effects in our testing framework.

3 Related Work

By focusing on the aforementioned considerations, and borrowing from the semantic priming paradigm, we build on a growing precedent of using psycholinguistics-inspired tests which focus on discovering the underlying mechanisms and linguistic competence of neural network based models, and how closely they approximate language processing phenomena observed in humans. For example, syntactic phenomena have been studied within recurrent neural network (RNN) LMs by supplying controlled, hand-crafted inputs to compare word probabilities in context across syntactically correct and anomalous instances (Futrell et al., 2019). This methodology has been applied to study subject-verb agreement (Linzen et al., 2016; Gulordava et al., 2018), garden-path effects (van Schijndel and Linzen, 2018; Frank and Hoeks, 2019; Futrell et al., 2019), and filler-gap dependencies (Wilcox et al., 2018). Deviating from prior work that has predominantly focused on investigating syntactic phenomena in LMs, Ettinger (2020) investigates BERT's semantic and pragmatic inference knowledge by using stimuli from N400 experiments (Kutas and Hillyard, 1980). The findings suggested that BERT accurately attributes nouns to their hypernyms, but struggles in presence of negation, highlighting a limitation of LM-based training objectives.

Syntactic Priming in LMs Prasad et al. (2019) draw on the syntactic priming paradigm—priming observed for sentence structure rather than word association—to investigate the ability of LMs to represent syntactic regularities. They define priming as adaptation to new stimuli by fine-tuning models on similarly structured sentences using the language model objective and investigating cumulative sentence surprisals before and after adaptation. In addition to focusing on a different type of priming, our work differs in operating directly on pre-trained BERT, without relying on any fine-tuning, which allows us to investigate the outcomes of the model's pre-training process itself.

Mispriming in LMs Building upon work by Petroni et al. (2019), which queries LMs by analyzing output over knowledge base queries recast as cloze questions, Kassner and Schütze (2020) introduce the "mispriming" probe, which shows BERT to be easily distracted by misprimes—words chosen to be prepended to cloze-like sentences. For instance, BERT-large predicts *Cicero* as the completion in place of the correct answer, *Plato*, when the previous query is modified to "*Cicero? Platonism is named after [MASK].*" While their setup is similar to the one discussed in this paper, our work differs methodologically in two ways: 1) we base our experiments on word pairs with clear, cognitively-based lexical relationships, for which we can explore fine-grained relation differences, and 2) we compare related to unrelated primes (rather than comparing primed to unprimed contexts, as do Kassner and Schütze (2020)), thus keeping constant the prepending of a word, so as to target lexical relation effects more precisely. Furthermore, in the present work we are focused additionally on the effects of contextual constraint on BERT's lexical sensitivity during inference.

4 Methods

4.1 Model Investigated: BERT

BERT (Devlin et al., 2019) is a deep bidirectional transformer (Vaswani et al., 2017) network, trained on pairs of sentences. It is pre-trained on: (1) the Masked Language Model objective (predicting missing words in context), and (2) the Next Sentence Prediction objective (predicting whether the first sentence of the pair follows the second). We test on two variants: BERT-base (110M parameters) and BERT-large (340M parameters).

4.2 Data

We use the Semantic Priming Project (SPP) (Hutchison et al., 2013) as our source of human priming experiment data. This resource has previously been used to evaluate word embedding models such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) by measuring the amount of variance in priming response times ex-

plained by cosine similarity between words as a predictor (Ettinger and Linzen, 2016; Auguste et al., 2017). The SPP is a large collection of priming data for 768 subjects for 3322 triples, represented as ($\mathcal{T}$, $\mathcal{R}, \mathcal{U}$), where $\mathcal{T}$ is the target word, and $\mathcal{R}$ and $\mathcal{U}$ are the related and unrelated primes, respectively. To enable fair comparison, we filter out target words that do not occur in BERT's vocabulary, as well as instances in which some of the RTs were missing, leaving us with 92% of the total triples (n = 3058).

Stimulus Construction In addition to the SPP triples, we introduce another component to accommodate the nature of the BERT model: a context, $\mathcal{C}$, which is a naturally-occurring sentence originally containing the target word $\mathcal{T}$, now with $\mathcal{T}$ replaced by the "[MASK]" token. We test the model's expectation for $\mathcal{T}$ in the masked position when $\mathcal{C}$ is preceded by a related prime $\mathcal{R}$, as well as when it is preceded by an unrelated prime $\mathcal{U}$, denoted as $(\mathcal{R}, \mathcal{C})$ and $(\mathcal{U}, \mathcal{C})$ respectively. We choose to embed $\mathcal{T}$ in $\mathcal{C}$ in order to better simulate BERT's standard usage, given that the model is pre-trained to predict words in sentence contexts. We choose the contexts $\mathcal{C}$ to be naturally-occurring sentences, since BERT is trained on well-formed sentences that affect its word level expectation. Our target contexts are sampled from the concatenation of the ROCstories Corpus (Mostafazadeh et al., 2016), and the train and test sets used in the "Story Cloze Test" task (Mostafazadeh et al., 2017), primarily due to the simplistic nature of the sentences.

For our prime contexts, we experiment with two scenarios: **(a) WORD**: where the prime word, followed by a period, '.' is prepended to the target context, and **(b) SENTENCE:** where a neutral context, *"the next word is "* followed by the prime word and a '.', is prepended to the target context. We add the [CLS] and [SEP] tokens at the beginning and the end of each stimulus, respectively, following previous studies with a similar setup. Table 1 shows full example items from these different settings. We limit to single word or neutral sentence contexts for our prime words because any naturalistic sentence containing $\mathcal{R}$ would be different from that containing $\mathcal{U}$, thus adding imbalanced noise from the non-prime words. The context $\mathcal{C}$ for the target, by contrast, will remain constant given that the target is constant (for any pair of primes).

Contextual Constraints We analyze BERT's reliance on single-word lexical cues (our primes) to inform its target word probabilities under various predictive constraints placed on the [MASK] token. To compute constraint of a context, we take the most expected words under BERT-base and BERT-large, and average their probabilities. This effectively represents how predictable the masked word is in the un-primed context. Our notion of constraint is grounded in psycholinguistic studies examining effects of sentence contexts (Schwanenflugel and LaCount, 1988; Federmeier and Kutas, 1999), which estimate sentence constraint based on the cloze probability of the most expected word in context. Mathematically, the constraint of a context $\mathcal{C}$ is defined as:

$$\text{constraint}(\mathcal{C}) = \frac{1}{2} \sum_{m \in \{b,l\}} \max_{x \in \mathcal{V}} P_m([\text{MASK}] = x \mid \mathcal{C}),$$

where P_m represents the probability distribution for [MASK] in the output of the BERT model, either base (b) or large (l), and x is a token belonging to BERT's vocabulary, $\mathcal{V}$. Our proposed constraint scores are thus bounded by $[0, 1]$. We calculate the constraint for all sentences in our corpus that contain the target words, and group them into 10 equal bins of width 0.1 each, i.e, a constraint score of 0.38 would be in bin 4. Additionally, as a control, we also use a synthetic and unconstraining target context that we refer to as neutral[2]: "[CLS] the last word of this sentence is [MASK]. [SEP]". This neutral context provides the lowest constraint, as it contains no information about what the masked target word can be—any word in BERT's vocabulary can fit in its [MASK] position. To make robust conclusions about the effect of constraint, we only sample triples that have at least one target context in each of the 10 bins. We faced polysemy issues

Scenario	Stimulus
WORD	[CLS] *airplane.* I wanted to become a [MASK]. [SEP] [CLS] *table.* I wanted to become a [MASK]. [SEP]
SENTENCE	[CLS] *The next word is airplane.* I wanted to become a [MASK]. [SEP] [CLS] *The next word is table.* I wanted to become a [MASK]. [SEP]

Table 1: Example Stimuli, with prime contexts in italics. Here, $\mathcal{T}$ = *pilot*, $\mathcal{R}$ = *airplane*, and $\mathcal{U}$ = *table*.

[2]Our choice of neutral prime context follows Schwanenflugel and LaCount (1988).

for 72 target words, in which the sense of the target in the originally sampled $\mathcal{C}$ did not fit the lexical relation with the primes—we manually corrected these by re-selecting appropriate contexts from the corpus. We could not resolve this issue for 28 items, which we discarded. This further reduces the number of unique triples to 2112 (69% of the valid instances), with each triple being associated with 11 (10 bins and a neutral context) stimuli.

Constraint Scores and Entropy While we follow psycholinguistic precedent in defining contextual constraint based on the highest-probability completion of a given context, another obvious candidate for defining contextual constraint would be the entropy of the probability distribution for the [MASK] token. In this setting, the entropy would quantify the amount of uncertainty about the [MASK] token when conditioning on the context: low-constraint contexts would produce high uncertainty, and therefore a high entropy value, while high-constraint contexts would produce lower entropy values. To establish the consistency of our chosen constraint measure with an entropy-based definition of constraint, for every context ($\mathcal{C}$) in our experiments we compute the entropy of the probability distribution on the [MASK] token, averaging the entropies from the two BERT models (*b, l*):

$$H_{\text{constraint}}(\mathcal{C}) = -\frac{1}{2}\sum_{m\in\{b,l\}}\sum_{x\in\mathcal{V}} P_m(x \mid \mathcal{C}) \log P_m(x \mid \mathcal{C})$$

The Pearson correlation between our constraint measure and $H_{\text{constraint}}(\mathcal{C})$ is -0.89, indicating a strong empirical relationship between constraint measured as the probability of the best completion and entropy of the predicted distribution.

4.3 Measuring Priming in BERT

We use **surprisal** as our measure of the model's expectation for $\mathcal{T}$ in the given context. The surprisal of a language model denotes the level of "surprise" of the model for a word w, in context $\mathcal{C}$:

$$\textit{Surp}(w \mid \mathcal{C}) = -\log_2 P(w \mid h_{\mathcal{C}}),$$

where $h_{\mathcal{C}}$ is the hidden state of the model for the context. Surprisal is an effective linking hypothesis between language model probabilities and measures of human language processing. For instance, surprisal derived from n-gram and RNN LMs was shown to be a significant predictor of (1) self-paced reading times, a measure of cognitive load incurred during sentence comprehension in humans (Hale, 2001; Levy, 2008; Smith and Levy, 2013); and (2) the amplitude of the N400 event related potential (ERP) (Frank et al., 2013), an electrical response that corresponds to lexical and semantic processing in human brains (Kutas and Hillyard, 1980).

In our experiments, we define the level of priming in BERT, which we call "Facilitation", as:

$$\mathbb{F} = \textit{Surp}(\mathcal{T} \mid \mathcal{U}, \mathcal{C}) - \textit{Surp}(\mathcal{T} \mid \mathcal{R}, \mathcal{C}).$$

Due to the setup of our stimuli, the difference in BERT's surprisals for the target word $\mathcal{T}$ between the context pairs (related vs. unrelated) quantifies the degree to which the model is influenced by one isolated prime word over the other. This can be considered analogous to the difference in human response times in the context of related versus unrelated primes, reflecting differing strengths of lexical association between the prime and target words. If BERT is sensitive to the presence of a related prime, as humans are, such that $\mathcal{R}$ primes the model to predict $\mathcal{T}$ more than $\mathcal{U}$ does, then BERT should show less "surprise"—i.e., produce higher probability—for $\mathcal{T}$ in the context $(\mathcal{R}, \mathcal{C})$, than in $(\mathcal{U}, \mathcal{C})$. In such cases, $\mathbb{F}$ will be positive.

5 Analysis and Results

To test for statistical significance between facilitation in BERT and the contextual constraints imposed by stimuli, we use a linear mixed-effects model with constraint scores as fixed effects and include random intercepts for target words. The pre-trained BERT models were accessed using the Transformers library (Wolf et al., 2019).

5.1 How Facilitation is affected by Constraint

Figure 1 shows the average facilitation effects and proportion of instances showing facilitation, for both models in each prime context setting.

Overall, we find that priming in BERT decreases as the predictive constraint placed on the [MASK] position increases. This is evidenced by decrease in both the facilitation effect ($p < .001$ for both models in both scenarios),[3] as well as the decrease in raw proportion of instances in which the facilitation was positive. This indicates that the information provided by the related prime word (relative to the

[3] While we plot facilitation against binned constraint scores, the significance was derived using raw constraint value as a predictor in the linear mixed effects model.

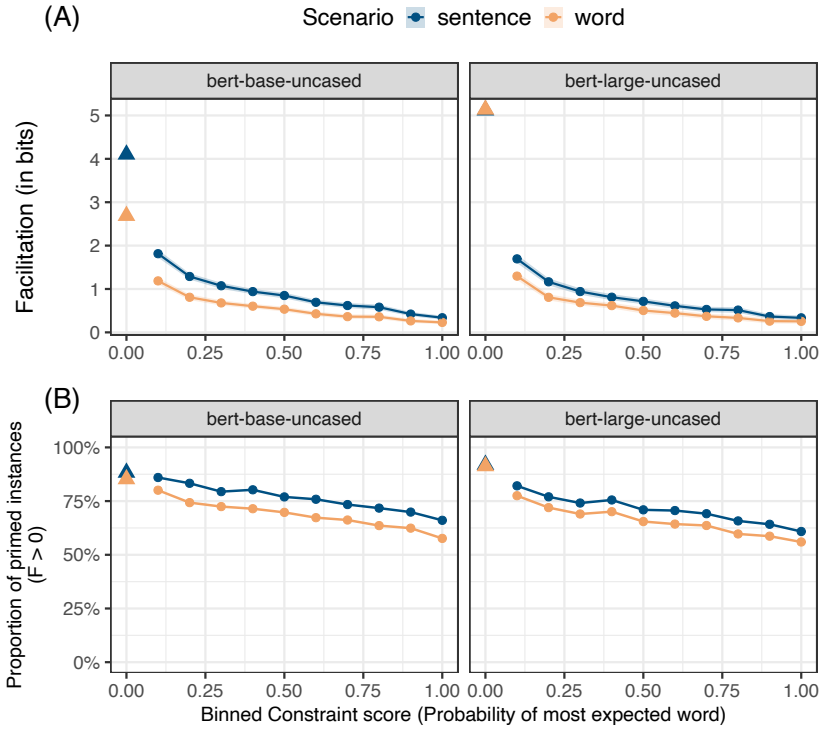

Figure 1: Average facilitation (A) and proportion of primed instances, i.e., $\mathbb{F} > 0$ (B) vs. binned constraint score. Error bands represent 95% confidence intervals. **Note:** Results for neutral contexts are shown separately as **triangles** (▲, ▲), and do not correspond to a constraint score of 0.0 (actual constraint score = 0.02).

unrelated one) is increasingly outweighed by the information provided by the predictive constraints as the level of constraint increases. At lower levels of contextual constraint, BERT takes substantially more advantage of the lexical association of the prime word to predict the target word. This is particularly apparent in neutral contexts, where BERT receives almost no context information from non-prime words, and shows considerably larger facilitation. Comparing settings with and without sentence context for the prime word, we see that BERT consistently shows greater facilitation effects when the prime context is a sentence rather than a single word, across every constraint bin ($p <$ .001), with the exception of BERT-large for neutral contexts, where the magnitudes of the facilitation are the largest (as shown in Figure 1), but not significantly different between sentence and word prime contexts ($t(2111)$ = -0.3402, $p = 0.6331$).

5.2 Facilitation across Lexical Relations

We have established above that BERT's predictions are sensitive to the addition of single related words in the context, particularly in contexts that are weakly constraining. In this section we investigate whether these sensitivity patterns are consistent across different types of lexical relations between the related prime and the target. We test priming effects for the 10 most frequent lexical relations annotated in the SPP, examples of which are shown in Table 2. As in section 5.1, we test how facilitation changes with contextual constraints.

Relation	n	$\mathcal{T}, \mathcal{R}$
Synonym	418	*anger, fury*
Forward Phrasal Associate	263	*ache, stomach*
Category	164	*bed, sofa*
Antonym	153	*deep, shallow*
Backward Phrasal Associate	151	*cause, effect*
Supraordinate	131	*spaghetti, pasta*
Script	124	*judge, court*
Perceptual property	90	*leaf, tree*
Functional property	73	*bell, ring*
Instrument	35	*bow, arrow*

Table 2: Top-10 relations within our subset of SPP.

Figure 2 shows facilitation effects, averaged for BERT-base and BERT-large. We find facilitation effects across our subset of lexical relations to be consistent with the results in section 5.1—facilitation decreases as the contextual constraint increases ($p <$ .001 across all lexical relations and prime con-

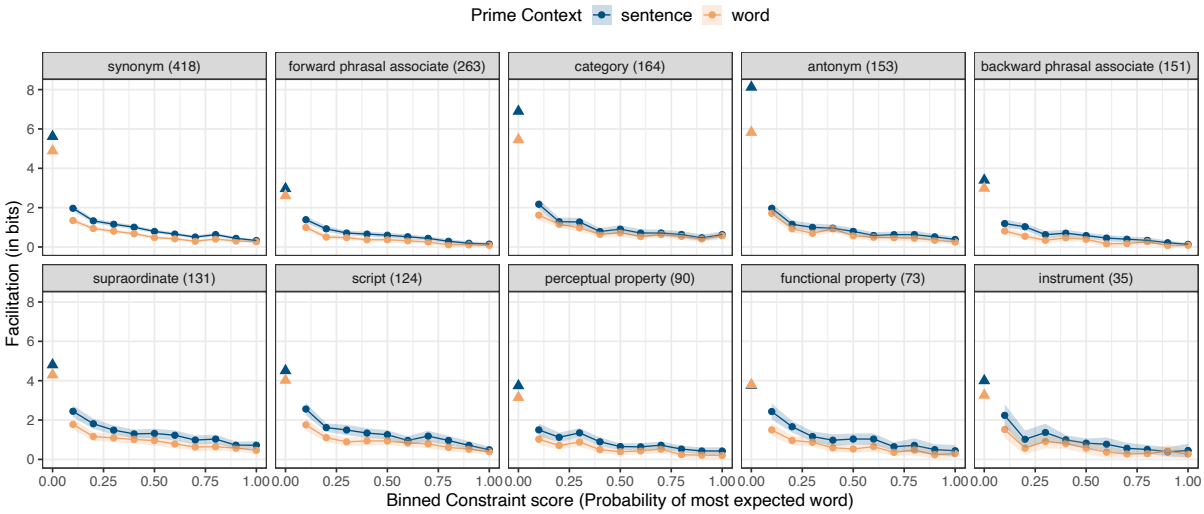

Figure 2: Facilitation effects across top 10 lexical relations in our subset of SPP (averaged for BERT-base and BERT-large). Error bands represent 95% confidence intervals. **Note:** Results for neutral contexts are shown separately as **triangles** (▲, ▲), and do not correspond to a constraint score of 0.0 (actual constraint score = 0.02).

text scenarios). Among different lexical relations, we see the largest variation in BERT's sensitivity on the lower constraint items, which impose fewer restrictions on the identity of [MASK]. Synonymy, category, and antonymy relations show the most pronounced differences, with BERT showing considerably larger facilitation in the neutral context than for other relations. This suggests that BERT's word predictions in context may be more strongly attuned to relations of synonymy, category membership, and antonymy than to other lexical relations.

5.3 On Primes and Distractors

The preceding results show a decrease in number of primed instances as contextual constraint increases. This means that 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 *less* than that in presence of an unrelated word. For example, the first row of Table 3 shows an instance for a target, *bacon*, with a constraint score of 0.89 (i.e., the 9th bin). Contrary to priming patterns observed in low-constraint contexts, the probability of *bacon* is quite low when BERT is primed by *pork*, and very high when the unrelated word, *meteorite*, is the prime. Here, the related prime acts as a **distractor**,[4] similar to the **mispriming** reported in Kassner and Schütze (2020). Upon further investigation, we observe that

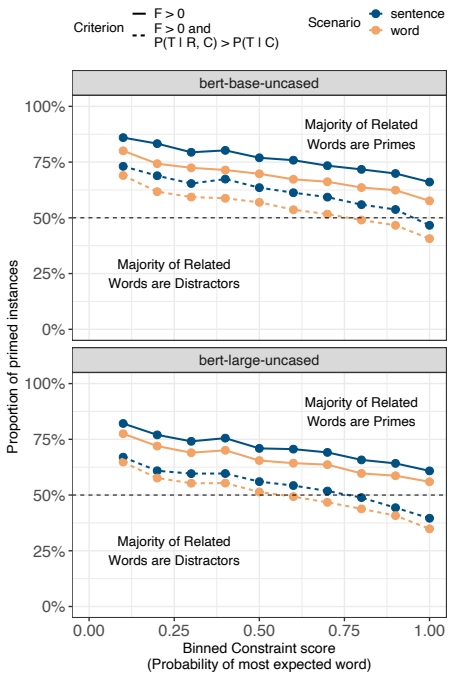

Figure 3: Proportion of primed instances under more (dashed) and less (solid) stringent priming criteria.

the probability of the target word in presence of the related word is in fact also lower than that in an un-primed context, i.e., $P(\mathcal{T} \mid \mathcal{R}, \mathcal{C}) < P(\mathcal{T} \mid \mathcal{C})$. The related word "distracts" BERT, thereby reducing the probability of the target. To account for such cases, we make our criterion more stringent and count an instance as "primed" if the facilitation is positive ($\mathbb{F} > 0$) **and** if the presence of the related word increases the probability of the target

[4]We refer to it as a distractor since the target word is not the absolute correct completion for our contexts, since they are not factual like in Kassner and Schütze (2020).

Target (Constraint)	(*Related / Unrelated*) Context	Top 5 Predicted Words (BERT-large probability)	
		Primed by Related	Primed by Unrelated
bacon (0.89)	(*pork/meteorite*). she cooked up some eggs, [MASK], and toast.	*eggs (0.20), potatoes (0.04), tea (0.04), pancakes (0.04), cheese (0.03)*	*bacon (0.78), sausage (0.06), ham (0.03), pancakes (0.02) toast (0.02)*
painting (0.75)	(*drawing/champagne*). dana was a young artist who spent many hours a day [MASK].	*drawing (0.88), painting (0.10), studying (<0.01), writing (<0.01), practicing (<0.01)*	*painting (0.79), drawing (0.06), working (0.03), studying (0.03), teaching (0.01)*

Table 3: Example high constraint instances that show "distraction" rather than priming in BERT-large.

over that in the un-primed instance ($P(\mathcal{T} \mid \mathcal{R}, \mathcal{C}) > P(\mathcal{T} \mid \mathcal{C})$). These changes are reflected in Figure 3.

The proportion of facilitatory instances is now substantially lower with this more robust notion of priming, but it follows the same pattern observed when only facilitation score was considered. At higher constraint scores, the proportions fall under 50%, giving us thresholds beyond which BERT shows more "distraction" from related prime words than facilitation. For example, starting at the 8th constraint bin, BERT-base shows priming only for 49% or fewer cases in the WORD prime context.

Qualitative Analysis We examine specific instances of model predictions in order to shed further light on the factors that contribute to BERT's distraction (as opposed to priming) effects. Table 3 shows two examples in which we observe such distraction patterns in BERT. In the example with *painting* as the target, we find BERT to show behavior akin to that discussed in Kassner and Schütze (2020). Here, the presence of a distractor (*drawing*), one that fits as a completion in the [MASK] position, leads BERT to predict the distractor with greater probability than the target (*painting*). However, in the example with *bacon* as the target, we observe a different kind of distraction: *pork* cannot replace *bacon* here as well as *drawing* can replace *painting* in the previous example, but *bacon* is still demoted in the probability distribution in favor of other foods related to *pork*. By contrast, in both examples the unrelated primes resemble "random misprimes" in Kassner and Schütze (2020): BERT isn't distracted by them—likely due to their degraded relevance to the context—and still predicts the target as the best completion.

6 General Discussion

In the experiments above, we show that when using word pairs informed by human semantic priming, the BERT model is reliably sensitive to individual lexical cues in its context—*if* the context is minimally constraining, such that there is little predictive information beyond that lexical cue. As the predictive constraint applied by the context increases, BERT's level of sensitivity to a given lexical cue decreases. These results suggest that BERT uses lexical cues as needed: when informative sentence cues are available, single lexical items are of less value, and so they exert less influence on BERT's expectations for a masked word.

Examining patterns across different types of lexical relations, we find that this general effect of constraint holds across relation types, but synonym, category, and antonym relations elicit larger lexical sensitivities in BERT, as compared to other relations (when the context is unconstraining). This suggests that BERT has identified these relations—or the particular words that share these relations—as being more reliably predictive. This may be because words sharing these relations are simply more likely to co-occur during BERT's training, or BERT may have formed higher-order relational associations that inform these sensitivities.

While we see that these priming-based lexical relations can have facilitatory effects on BERT's word predictions when the context is otherwise unconstraining, we see conversely that when the context *is* constraining, prime words can actually have a "distractor" effect—actively demoting the target word in the probability distribution. This finding builds on recent evidence of BERT's sensitivity to such distractions when predicting completions to factual queries (Kassner and Schütze, 2020). We find in our analyses that the nature of this distraction depends critically on the interaction of contextual constraint and the strength of the lexical relation: when the context is unconstraining, the probability of a word is likely to be promoted by a related lexical item more than by an unrelated lexical item. If the context is constraining, a related lexical item may demote the probability of a

target word in the predicted distribution, while an unrelated word is likely to have less impact.

The effectiveness of human priming pairs in influencing BERT's lexical sensitivities, as well as the impact of contextual constraint on BERT's use of lexical context cues, suggest possible parallels with mechanisms in human language processing. Not only do humans show priming with the same word pairs that we show to impact BERT's predictions here, but like BERT, humans also show more limited semantic priming in constraining contexts, and wider scope of priming in low-constraint contexts (Schwanenflugel and LaCount, 1988). This suggests that the mechanisms that dictate BERT's lexical sensitivity may be optimized in a manner—or at least to an outcome—comparable to those underlying priming effects in humans.

In practical terms, our results highlight the importance of contextual constraint in the dynamics of word prediction and information usage in the BERT model. Future work studying these dynamics should be mindful of this fact, as any observed prediction dynamics may change with the predictiveness of the context. This further emphasizes parallels with the study of human processing, as the predictive constraint of context has long been an important consideration and instrument in studying human sentence processing (Schwanenflugel and LaCount, 1988; Schwanenflugel, 1991; Federmeier and Kutas, 1999; McFalls and Schwanenflugel, 2002). Our results show a similarly important role played by the amount of constraint imposed on a masked word during word probability estimation, which can lead to substantially different outcomes in behavioral analysis of pre-trained models.

7 Conclusion and Future Work

In this paper, we presented a framework to test how BERT uses individual lexical relationships as cues for word prediction. Our framework is inspired by the psycholinguistic phenomenon of semantic priming, and our lexical cues are derived from a large collection of human priming experiments.

We examine the dynamics of BERT's word prediction in context, and relate its sensitivity towards lexical cues with contextual constraints and finer-grained lexical relations. Our findings establish the importance of considering predictive constraint effects of context in studies that behaviorally analyze language processing models, and highlight possible parallels with human processing.

The tests here are limited to the bidirectional masked language modeling framework used for training BERT, as opposed to autoregressive LM architectures such as RNNs, or GPT-2 (Radford et al., 2019). In future work it will be informative to establish whether different architectures and training objectives will produce differences in sensitivities towards contextual cues. Our paradigm can be extended by complementing our sampling procedure with hand-crafted templates of simple sentences that place all context to the left of target words. This will enable testing in the context of incremental language processing and help compare priming across various LM strategies.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful comments and suggestions. This work has also benefited from fruitful discussions with the members of the AKRaNLU lab at Purdue University, and the CompLing lab at the University of Chicago.

References

Jeremy Auguste, Arnaud Rey, and Benoit Favre. 2017. Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 21–26, Copenhagen, Denmark. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Allyson Ettinger and Tal Linzen. 2016. Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 72–77, Berlin, Germany. Association for Computational Linguistics.

Kara D Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, 41(4):469–495.

Stefan L Frank and John Hoeks. 2019. The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *CogSci 2019*, pages 337–343.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2013. Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 878–883.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT 2019*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of NAACL-HLT 2018*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Keith A Hutchison, David A Balota, James H Neely, Michael J Cortese, Emily R Cohen-Shikora, Chi-Shing Tse, Melvin J Yap, Jesse J Bengson, Dale Niemeyer, and Erin Buchanan. 2013. The semantic priming project. *Behavior Research Methods*, 45(4):1099–1114.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Marta Kutas and Steven A Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Elisabeth L McFalls and Paula J Schwanenflugel. 2002. The influence of contextual constraints on recall for words within sentences. *American Journal of Psychology*, 115(1):67–88.

Timothy P McNamara. 2005. *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT 2016*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *CogSci 2018*, pages 2603–2608.

Paula J Schwanenflugel. 1991. Contextual constraint and lexical processing. In *Advances in psychology*, volume 77, pages 23–45. Elsevier.

Paula J. Schwanenflugel and Kathy L. LaCount. 1988. Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(2):344–354.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764.