The 2020 Conference on Empirical Methods in Natural Language Processing

EMNLP 2020

Review Report

Title: Exploring BERT's sensitivity to lexical cues using tests from semantic priming
Authors: Kanishka Misra, Allyson Ettinger and Julia Rayz
Status: Undecided

Review #1

What is this paper about, what contributions does it make, and what are the main strengths and weaknesses?

The study continues the line of work using methods from psycholingusitics to investigate pretrained language models for what they understand. The paper is laudable for focusing on a semantic task, finding that semantic priming effects can also be found in BERT. The analysis, finding also distraction effects, is particularly well-executed. A primary weakness would be a lack of a clear take-home message: semantic priming purports to show that BERT relies on lexical cues, but not in constrained contexts. But the constrained contexts themselves could be constrained because of lexical items appearing elsewhere in the context.

Reasons to accept

A well-executed analysis experiment for investigating the nature of a pre-trained language model's understanding.

Reasons to reject

No "risks" per se

Reproducibility: 4
Overall Recommendation: 3.5

Questions for the Author(s)

- The concept of "priming" has a lot to do with left-to-right processing in humans, which Transformer-based LMs lack (i.e. stuff read before effects stuff read after). This makes me wonder what would happen if the priming task were done differently: e.g. would the same effects be observed if the prime was appended instead of prepended?

Missing References

- In the introduction, I would minimally recommend adding references to the ULMFiT (https://arxiv.org/abs/1801.06146) and ELMo papers (https://www.aclweb.org/anthology/N18-1202/)

Review #2

What is this paper about, what contributions does it make, and what are the main strengths and weaknesses?

This work provides insights into the behavior of the BERT language model (contextual embeddings) in the context of priming effects. It's a nice little experimental study of one semantic aspect of a well-known model. The experiment looks into the effect of priming cues in low-constraint and high-constraint context of a target word. Both the data (priming triples from human sentence processing experiments) and method (surprisal) are from previous work in computational linguistics and psycholinguistics; applying them to test BERT prediction mechanism is the main contribution of this work.

Reasons to accept

The paper is very clear and well written on a generally interesting topic; I consider it a nice read for both specialized and general audience in psycholinguistics and computational linguistics. BERT is a widely used model in NLP and understanding its predictions alignment with human language processing would help the community choose it or not for particular use cases.

No new dataset, no new NLP model or algorithm, and no new methodology is proposed. This work is an evaluation paper, and a rather simple one. So it would be nice if they can provided:

- an evaluation against one of the simpler sentence embedding models as a baseline for comparison
- an evaluation against one of the more recently proposed architectures such as GPT-2 (this is mentioned in the conclusion as a future direction though)
- rationalization of the results by discussing BERT architecture and learning algorithm. The only place in the paper that I could find a mention of "why" the priming effect is observed is "This may be because words sharing these relations are simply more likely to occur during BERT's training" which is rather simple and is more related to data than the model. Just the fact that BERT is a widely used NLP model these days, does not convince or motivate one to test this model against human data and expect similar results. I would like the authors to make a connection between each of the effects the have observed and explain "why" BERT is capable of reproducing them. Specially, in the case of mispriming effect, there's not much evidence to create a base hypothesis to whether and why we should see that in a given language model. Do we, for example, expect to see the same effect in another LM like GPT-2 or do we not, and why? What is different or special in BERT's learning algorithm or structure that enables it to learn/generate these effects?

By addressing the above points, the paper would have the potential of getting a higher overall score as a strong evaluation study (vs., an ordinary one, which it is currently).

Reproducibility: 4

Overall Recommendation: 3.5

Questions for the Author(s)

I have posed a few questions in the above.

Review #3

What is this paper about, what contributions does it make, and what are the main strengths and weaknesses?

Summary: This paper tests the semantic priming phenomenon from cognitive science/psychology for BERT pretrained models. It uses priming word pairs from semantic priming for humans literature to test how BERT responds, and finds that BERT does show priming effects, but with differing sensitivity depending on how open-ended (which they define more rigorously as "constraining") the context sentence is, and depending on the lexical relation of the related and unrelated word pairs.

Strengths: The methodology is solid, and scientifically rigorous, as are the results. The general discussion section (6) is very strong. The first paragraph particularly, but all of that section is very strong.

Weakesses: It is actually unclear what research question is being answered or the novel contribution to the field that the paper is making. It seems at first like a cogsci paper, so potentially it's just in the wrong track. But then in 5.3 they redefine primes to be only if they worked, which seems different from the human version of priming (though perhaps this is not the case as I am unfamiliar with the reference in line 740 that they use to motivate it).

Overall: This paper is impressive in its rigor but ambiguous in its positioning and the research question it answers. I can imagine many ways that looking into priming effects could be useful outside of cogsci -- as a security or a robustness study. However this has been done before in mispriming studies (as they mention in related work) so the novelty is unclear. They assert that there is a difference in contribution because they use official semantic priming pairs , but I think this is insufficient as a contribution of a long paper. Alternatively the contribution of this paper seems to truly be a study into how the levels of "constraint" interact with priming words and change priming effects, and also how the lexical relations of related and unrelated pairs change this effect. This seems to me to be quite valuable. Though if that is truly the focus, it would make sense to have more detail and analysis on the lexical relations, and to engage with information theory for the "constraint" thigs.

Reasons to accept

It rigorously outlines the effect of how constraining a sentence is on the effects of priming, which can further inform the experiments on priming in the field.

Reasons to reject

The takeaways are unclear as the paper is currently laid out -- I think the research is well done and the findings are valuable, but I would prefer it to take a clearer direction and then tailor the analysis better towards that direction. As it is I think most of the findings are better suited to a short paper, though it could easily be expanded to a long one with additional targeted experiments.

Reproducibility: 5
Overall Recommendation: 3

Questions for the Author(s)

What do you envision your core research question is, and how your work fits into the field? Can you see a way this can be reorganised to make that clearer?

Typos, Grammar, Style, and Presentation Improvements

Here are things that should be clearer: Abstract: This sentence it not yet clear till the rest of the paper has been read " Followup analysis shows BERT to be increasingly distracted by related prime words as context becomes more informative, assigning lower probabilities to related words. "

Lines: 078 -- contextual constraint is used but not defined. Maybe provide an example of constrained vs. unconstrained contexts. We don't get is till 173 -- you can refer forward or put the information earlier. And the technical definition isn't till 341...when it seems like this is the main thrust of the paper this needs to be clear early.

195 -- facilitation seems like a jargon term, define it. You don't define it till 435.

625 -- unnecessary, we of course value qualitative examples

START Conference Manager (V2.61.0 - Rev. 6099M)