Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context

Urvashi Khandelwal, He He, Peng Qi, Dan Jurafsky
Stanford University

An analytic study of how LSTM language models use prior linguistic context. We measure changes in LSTM performance, as a result of ablations applied to contextual features of the input, during evaluation.

Setup

- Perturbations applied only during evaluation.
- Datasets: **Penn Treebank** (PTB) and **Wikitext-2** (Wiki).
- Standard LSTM LM architecture (Merity et al., 2018).
- All results are reported on the development set (to protect the test set).
- Measuring changes in negative log likelihood:

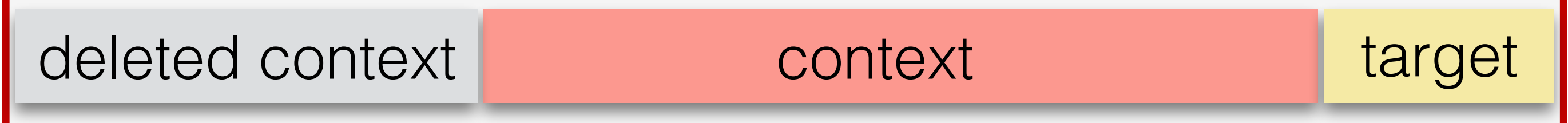
$$\mathrm{NLL} = -\frac{1}{T}\sum_{i=1}^{T} \log P(w_t | w_{t-1}, \ldots, w_1)$$

Implications

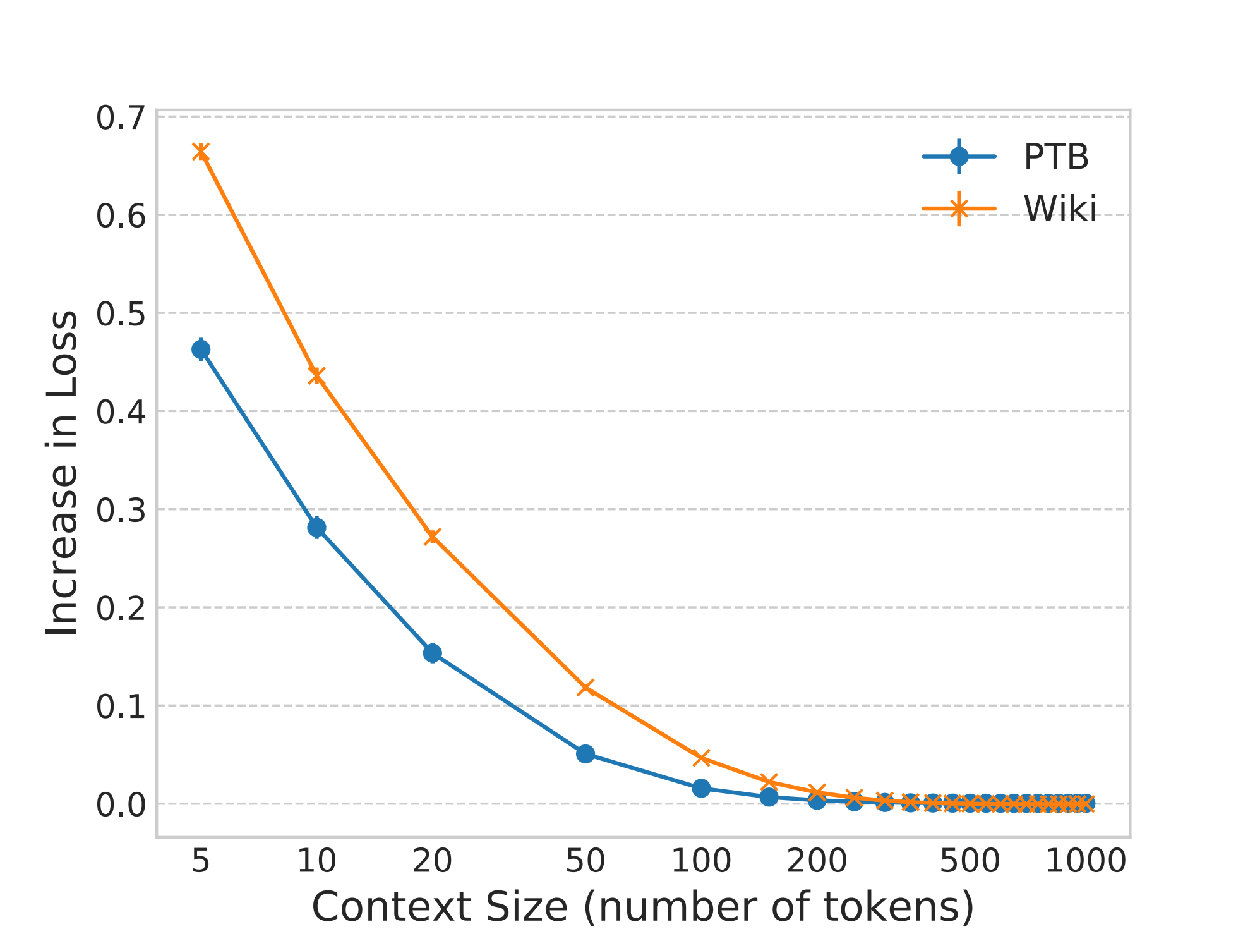
- Improve existing models!
- Compare model classes on more than just test set perplexities!
- Can we decouple the data from the models?
  Experiment with different model classes and different languages
- Theoretical justifications???

How much context is used?

Perturbation: *guess a context size, delete all prior tokens*

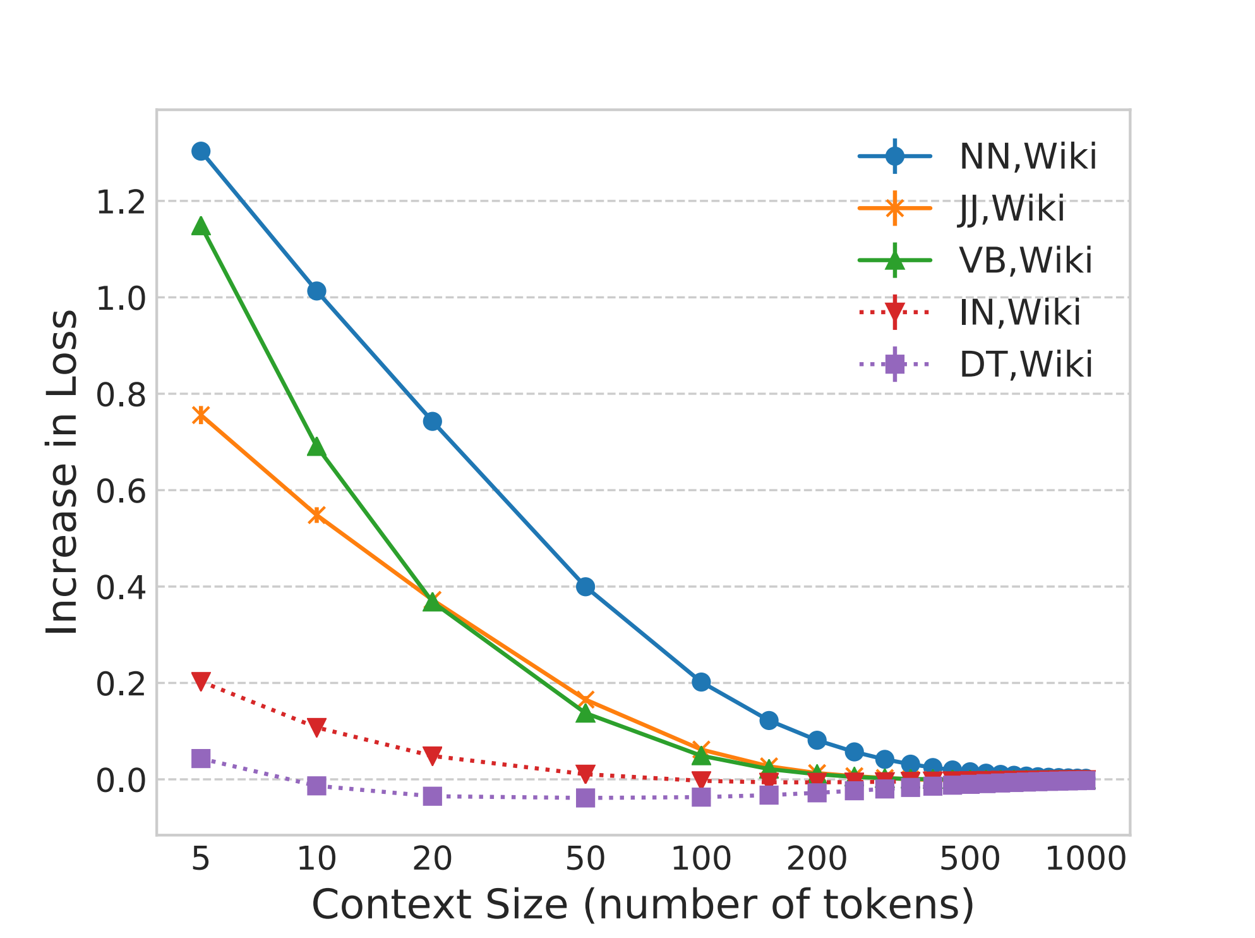

deleted context | context | target

LSTM language models can use at least about 200 tokens of context, on average.



Changing hyperparameters does not change the amount of context used.

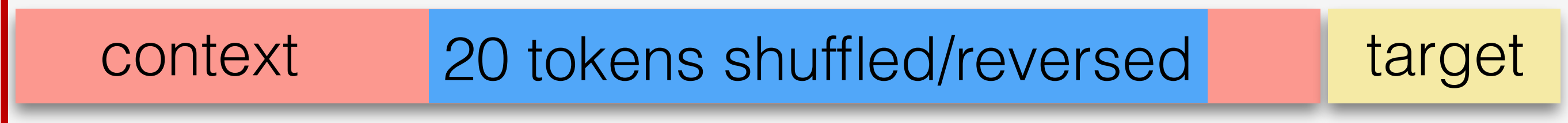
Content words (eg: nouns) need far more context than function words (eg: determiners).



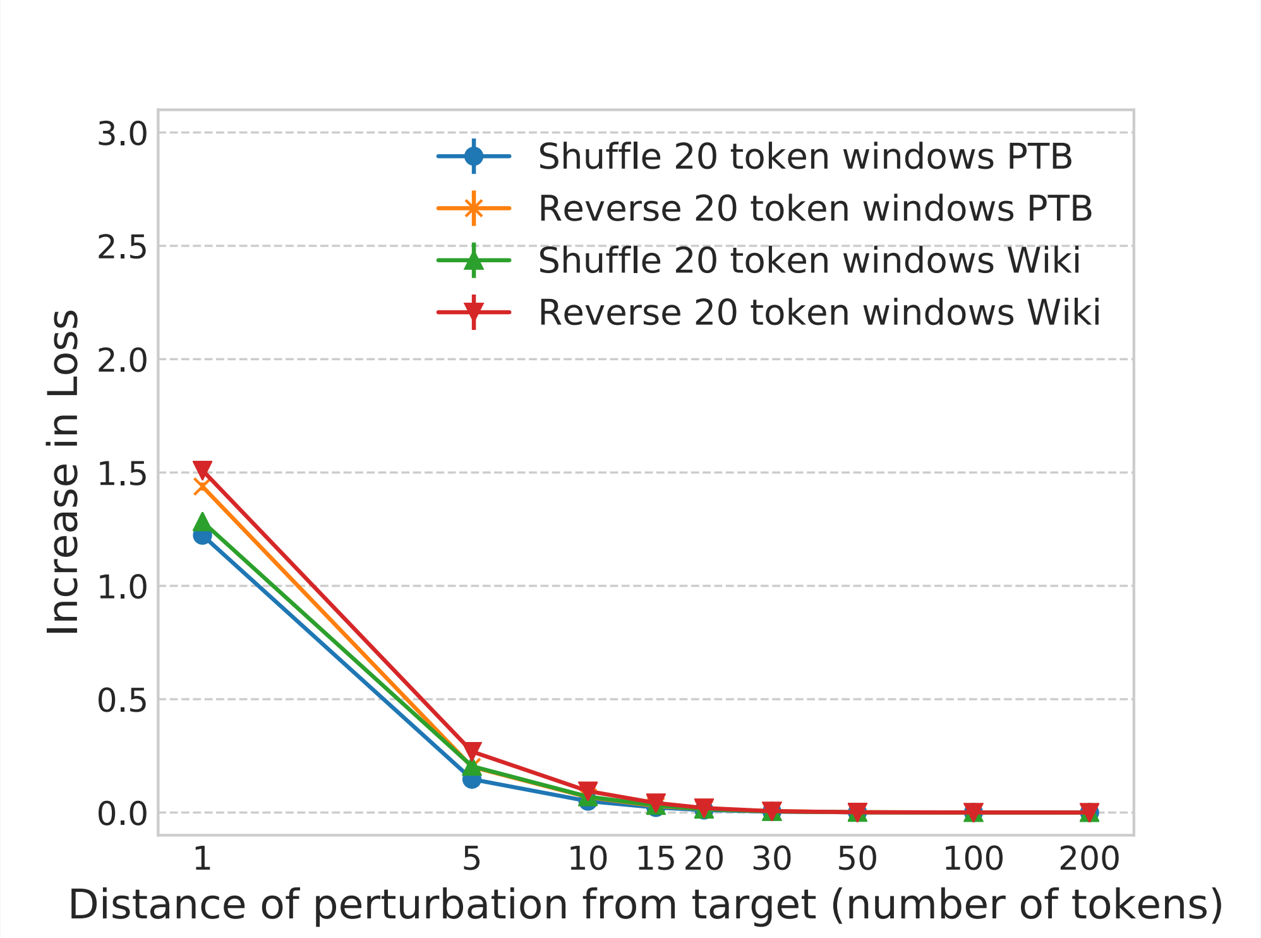
Does word order matter?

Perturbation: *shuffle/reverse spans in prior context*

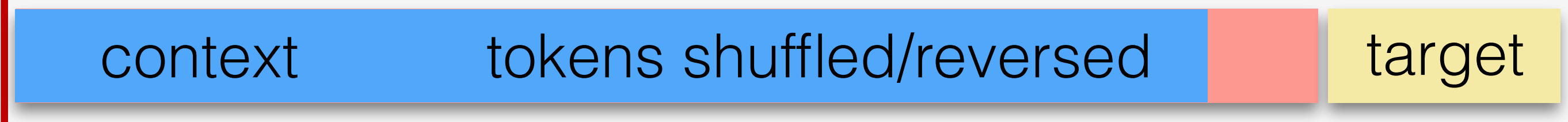
Local Word Order: *order of words only within a* ***20 token span***.


context | 20 tokens shuffled/reversed | target

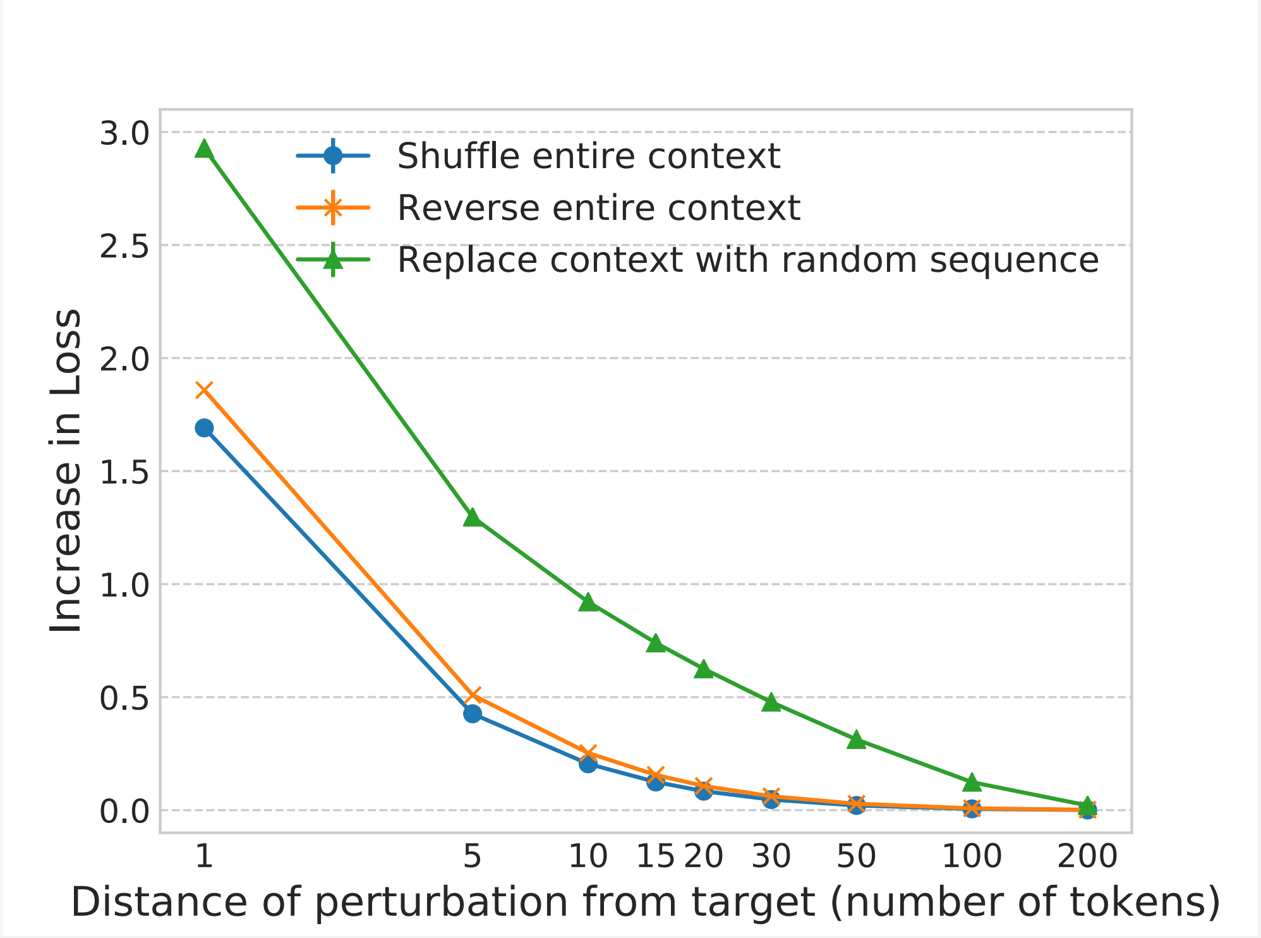
Local word order only matters within the most recent sentence, ~20 tokens.



Global Word Order: *order of words within the* ***entire sequence***.


context | tokens shuffled/reversed | target

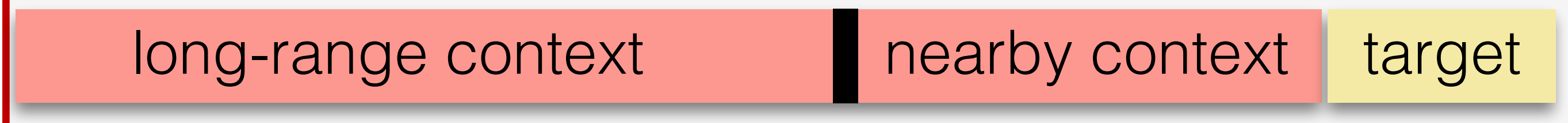
Global word order only matters for the most recent 50 tokens.



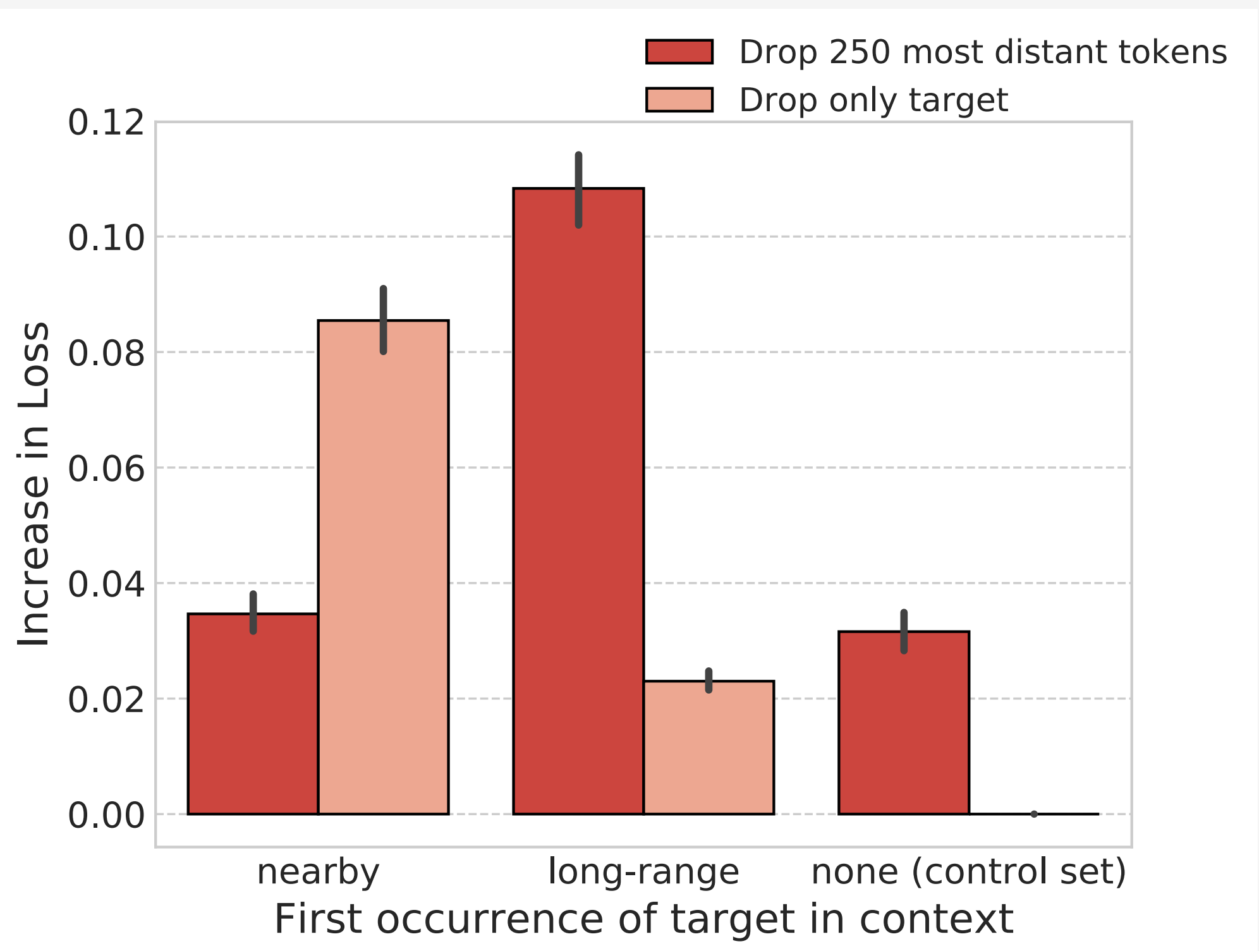
Can LSTMs copy words?

Three Categories of Target Words

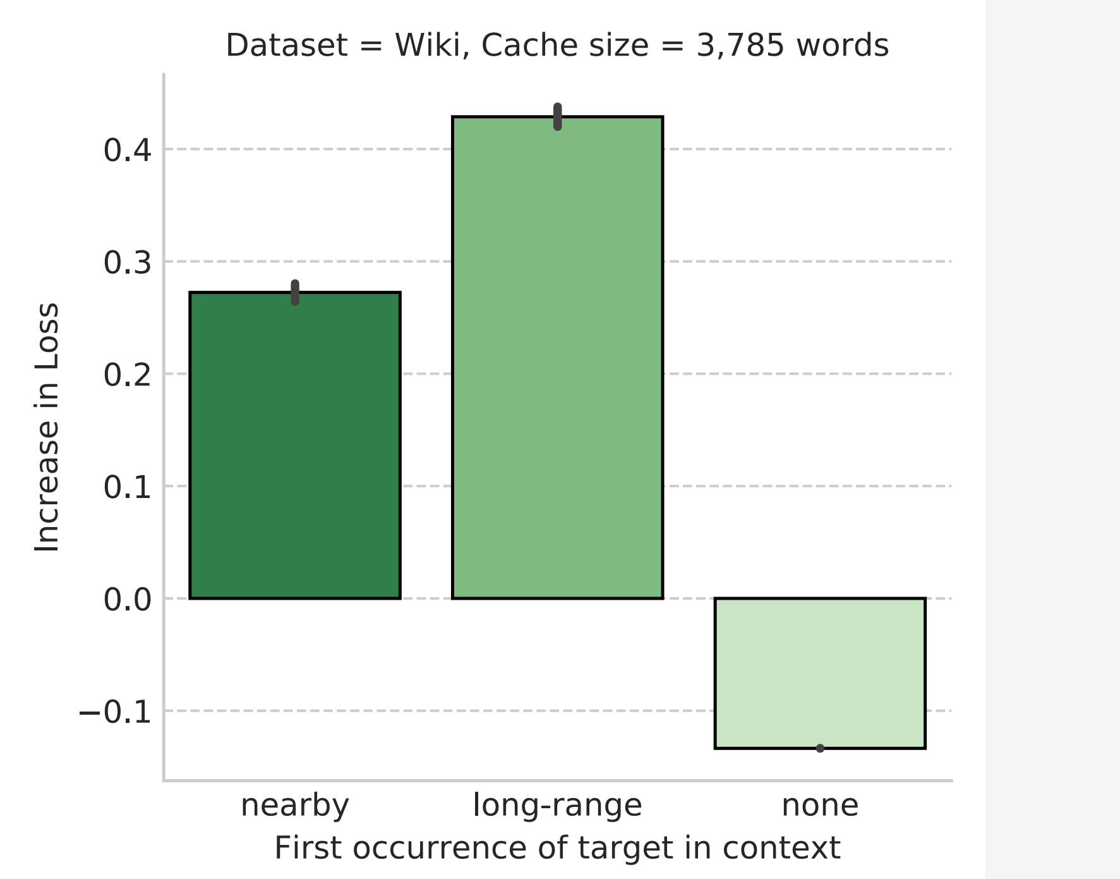
1. *Appear in their own nearby context (within 50 tokens).*
2. *Appear only in their own long-range context (beyond 50 tokens).*
3. *Never appear in their own context (none).*


long-range context | nearby context | target

LSTMs can regenerate words seen in nearby context.



Neural Caches (Grave et al., 2017b) help words that can be copied from long-range context, the most.



References

[1] Stephen Merity, Nitish Shirish Keskar, Richard Socher. 2018. *Regularizing and Optimizing LSTM Language Models*. In ICLR.

[2] Edouard Grave, Armand Joulin, Nicolas Usunier. 2017b. *Improving Neural Language Models with a Continuous Cache*. In ICLR.

Acknowledgements

We gratefully acknowledge support of the DARPA Communicating with Computers (CwC) program under ARO prime contract no. W911NF15-1-0462 and the NSF via grant IIS-1514268.

Code: https://github.com/urvashik/lm-context-analysis