Overview

- Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context ACL 2018
- Do Transformers Need Deep Long-Range Memory? ACL 2020
- What Context Features Can Transformer Language Models Use? ACL 2021

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

Urvashi Khandelwal, He He, Peng Qi, Dan Jurafsky
Computer Science Department
Stanford University
{urvashik,hehe,pengqi,jurafsky}@stanford.edu

Introduction

- Language models are an important component of natural language generation tasks, such as machine translation and summarization.
- However, all of the previous work studies LSTMs at the sentence level, even though they can potentially encode longer context.
- This work is to complement the prior work to provide a richer understanding of the role of context, in particular, long-range context beyond a sentence.

Method

The paper aims to answer the following questions:

- How much context is used by NLMs, in terms of the number of tokens?
- Within this range, are nearby and long-range contexts represented differently?
- How do copy mechanisms help the model use different regions of context?

Method

- Language Modeling

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

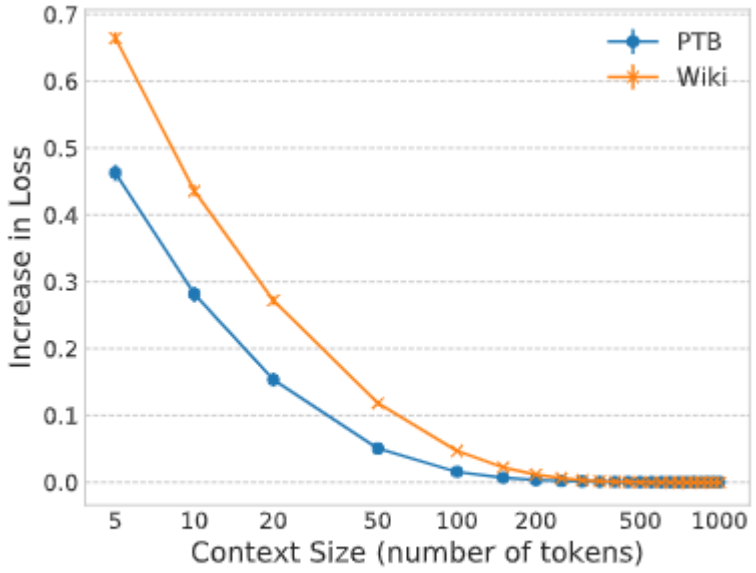
- Model: a standard LSTM language model
- Datasets: Penn Treebank (PTB) and Wikitext-2
- Method: training model with correct text, testing with perturbed text.
- Perturbations: dropping tokens, shuffling/reversing tokens, and replacing tokens with other words from the vocabulary.

Experiment

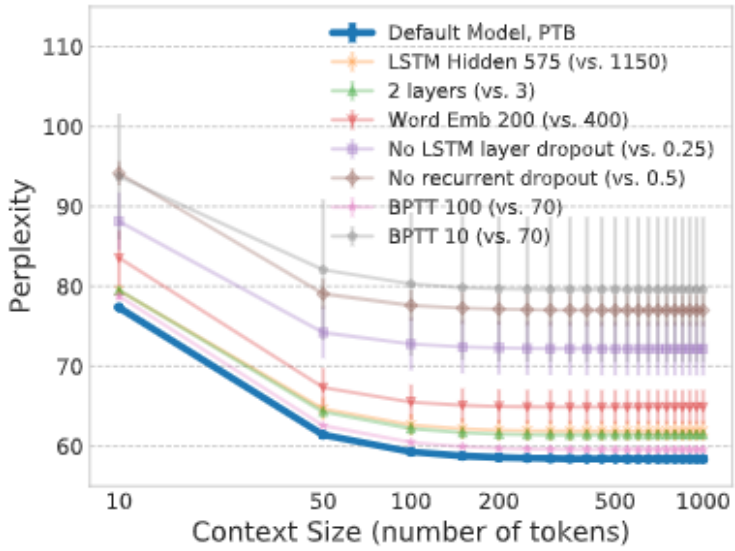
- How much context is used?

$$\delta_{\text{truncate}}(w_{t-1}, \ldots, w_1) = (w_{t-1}, \ldots, w_{t-n}),$$

- LSTM language models have an effective context size of about 200 tokens on average.
- Changing hyperparameters does not change the effective context size.



(a) Varying context size.



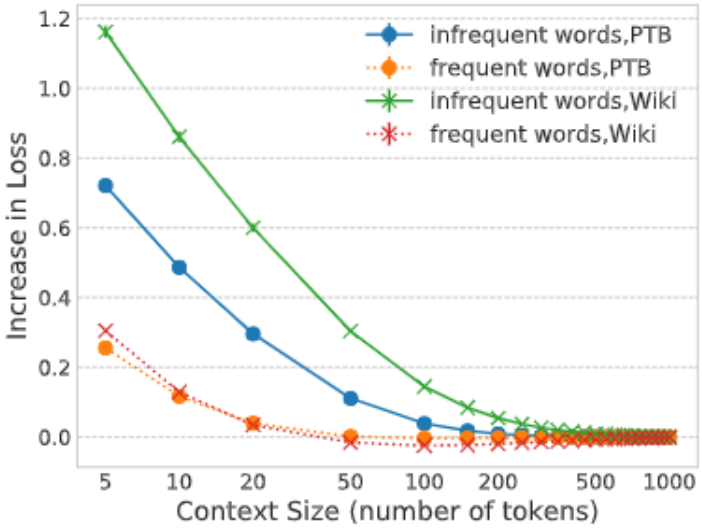
(b) Changing model hyperparameters.

Experiment

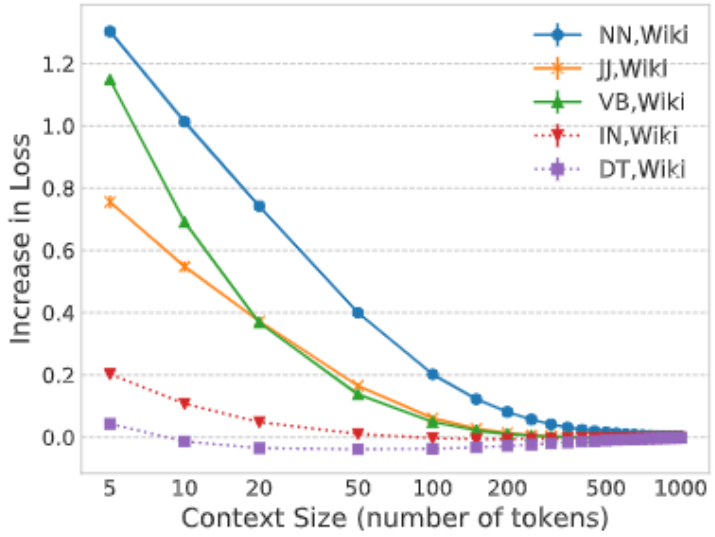
- Do different types of words need different amounts of context?

  frequency/parts-of-speech.

- Infrequent words need more context than frequent words.
- Content words need more context than function words.



(c) Frequent vs. infrequent words.



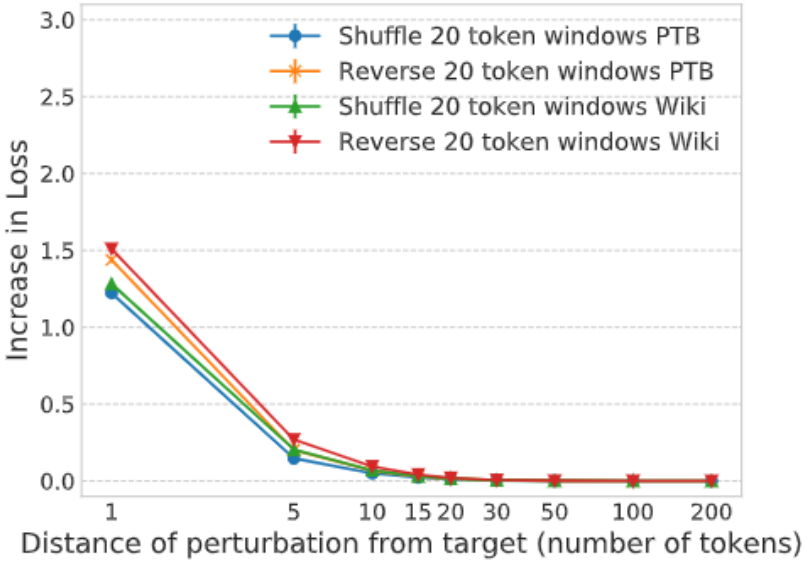
(d) Different parts-of-speech.

Experiment

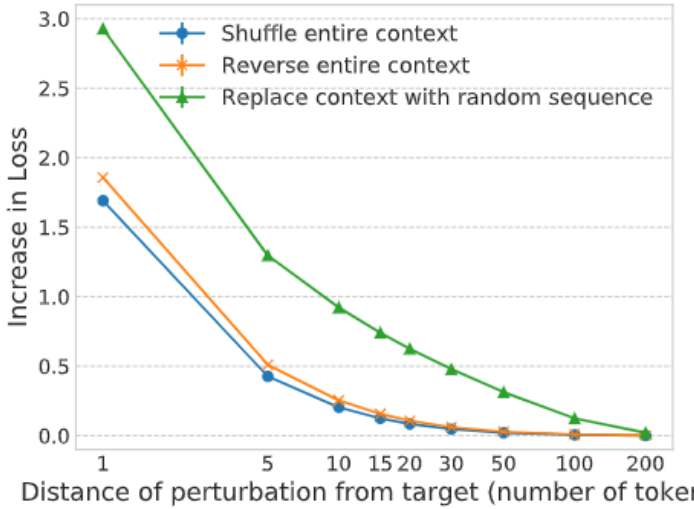
- Does word order matter?

$$\delta_{\text{permute}}(w_{t-1}, \ldots, w_{t-n}) =$$
$$(w_{t-1}, .., \rho(w_{t-s_1-1}, .., w_{t-s_2}), .., w_{t-n})$$

- Local word order only matters for the most recent 20 tokens.
- Global order of words only matters for the most recent 50 tokens.



(a) Perturb order locally, within 20 tokens of each point.



(b) Perturb global order, i.e. all tokens in the context before a given point, in Wiki.

Experiment

- Types of words and the region of context

$$\delta_{\text{drop}}(w_{t-1}, \ldots, w_{t-n}) =$$
$$(w_{t-1}, .., w_{t-s_1}, f_{\text{pos}}(y, (w_{t-s_1-1}, .., w_{t-n})))$$

- Content words matter more than function words.

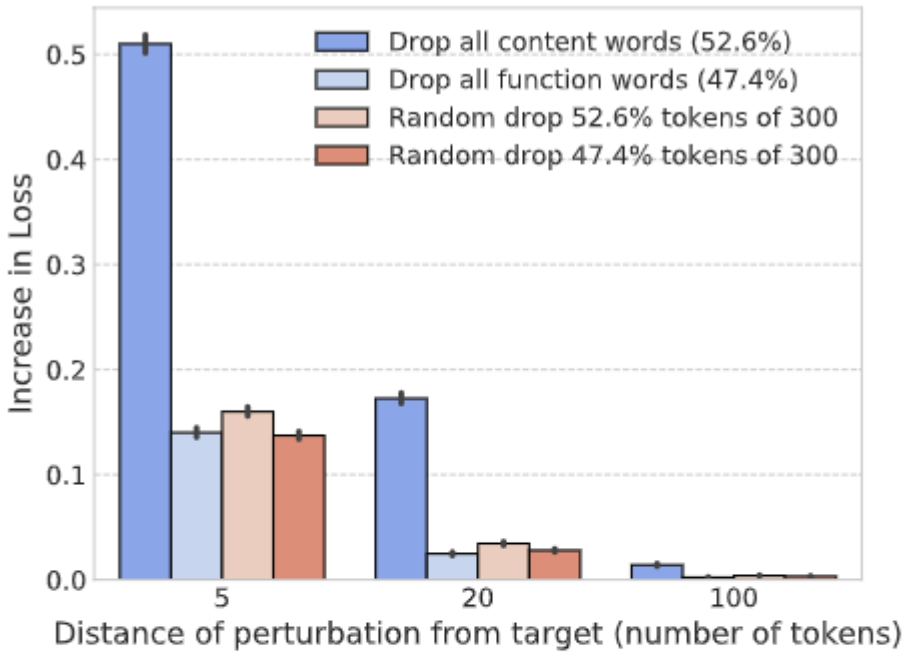


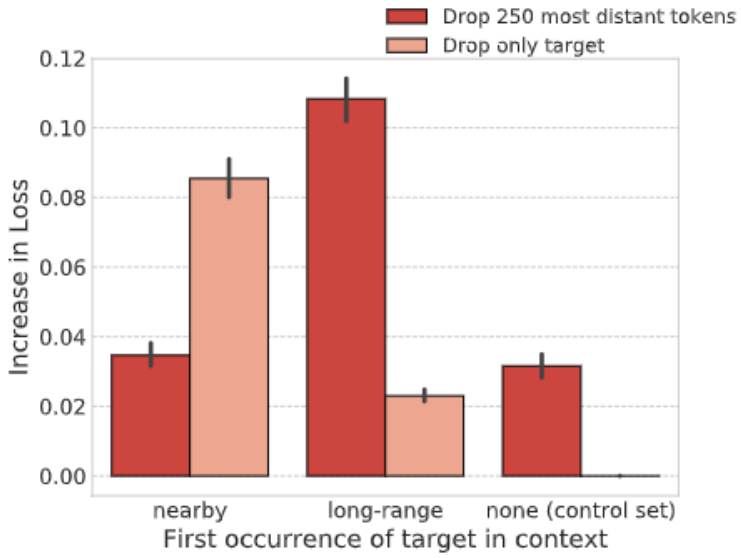
Figure 3: Effect of dropping content and function words from 300 tokens of context relative to an unperturbed baseline, on PTB. Error bars represent 95% confidence intervals. Dropping both content and function words 5 tokens away from the target results in a nontrivial increase in loss, whereas beyond 20 tokens, only content words are relevant.

Experiment

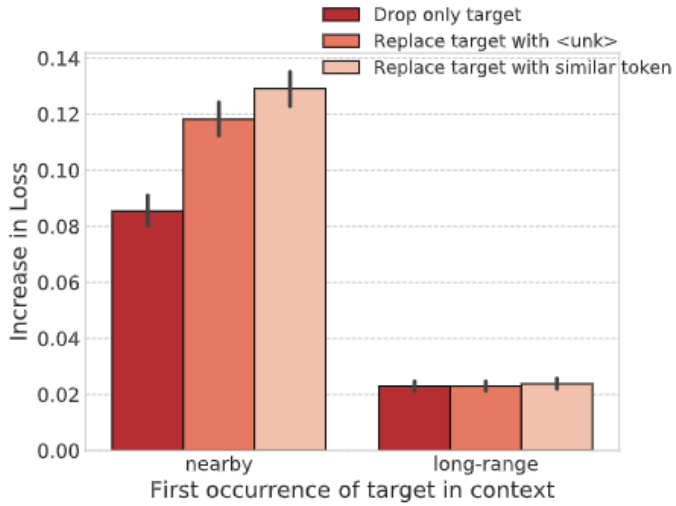
- Can LSTMs copy words without caches?

$$\delta_{\text{drop}}(w_{t-1}, \ldots, w_{t-n}) = f_{\text{word}}(w_t, (w_{t-1}, \ldots, w_{t-n})),$$

- LSTMs can regenerate words seen in nearby context.



(a) Dropping tokens



(b) Perturbing occurrences of target word in context.

Experiment

- How does the cache help?

$$P_{\text{cache}}(w_t|w_{t-1},\ldots,w_1;h_t,\ldots,h_1)$$
$$\propto \sum_{i=1}^{t-1} \mathbb{1}[w_i = w_t]\exp(\theta h_i^T h_t),$$

- Caches help words that can be copied from long-range context the most

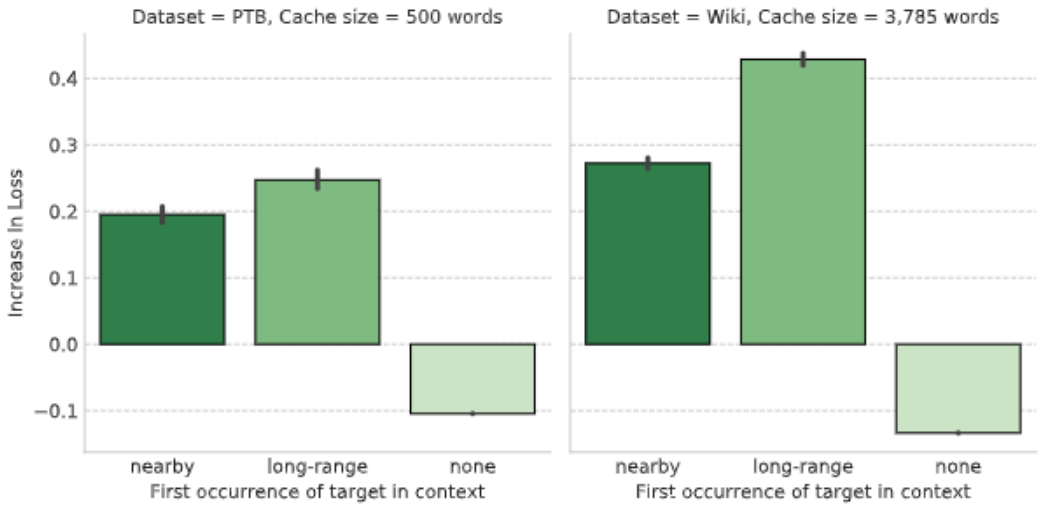


Figure 7: Model performance relative to using a cache. Error bars represent 95% confidence intervals. Words that can only be copied from the distant context benefit the most from using a cache.

Summary

- A standard LSTM language model can effectively use about 200 tokens of context.
- It is sensitive to word order in the nearby context, but less so in the long-range context.
- The model is able to regenerate words from nearby context, but heavily relies on caches to copy words from far away.

Do Transformers Need Deep Long-Range Memory?

Jack W. Rae
DeepMind & UCL
London, UK
jwrae@google.com

Ali Razavi
DeepMind
London, UK
alirazavi@google.com

Introduction

- Deep attention models have advanced the modelling of sequential data across many domains.
- For language modelling in particular, the Transformer-XL has been shown to be state-of-the-art across a variety of well-studied benchmarks.
- However it is unclear whether this is necessary.

Method

- replace the long-range memory, for a given layer, with a short-range memory (SRM)

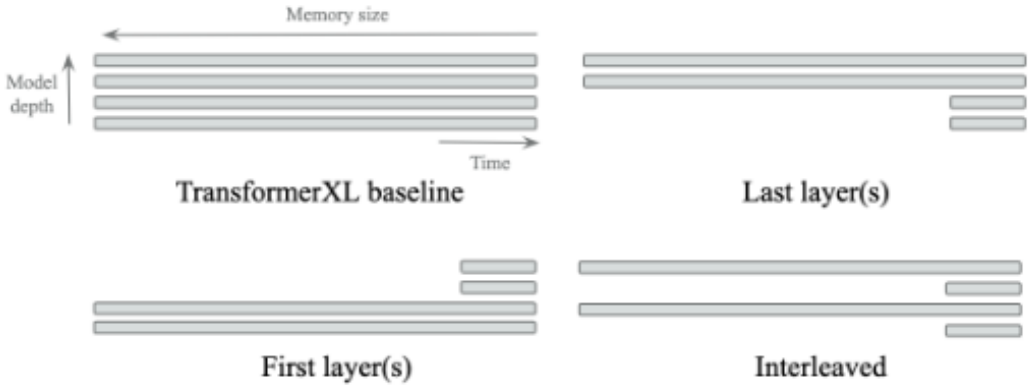


Figure 1: Comparison of arrangement patterns for long-range and short-range memories across the layers of a Transformer. Baseline contains equally long-range memories at every layer.

Experiment

- Position clearly matters, if we place long-range memories in the first layers then performance is significantly worse.
- The full TXL with 24 LRMs is seemingly identical to the 12 LRM models, with either LRMs interleaved across the whole model or LRMs placed in the final 12 layers

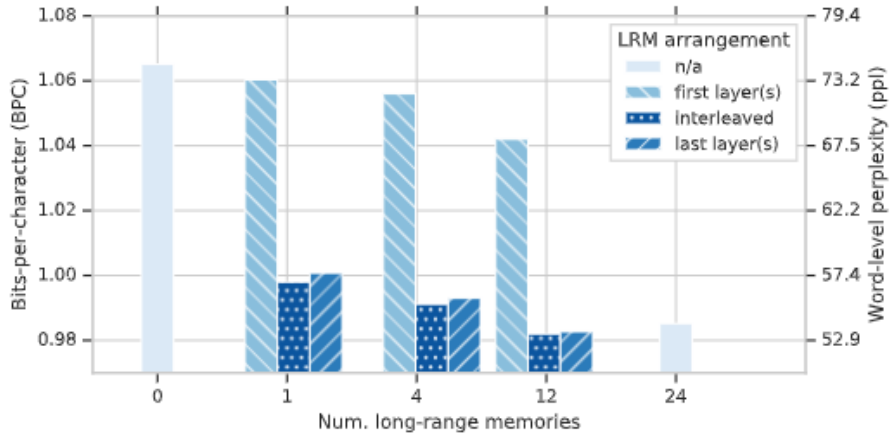


Figure 3: Enwik8 test performance over a varying number of long-range memories and arrangement patterns. Lower is better. Model: 24-layer Transformer-XL, evaluation long-range memory size: 6000 (trained with 2304) and short-range memories size: 128.

Experiment

- Increasing the memory size beyond 512 further slows the model down and reduces modelling performance.
- limiting the range of attention can not only speed up the model but improve performance.

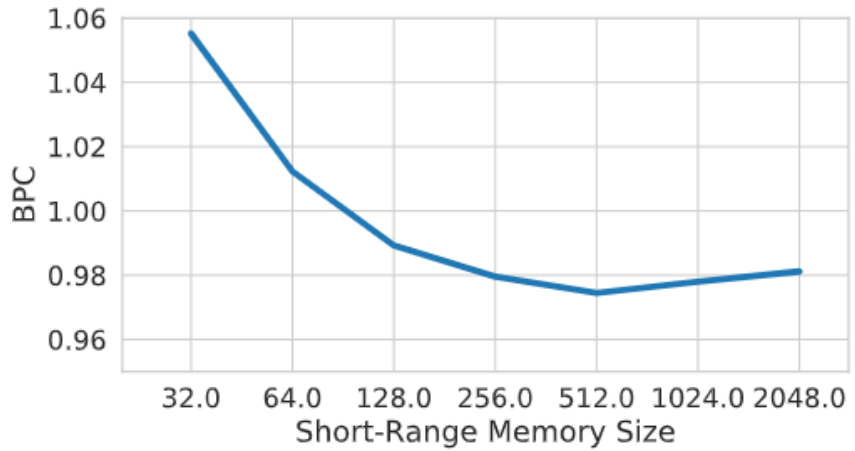


Figure 4: Enwik8 test performance for varying short-range memory length (at both train and test). TransformerXL model uses 4 interleaved long-range memories (trained 2304, tested 6000) and 20 short-range memory layers.

What Context Features Can Transformer Language Models Use?

Joe O'Connor **Jacob Andreas**
Massachusetts Institute of Technology
{joeoc,jda}@mit.edu

Introduction

- Recent years have seen a significant improvement in the predictive accuracy of neural language models (LMs).
- But despite empirical evidence that long contexts are helpful, little is understood about why.
- If the future of language modeling will include a focus on contexts of increasing size, it is important to first understand what contextual information contributes to accurate prediction in current models.

Method

- language model

$$p(x) = \prod_i p(x_i \mid x_0, x_1, \ldots, x_{i-1}) \ .$$

- Usable information

Definition 1. The usable predictive information (formally, predictive **$\mathcal{V}$-information**) from a random variable X to a random variable Y as:

$$I_{\mathcal{V}}(X \to Y) = \big[\inf_{p_1 \in \mathcal{V}} -\mathbb{E} \log p_1(Y)\big] - \big[\inf_{p_2 \in \mathcal{V}} -\mathbb{E} \log p_2(Y \mid X)\big] \quad (2)$$

for a class $\mathcal{V}$ of distributions p.

- Measuring what is used

Definition 2. The **ablated information** due to an ablation f at an offset k is:

$$\mathcal{A}(f,k) = \frac{I_{\mathcal{V}}(X_{0:n} \to X_n) - I_{\mathcal{V}}(f_k(X_{0:n}) \to X_n)}{I_{\mathcal{V}}(X_{0:n} \to X_n) - I_{\mathcal{V}}(X_{n-k:n} \to X_n)} \quad (8)$$

$$= \frac{\inf_\theta \mathcal{L}(\theta, f, k) - \inf_{\theta'} \mathcal{L}(\theta', n)}{\inf_{\theta''} \mathcal{L}(\theta'', n-k) - \inf_{\theta'} \mathcal{L}(\theta', n)} \ , \quad (9)$$

where $\mathcal{L}(\theta, i)$ is the (unablated) negative log-likelihood $-\mathbb{E} \log p_\theta(X_n \mid X_{n-i:n})$.

Figure 1: Calculation of the ablated likelihood $\mathcal{L}(\text{nouns}, \ell : m \sim n)$ (Eq. (10)). A context ablation

$$\mathcal{L}(\theta, f, \ell : m \sim n) = -\frac{1}{|\mathcal{X}|(n-m)} \sum_x \sum_{i=\ell+m}^{\ell+n} \log p_\theta(X_i \mid [f(X_{0:\ell}), X_{\ell:i}])$$

Experiment

- no information model to minimize L(θ,0~512)
- a full information model to minimize L(θ,512~1024).
- For each context ablation f, we train a model to minimize L(θ,f,512 : 0~512).

Experiment

- Does order matter?

Overall word order

shuffle all

61 N.V., director the of Mr. Vinken Dutch group. as nonexecutive the 29. is Vinken, years Elsevier join old, publishing a Nov. will Pierre board chairman

shuf. trigrams globally

publishing group. N.V., the Dutch Mr. Vinken is join the board as a nonexecutive years old, will chairman of Elsevier Pierre Vinken, 61 director Nov. 29.

Word order within sentences

shuf. within sent.

61 director as the old, join will a Nov. board nonexecutive years Vinken, 29. Pierre is publishing the Vinken N.V., Mr. group. chairman Elsevier of Dutch

shuf. within trigrams

Vinken, Pierre 61 will old, years the board join a nonexecutive as Nov. director 29. Mr. Vinken is of Elsevier chairman the Dutch N.V., group. publishing

shuf. trigrams within sent.

years old, will as a nonexecutive join the board Pierre Vinken, 61 director Nov. 29. N.V., the Dutch chairman of Elsevier Mr. Vinken is publishing group.

Sentence order

shuf. sent.

Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group. Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Order of entire sections

replace w/ old

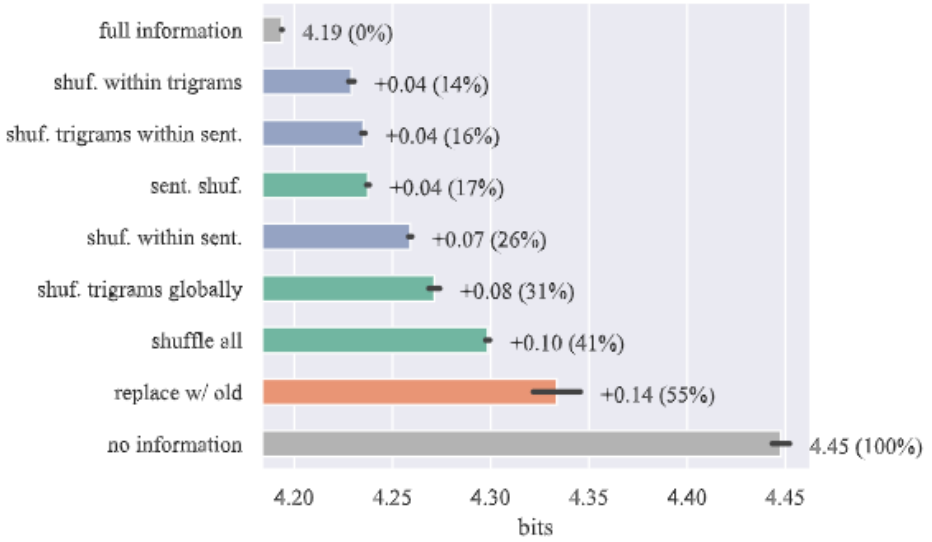
Rudolph Agnew, 55 years old and former chairman of Consolidated Gold Fields PLC, was named a nonexecutive director of this British industrial conglomerate.

Experiment

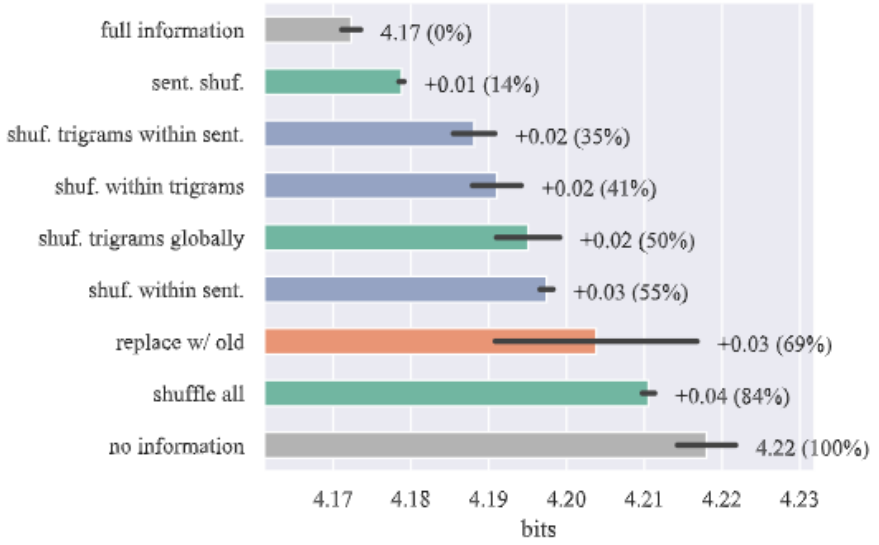
- Does order matter?

Overall word order

- ordering information is important even very far from the target.
- local co-occurrence statistics carry a significant amount of usable information.



(a) Mid-range condition (first 256 tokens after ablation)



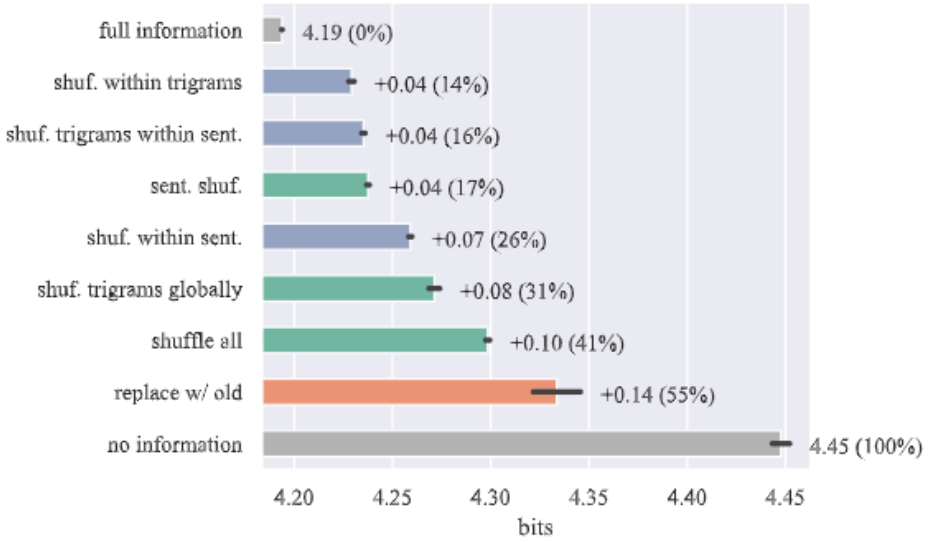
(b) Long-range condition (tokens 256-512 after ablation)

Experiment

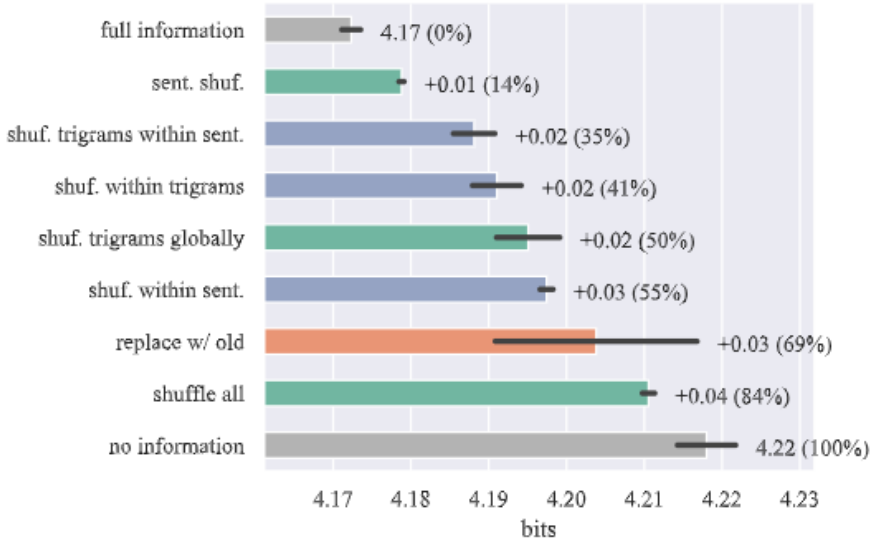
- Does order matter?

Word order within sentences

- Usable information is decreased only slightly by ablations that preserve local co-occurrence statistics and/or linear information flow.



(a) Mid-range condition (first 256 tokens after ablation)



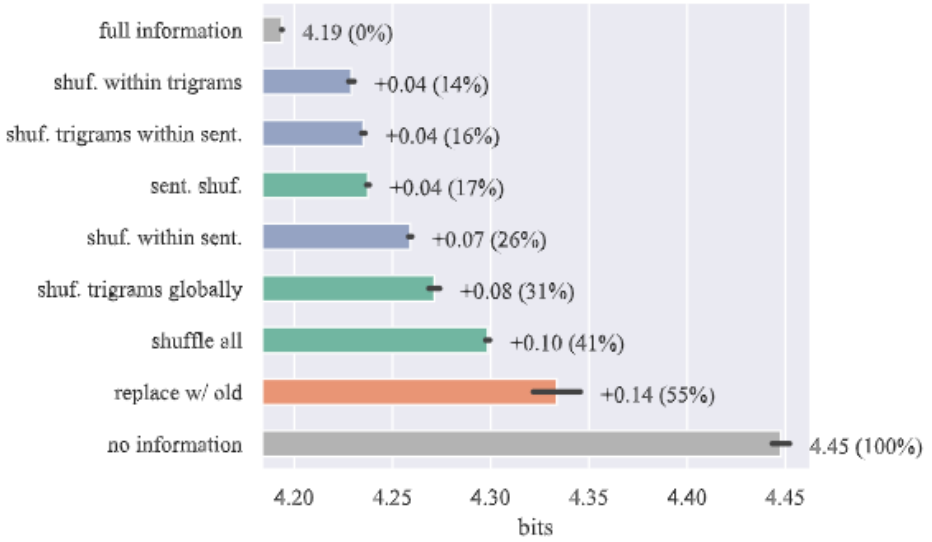
(b) Long-range condition (tokens 256-512 after ablation)

Experiment

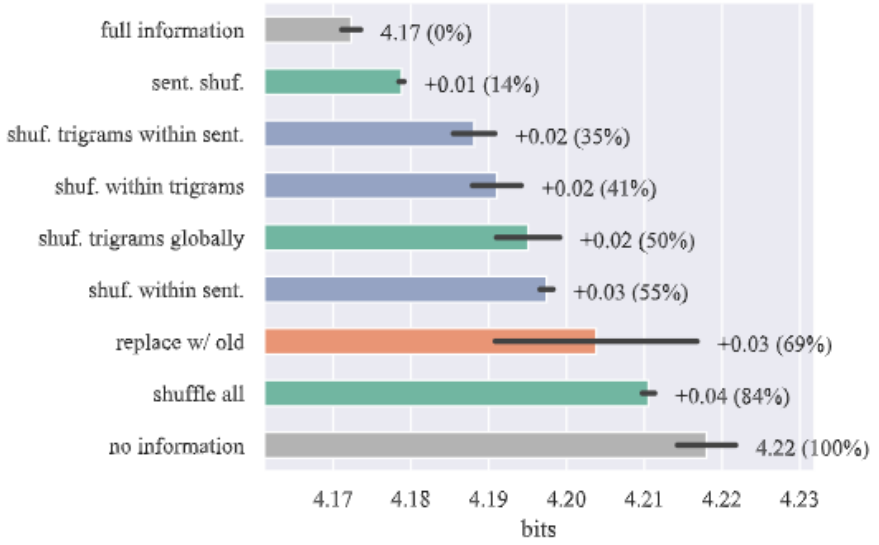
- Does order matter?

Sentence order

- prediction accuracy depends on information about local word co-occurrence, but not fine-grained word order or global position.



(a) Mid-range condition (first 256 tokens after ablation)



(b) Long-range condition (tokens 256-512 after ablation)

Experiment

- Do all words matter?

Named entities

named entities
Pierre Vinken 61 years old Nov. 29 Vinken Elsevier N.V. Dutch

Word frequency

common
Pierre years old join board director . Mr. chairman Dutch publishing group .

rare
Vinken nonexecutive Nov. Vinken Elsevier N.V.

Parts of speech

N
Pierre Vinken years board director Nov. Mr. Vinken chairman Elsevier N.V. publishing group

N & VB
Pierre Vinken years will join board director Nov. Mr. Vinken chairman Elsevier N.V. publishing group

N & VB & ADJ
Pierre Vinken years old will join board nonexecutive director Nov. Mr. Vinken chairman Elsevier N.V. Dutch publishing group

cont. words (N & VB & ADJ & ADV)
Pierre Vinken years old will join board nonexecutive director Nov. Mr. Vinken chairman Elsevier N.V. Dutch publishing group

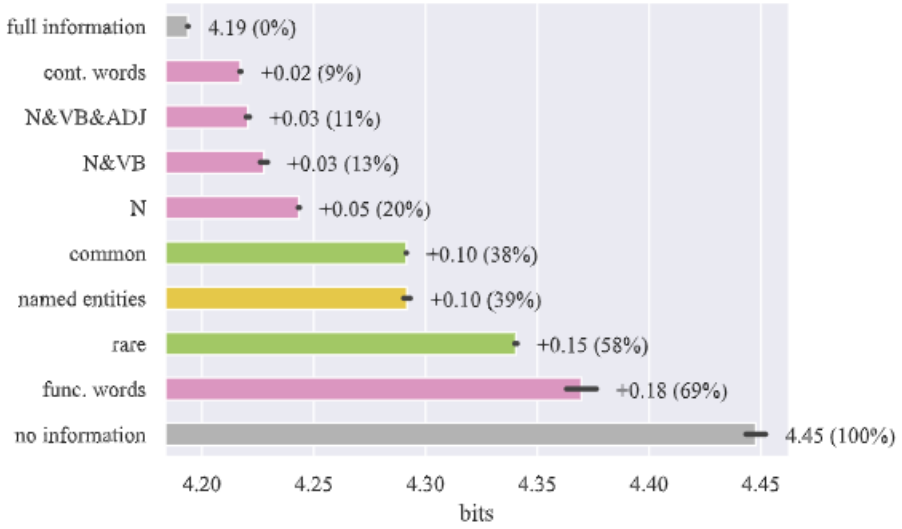
func. words
, 61 , the as a 29 . is of , the .

Experiment

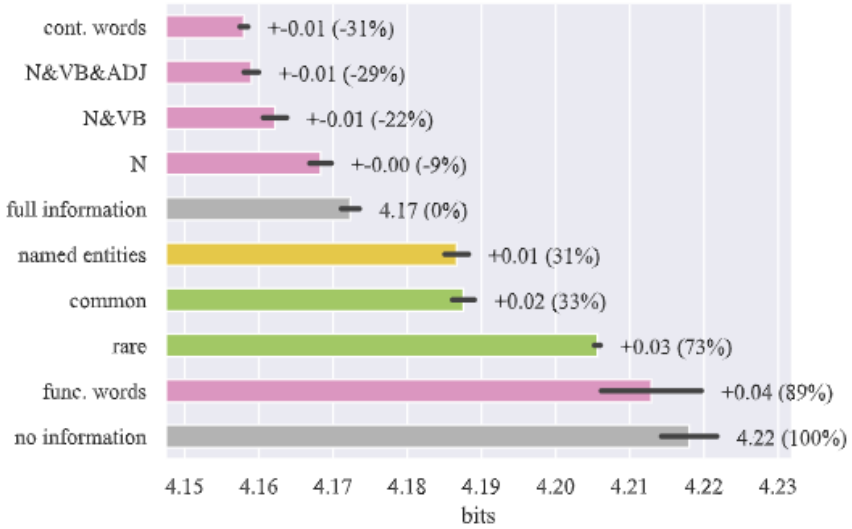
- Do all words matter?

Part-of-speech

- Most usable information, even in mid-range context, appears to be captured by nouns and verbs.
- In the long-range condition, results are even more striking: retaining only content words improves predictions over the “full information” experiment.



(a) Mid-range condition (first 256 tokens after context)

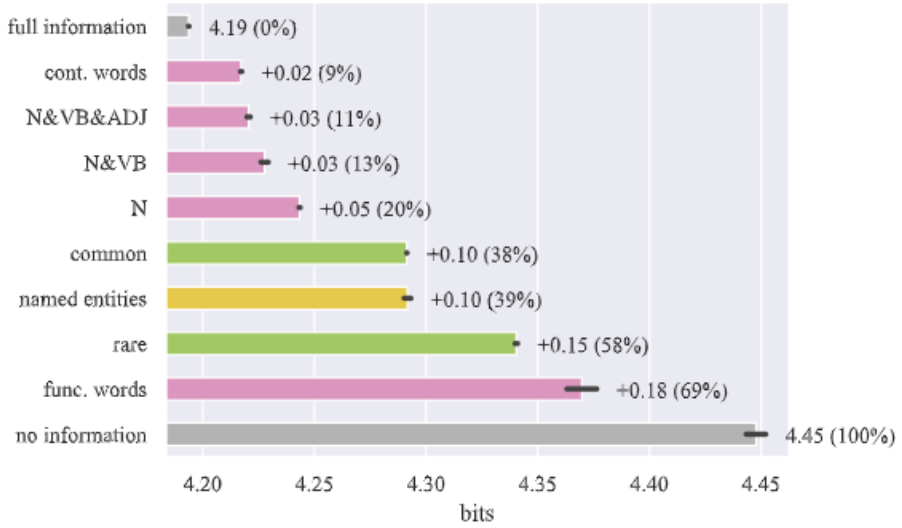


(b) Long-range condition (tokens 256-512 after context)

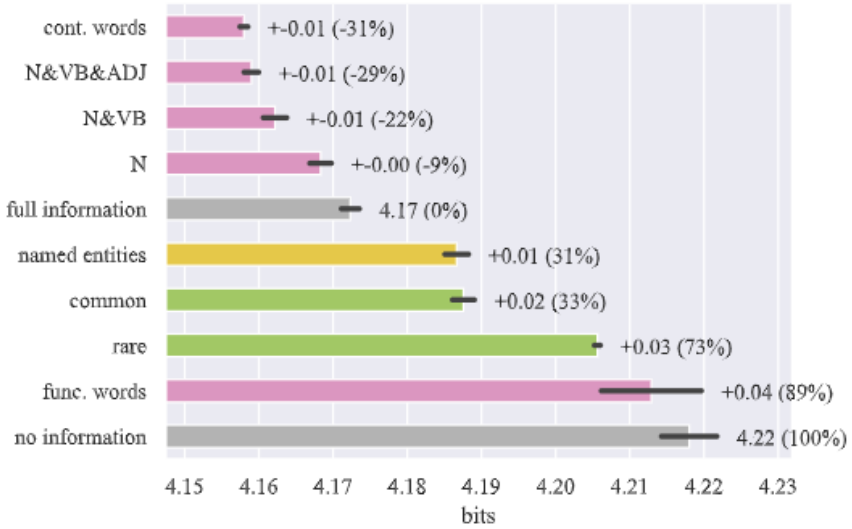
Experiment

- Do all words matter?

- **Named entities:** retaining only entities results removes only about a third of usable information in both conditions (39% and 31%)
- **Word frequency:** Both ablations remove a significant amount of information relative to the POS-based ablations above, but retaining only frequent words improves perplexity relative to rare words in both the mid- and long-range conditions



(a) Mid-range condition (first 256 tokens after context)



(b) Long-range condition (tokens 256-512 after context)

Experiment

- Making better language models?

- The lexical ablation experiments in Section 3.2 indicated that model accuracy could be improved by selective deletion of context words. Can this effect be exploited to further improve models?
- Longer contexts, even of a kind previously found to be informative, did not provide additional usable information.

Summary

- For current models, the primary carriers of information in long-range context are content words and local cooccurrence statistics: deleting function words and shuffling within local windows both have very little effect on models' predictive power.
- long contexts, but not their detailed syntactic and propositional content, are important for the low perplexity of current transformer language models.

Thanks