@ukhndlwl

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

Urvashi Khandelwal

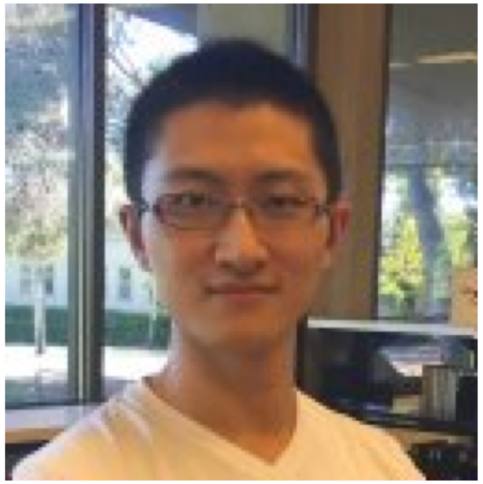
Stanford University

Bay Area Research in NLP and ML Meetup
March 28, 2019

Presented at ACL 2018

Collaborators

He He
Amazon/NYU



Peng Qi
Stanford

Dan Jurafsky
Stanford

Language Models

assign probabilities to sequences of words

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | \underbrace{w_{t-1}, \ldots, w_1}_{\text{Context}})$$

Language Models

assign probabilities to sequences of words

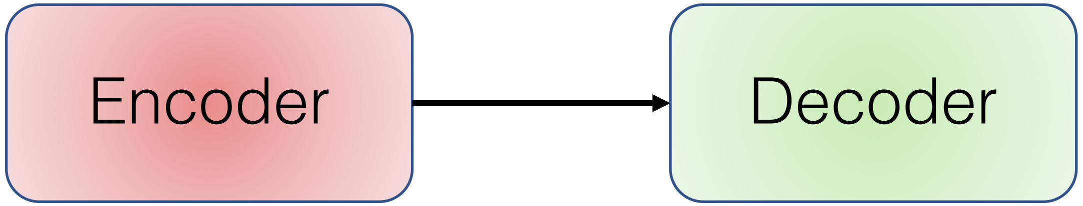
$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | \underbrace{w_{t-1}, \ldots, w_1}_{\text{Context}})$$

Iron Man is a character in the Marvel ___???___

Vocabulary

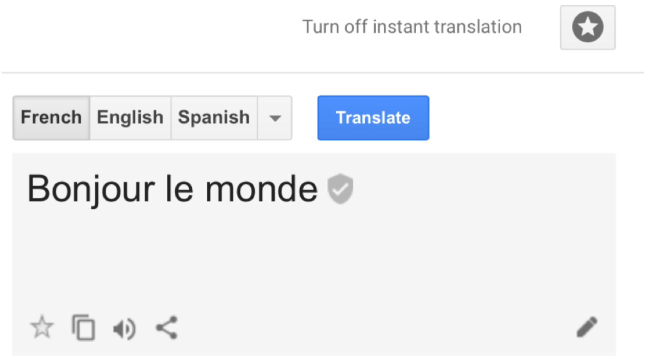
aardvark – 0.00
…
comics – 0.10
…
…
universe – 0.30
…
zyzzyva – 0.00

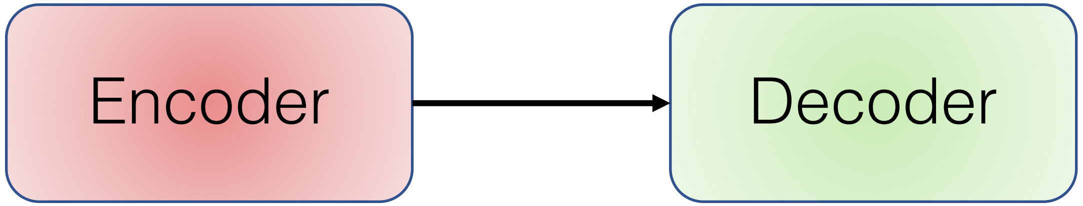
Language Models – for Generation



Sequence to Sequence

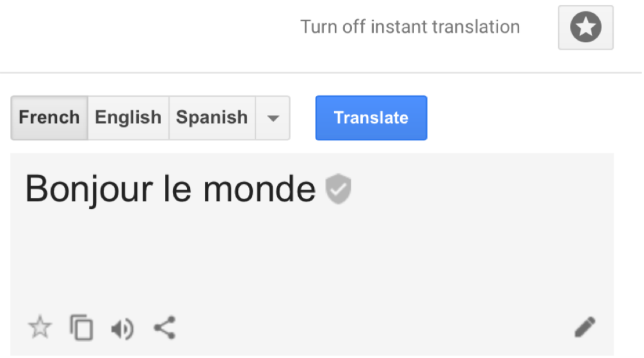
Language Models – for Generation

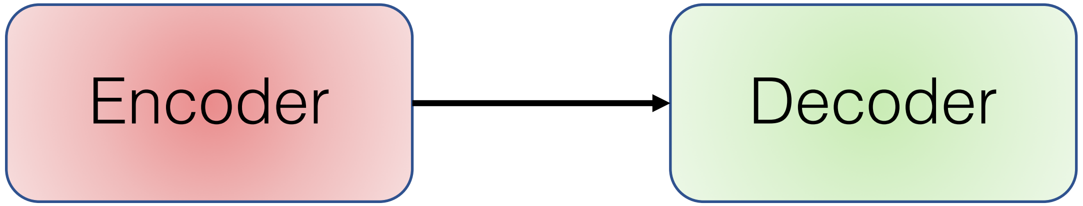




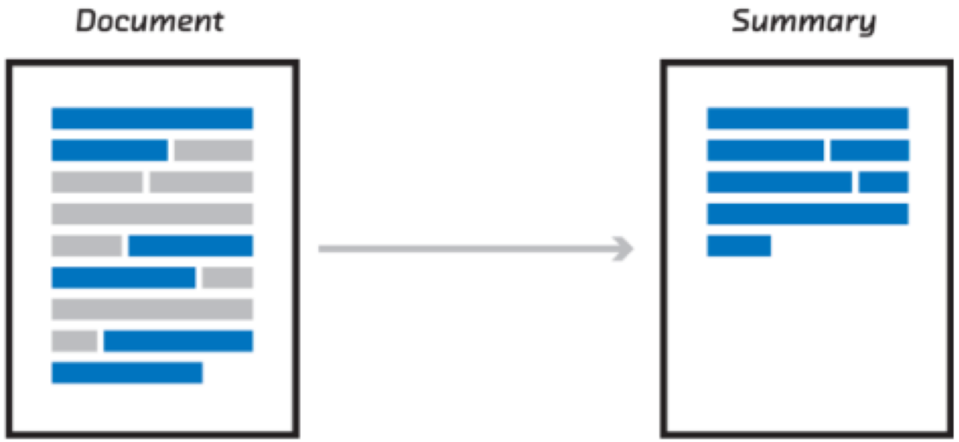
Sequence to Sequence

Language Models – for Generation

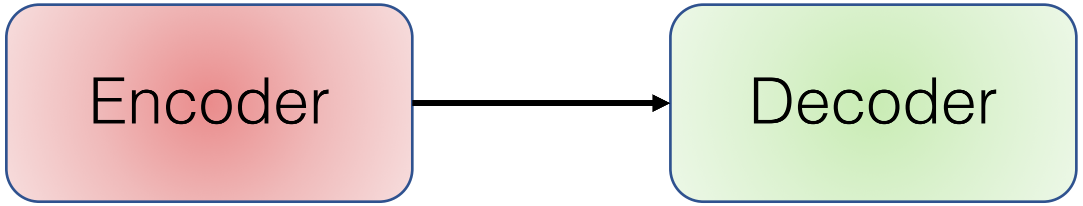




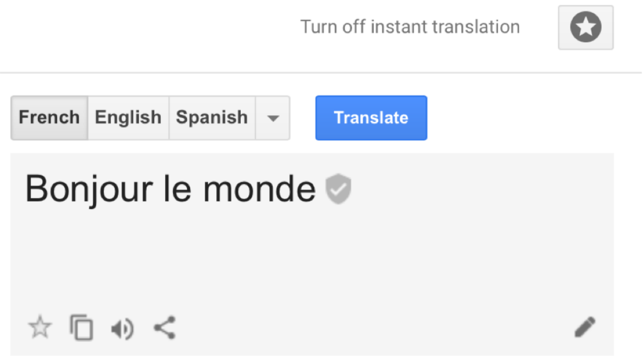
Sequence to Sequence

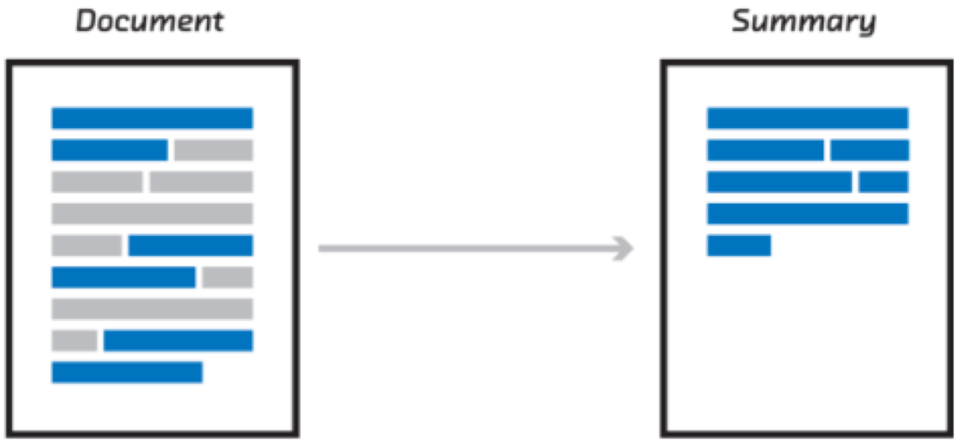


Language Models – for Generation



Sequence to Sequence





Language Models – for Transfer Learning

Pre-trained Language Model → Fine-tuning on Target Task

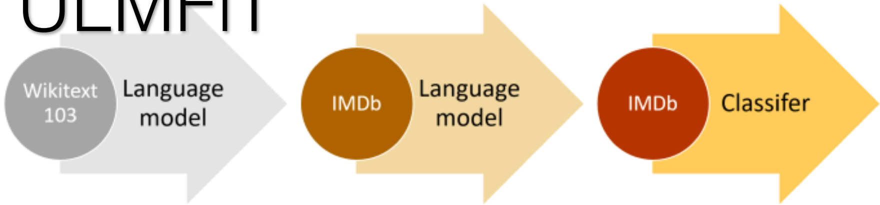
ELMo

BERT

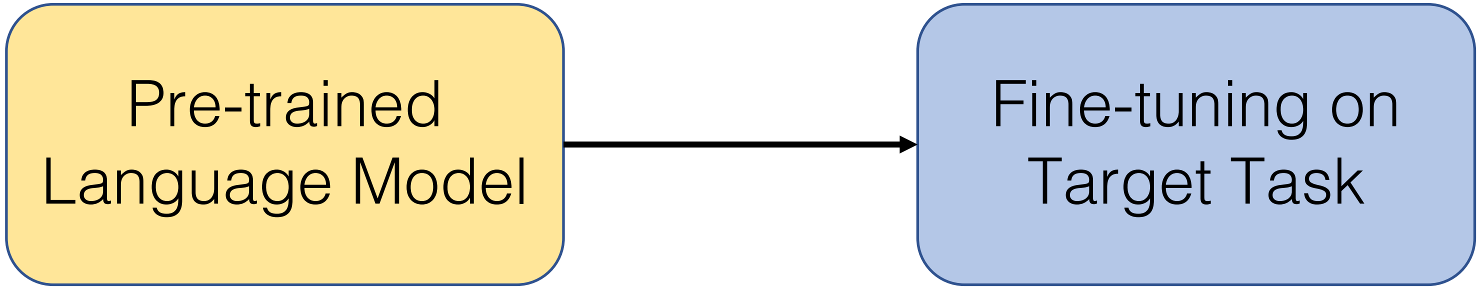
OpenAI

GPT

ULMFiT



Language models – for Transfer Learning



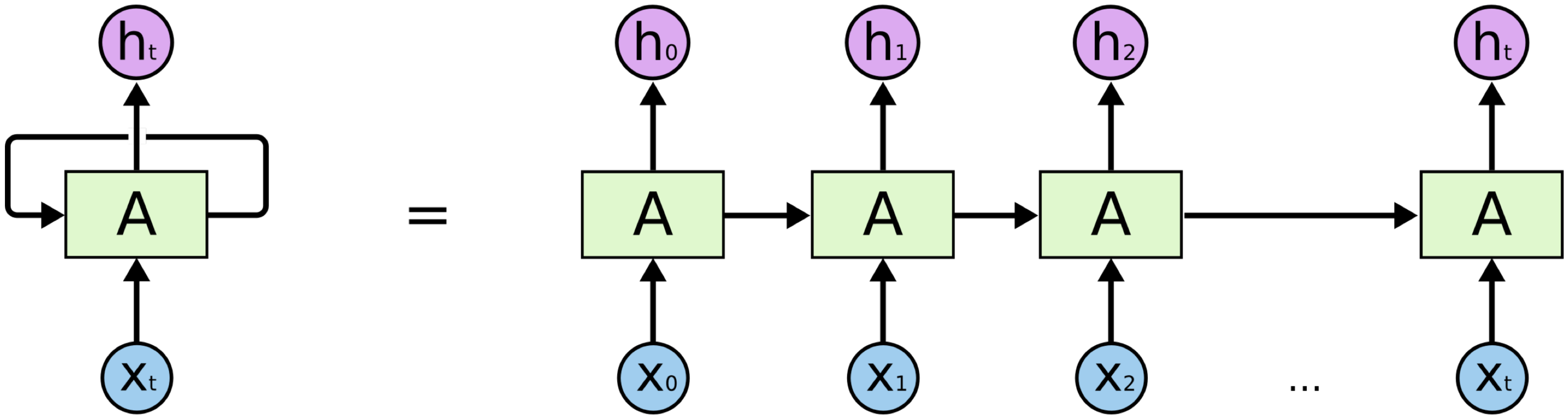
- Large amounts of unlabeled data
- Good downstream task performance without fine-tuning (Radford et al., 2019) or without adding too many task-specific parameters (Devlin et al., 2019)

Analysis of Language Models

Understanding how language models operate allows us to

- Create architectures that encode inductive biases better
- Build explainable models
- Address some legal and policy concerns

Language Models - LSTMs



Language models assign probabilities to sequences of words

Language models assign … words

Figure from colah.github.io

Language Models

assign probabilities to sequences of words

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | \underbrace{w_{t-1}, \ldots, w_1}_{\text{Context}})$$

Language Models

assign probabilities to sequences of words

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | \underbrace{w_{t-1}, \ldots, w_1}_{\text{Context}})$$

N-gram LMs

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, \ldots, w_{t-n+1})$$ *Context Size* = $n - 1$

Language Models

assign probabilities to sequences of words

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | \underbrace{w_{t-1}, \ldots, w_1}_{\text{Context}})$$

N-gram LMs

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, \ldots, w_{t-n+1})$$ *Context Size* = $n - 1$

LSTM LMs

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, \ldots, w_1)$$ *Context Size* = ∞

Language Models

assign probabilities to sequences of words

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | \underbrace{w_{t-1}, \ldots, w_1}_{\text{Context}})$$

N-gram LMs

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, \ldots, w_{t-n+1})$$ *Context Size* = $n - 1$

LSTM LMs

$$P(w_1, \ldots, w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, \ldots, w_1)$$ *Context Size* = ∞

Some things we know about LSTMs

- LSTMs can remember properties such as sentence lengths, word identity and word order
(*Adi et al., 2017*)

- LSTMs can capture syntactic information such as subject-verb agreement
(*Linzen et al., 2016*)

- …and more.

Our goal is to study…

…how LSTM LMs use contextual features, such as word order or word identities, while modeling long sequences.

Our approach

Measure changes in LSTM performance, as a result of perturbing contextual features of the input, during evaluation.

Key Questions

- How much context is used by LSTM LMs?
  About 200 tokens.

- Are nearby and long-range contexts represented differently?
  Yes!

- How do copy mechanisms help the model?
  By copying words from far away.

Setup

- Perturbations applied only during evaluation.
- Datasets (English only): Penn Treebank (PTB) and Wikitext-2 (Wiki).
- Standard LSTM LM (*Merity et al., 2018*).
- All results are reported on the development set.

Evaluation of LMs

- Loss = Negative Log Likelihood (NLL)

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

- Perplexity = $\exp(\mathrm{NLL})$

Key Questions

- How much context is used by LSTM LMs?
  About 200 tokens.

- Are nearby and long-range contexts represented differently?
  Yes!

- How do copy mechanisms help the model?
  By copying words from far away.

How much context?

Effective Context Size: number of tokens of context such that

$$-\frac{1}{T}\sum_{t=1}^{T}\log P(w_t|w_{t-1},\ldots,w_{\text{ecs}}) \approx -\frac{1}{T}\sum_{t=1}^{T}\log P(w_t|w_{t-1},\ldots,w_1)$$

… Language models assign probabilities to sequences of <u>words</u>

How much context?

Effective Context Size: number of tokens of context such that

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

… Language models assign probabilities to sequences of <u>words</u>

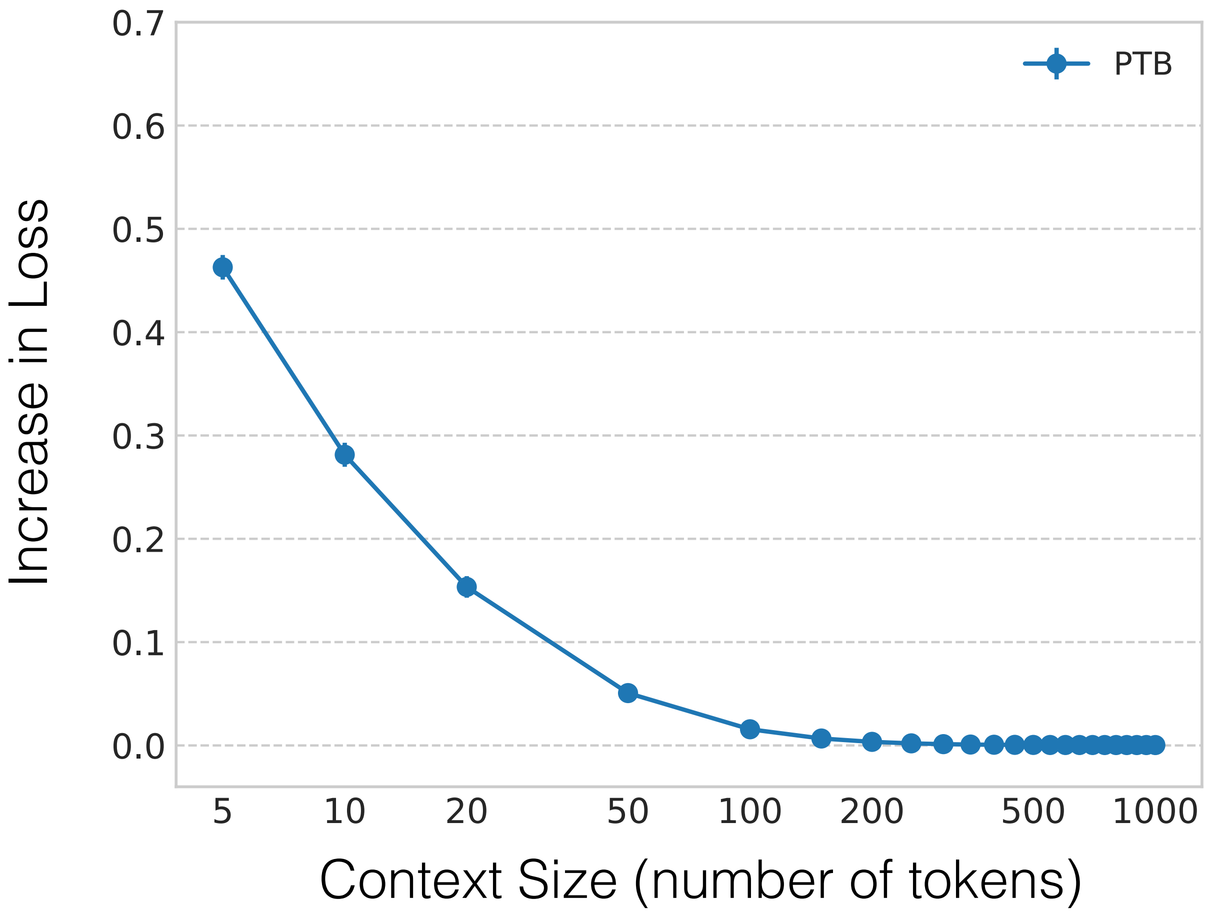
How much context?

Effective Context Size: number of tokens of context such that

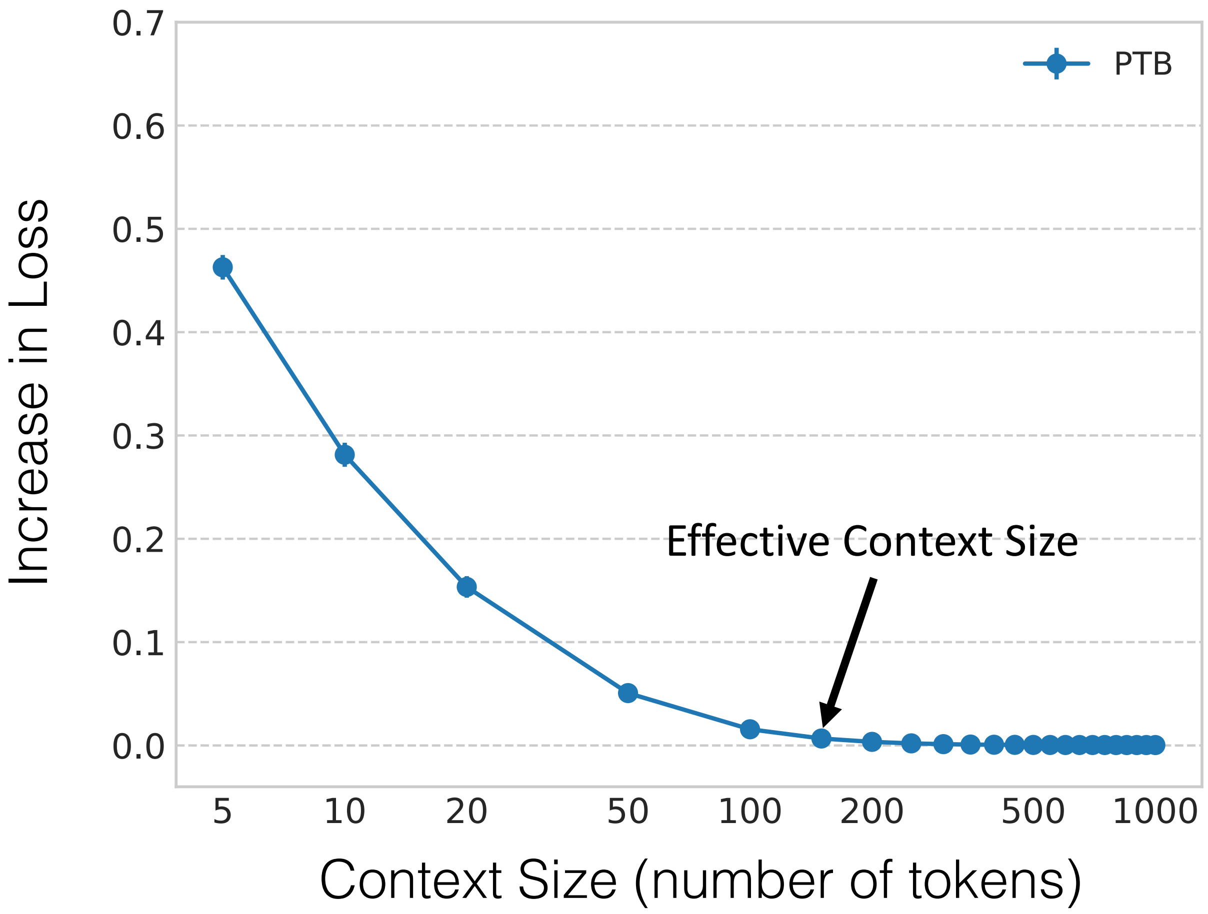
$$-\frac{1}{T}\sum_{t=1}^{T}\log P(w_t|w_{t-1},\ldots,w_{\text{ecs}}) \approx -\frac{1}{T}\sum_{t=1}^{T}\log P(w_t|w_{t-1},\ldots,w_1)$$

~~… Language models assign~~ probabilities to sequences of words

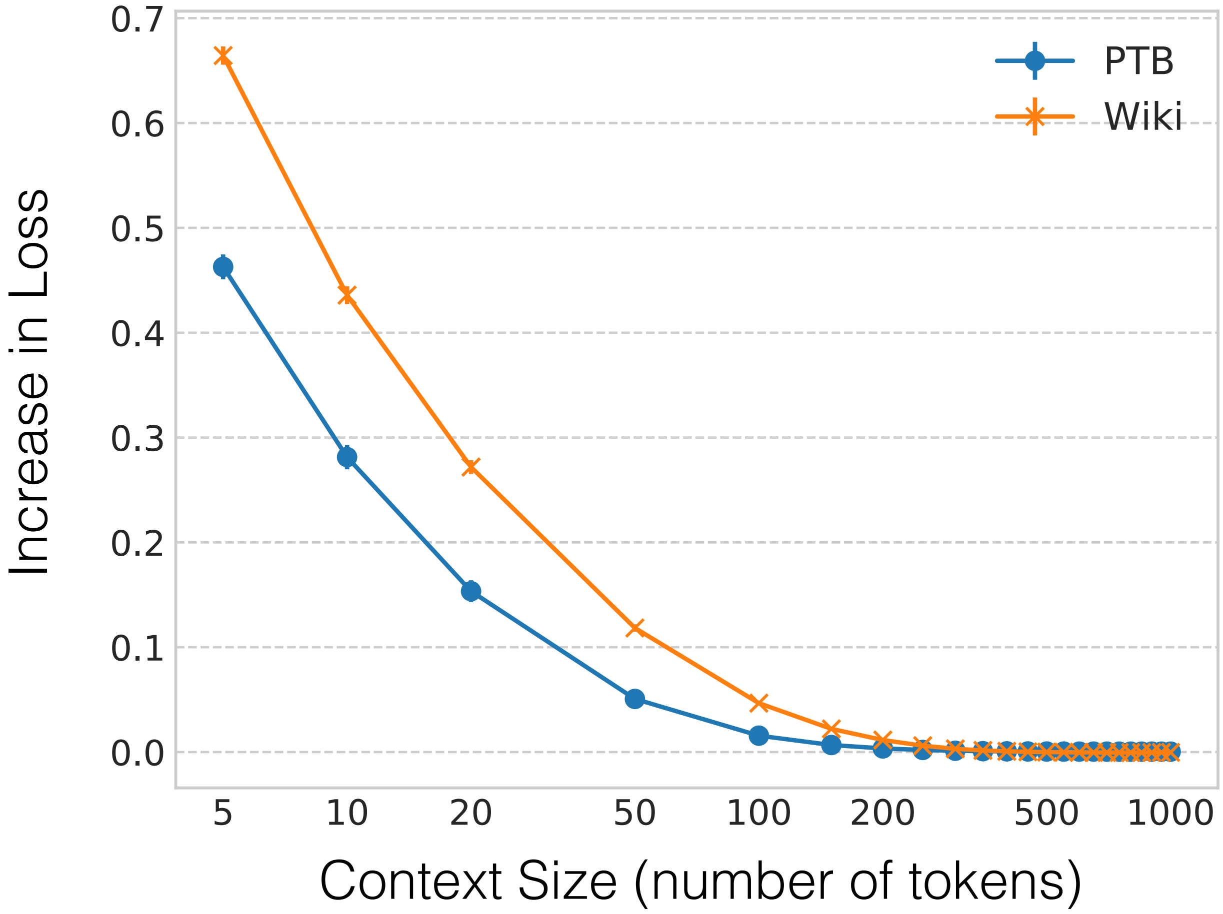
LSTM language models have an effective context size of about 200 on average



LSTM language models have an effective context size of about 200 on average



LSTM language models have an effective context size of about 200 on average

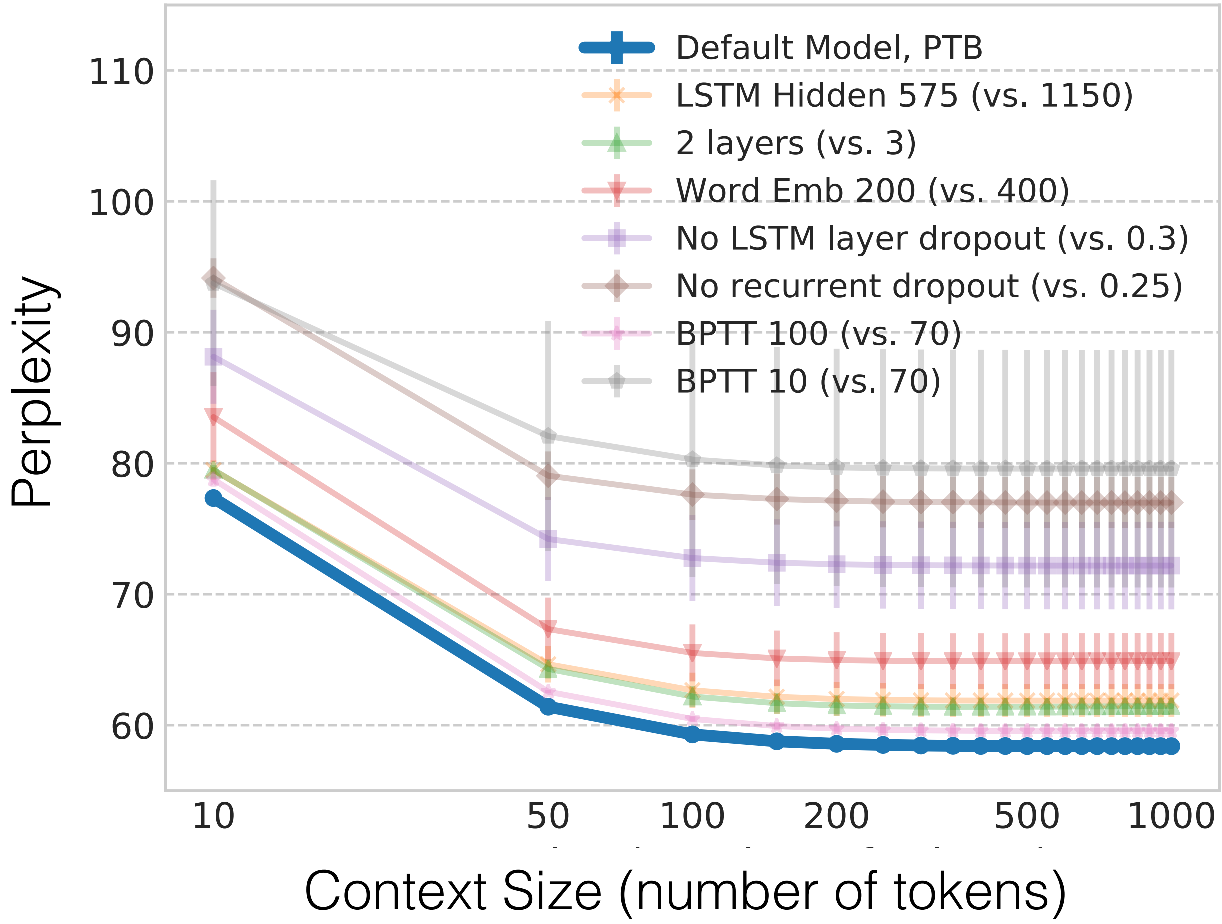


What about hyperparameters?

The model is trained with specific hyperparameters. But what if we changed them?

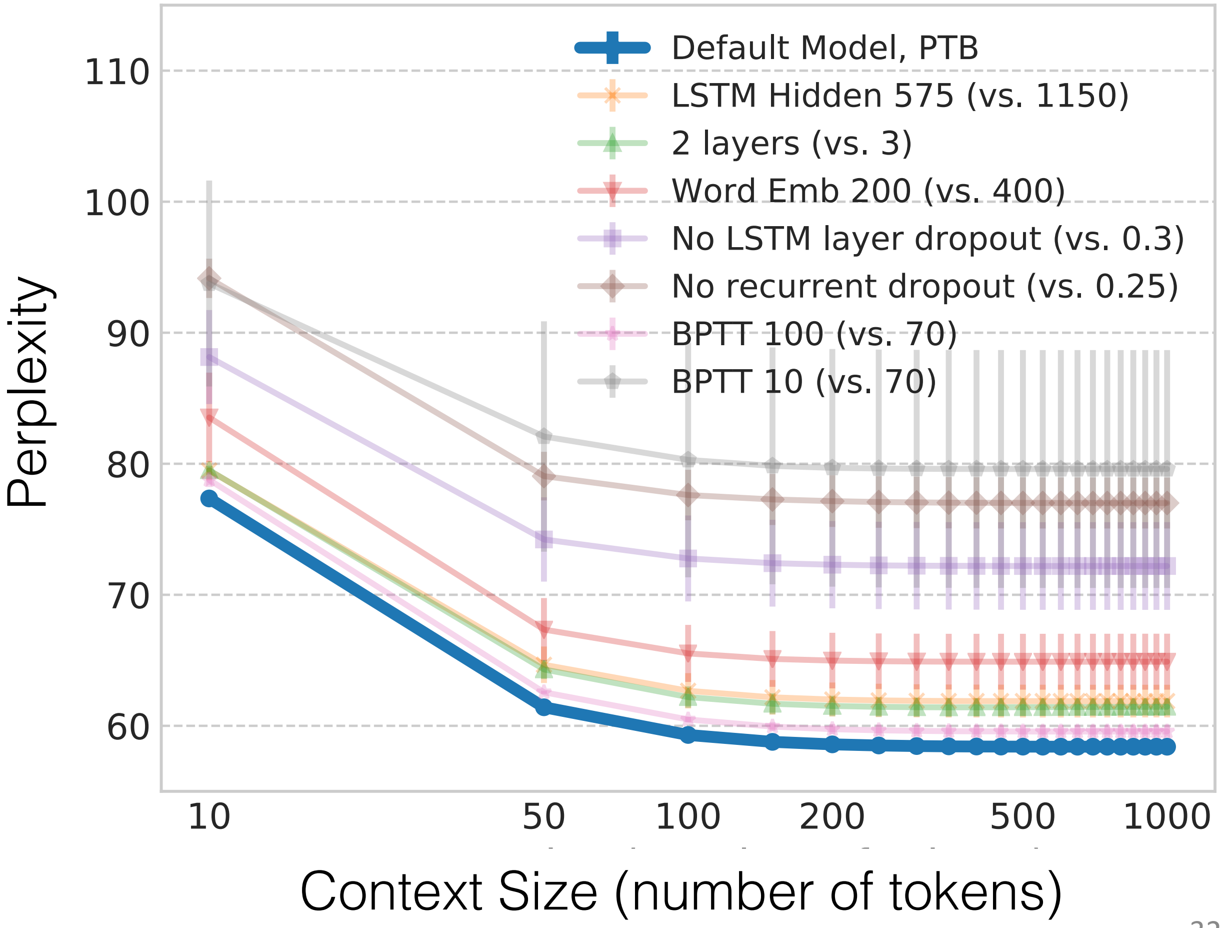
- Does the amount of dropout matter?
- Does the size of the hidden states or the word embeddings matter?
- Does the number of timesteps used backpropagation matter?

Changing the model hyperparameters does not change the effective context size



Changing the model hyperparameters does not change the effective context size

- Default model has best performance.
- Changing hyperparameters changes perplexity – models are clearly different
- Trend for effective context size remains the same

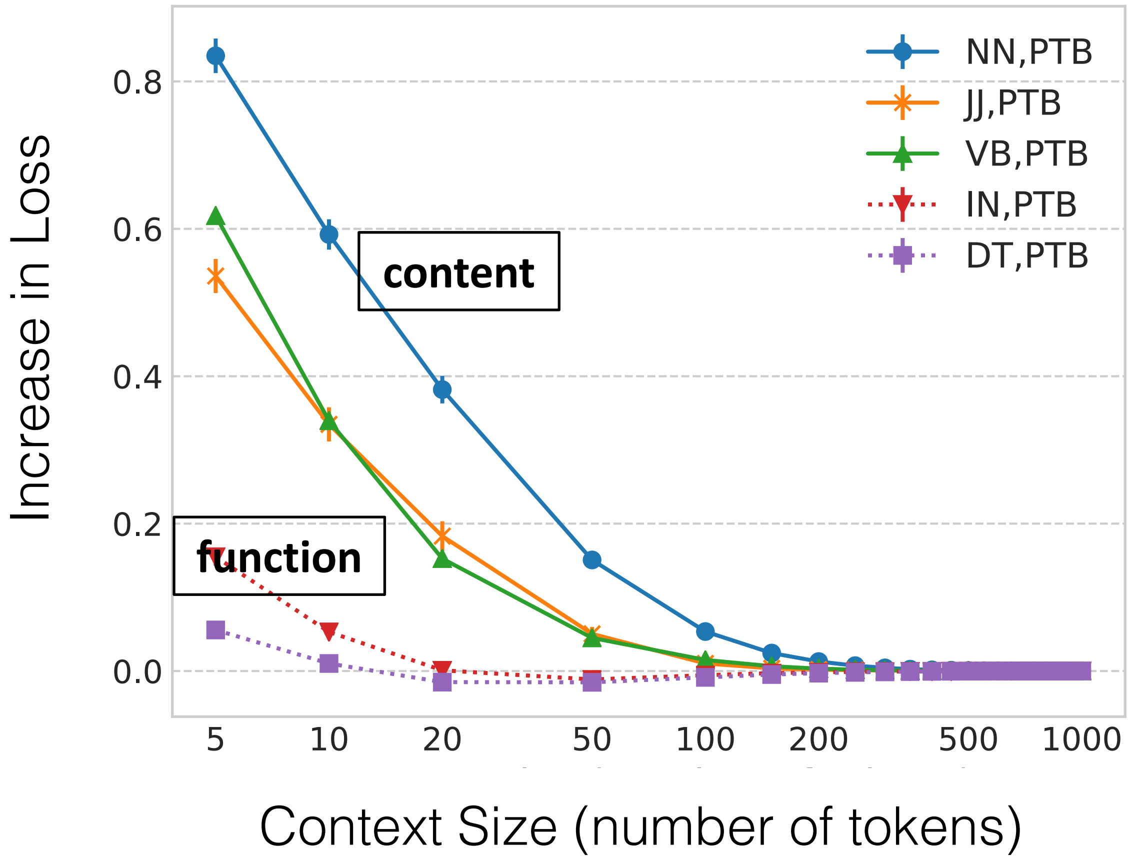


Does the target word's type matter?

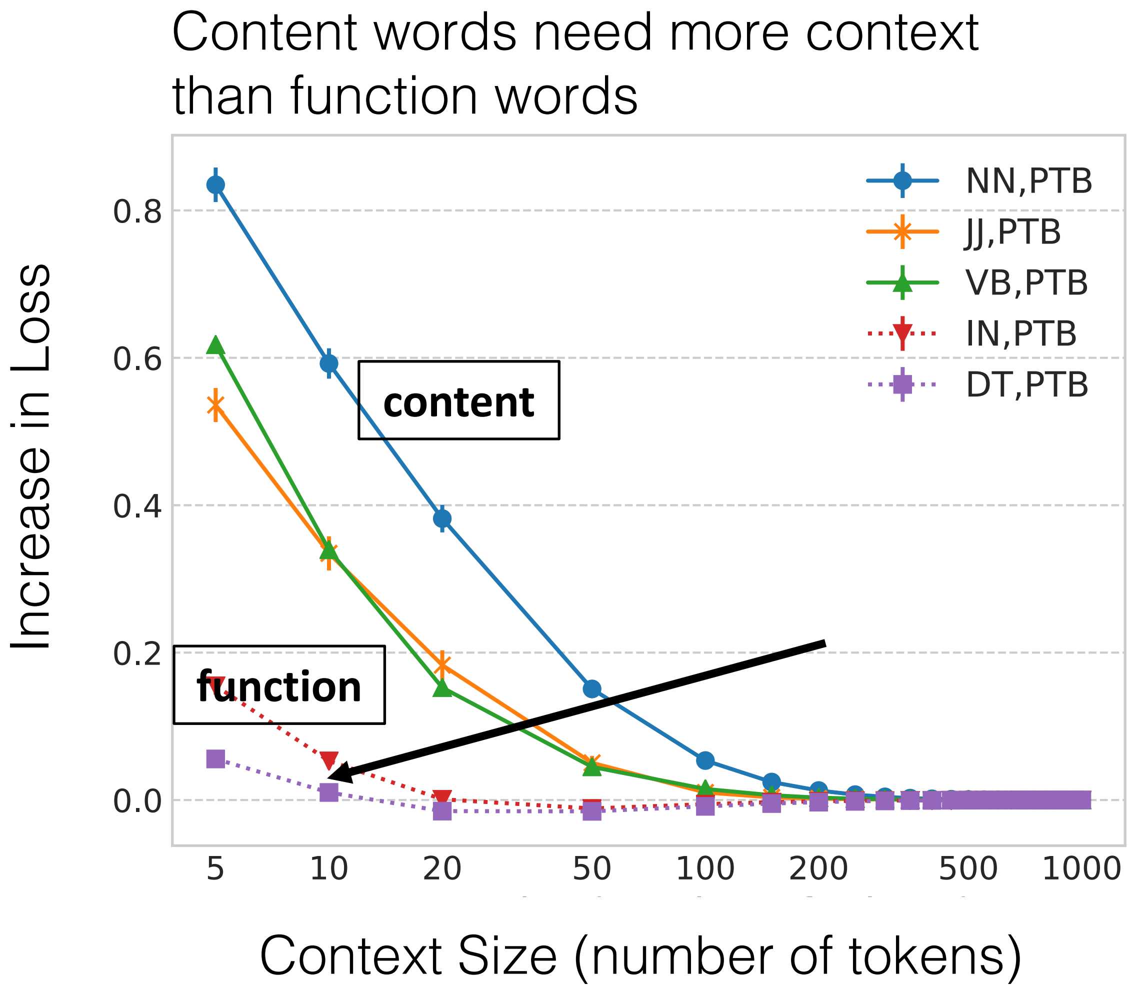
Nouns are not the same as determiners. Does the model know this?

The LSTM’s effective context size is dynamic and depends on the target word

Content words need more context than function words

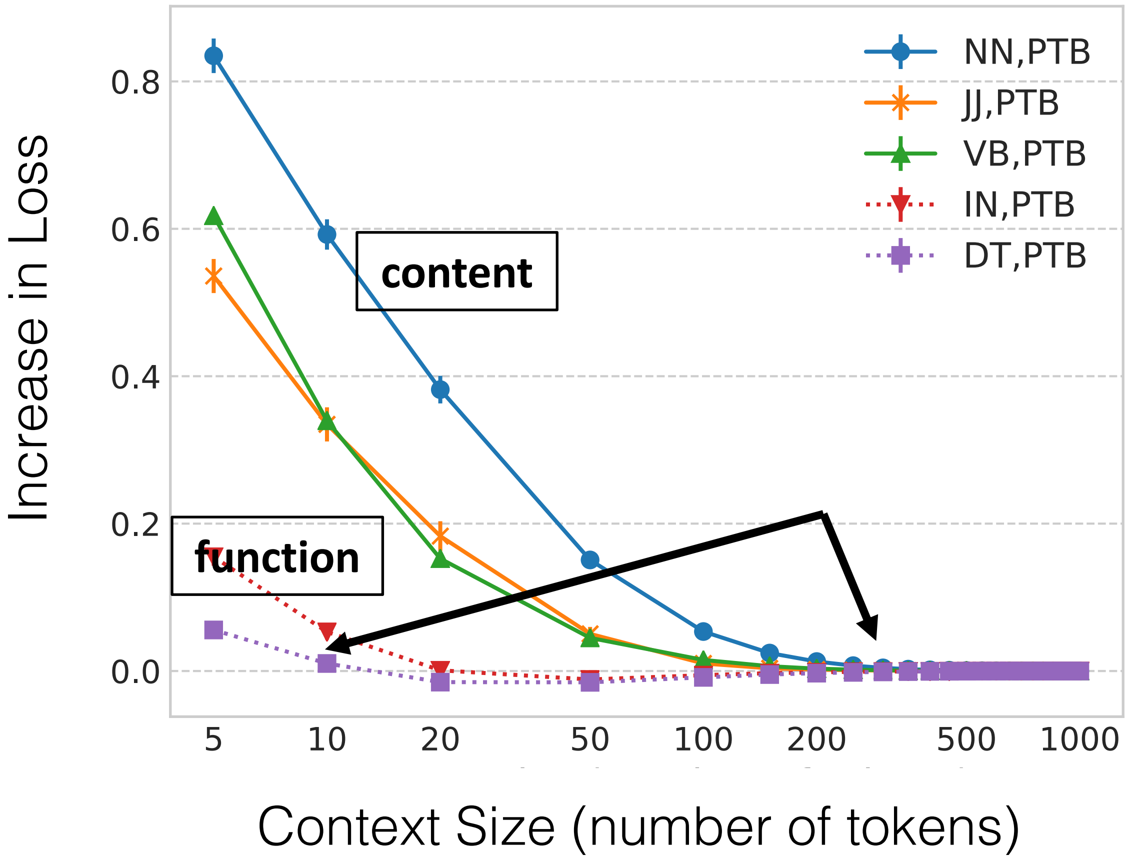


The LSTM’s effective context size is dynamic and depends on the target word



The LSTM's effective context size is dynamic and depends on the target word

Content words need more context than function words



Key Questions

- How much context is used by LSTM LMs?

  About 200 tokens.

  Agnostic to changes in hyperparameters.

  Context use is dynamic.

- Are nearby and long-range contexts represented differently?

  Yes!

- How do copy mechanisms help the model?

  By copying words from far away.

Key Questions

- How much context is used by LSTM LMs?

  About 200 tokens.

  Agnostic to changes in hyperparameters

  Context use is dynamic

- Are nearby and long-range contexts represented differently?

  Yes!

- How do copy mechanisms help the model?

  By copying words from far away.

Does word order matter?

Local Word Order: Order within 20 token spans (about the length of a sentence)

In this analytic study , we investigate the use of context by LSTM language models , using ablations . A language model assigns probabilities to sequences of words

Does word order matter?

Local Word Order: Order within 20 token spans (about the length of a sentence)

In this analytic study , we probabilities the by of investigate using to context sequences models language . assigns use model A LSTM language ablations , of words

Does word order matter?

Local Word Order: Order within 20 token spans (about the length of a sentence)

In this analytic study , we probabilities the by of investigate using to context sequences models language . assigns use model A LSTM language ablations , of words

Global Word Order: Order within the entire context

In this analytic study , we investigate the use of context by LSTM language models , using ablations . A language model assigns probabilities to sequences of words

Does word order matter?

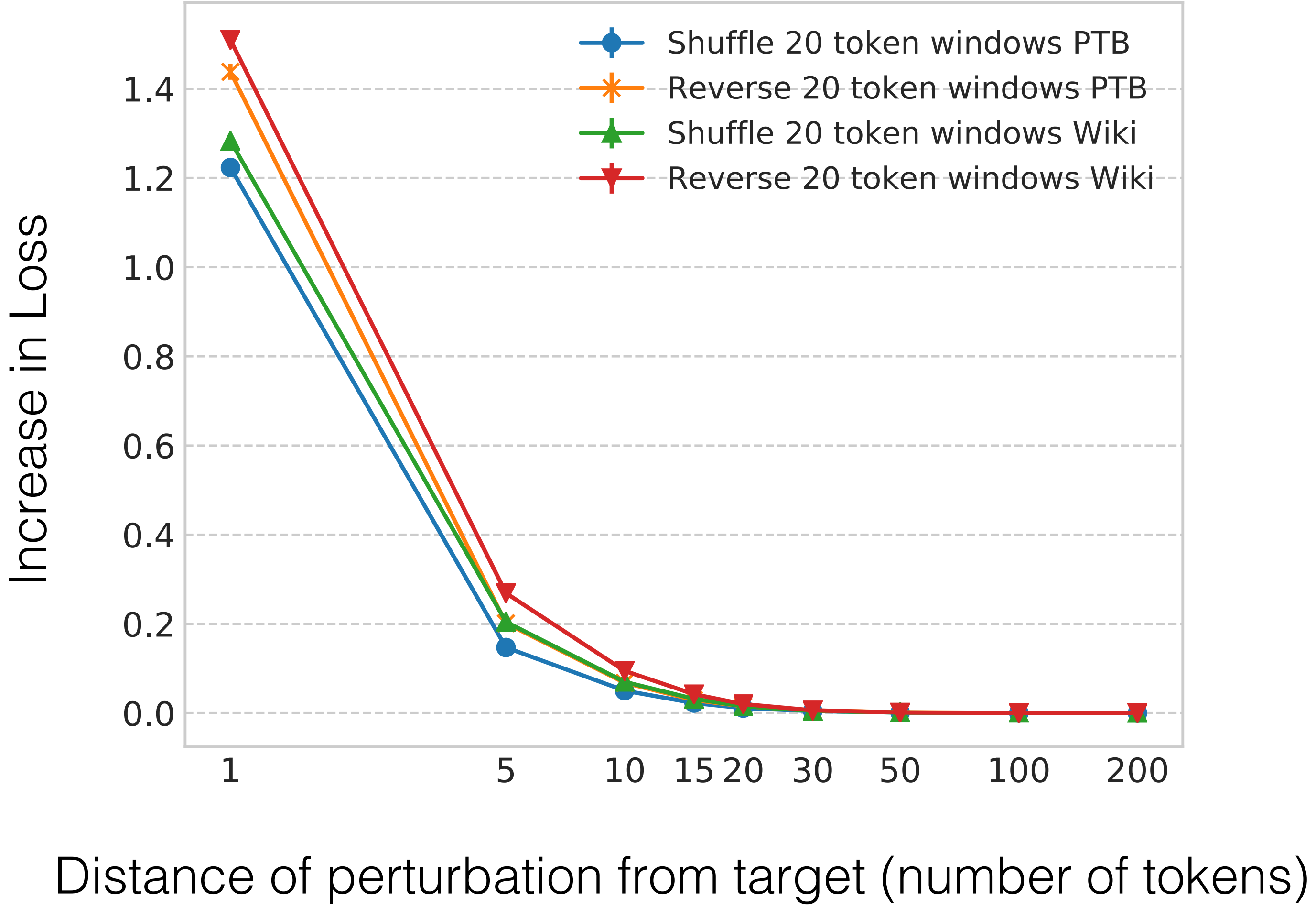
Local Word Order: Order within 20 token spans (about the length of a sentence)

In this analytic study , we probabilities the by of investigate using to context sequences models language . assigns use model A LSTM language ablations , of words

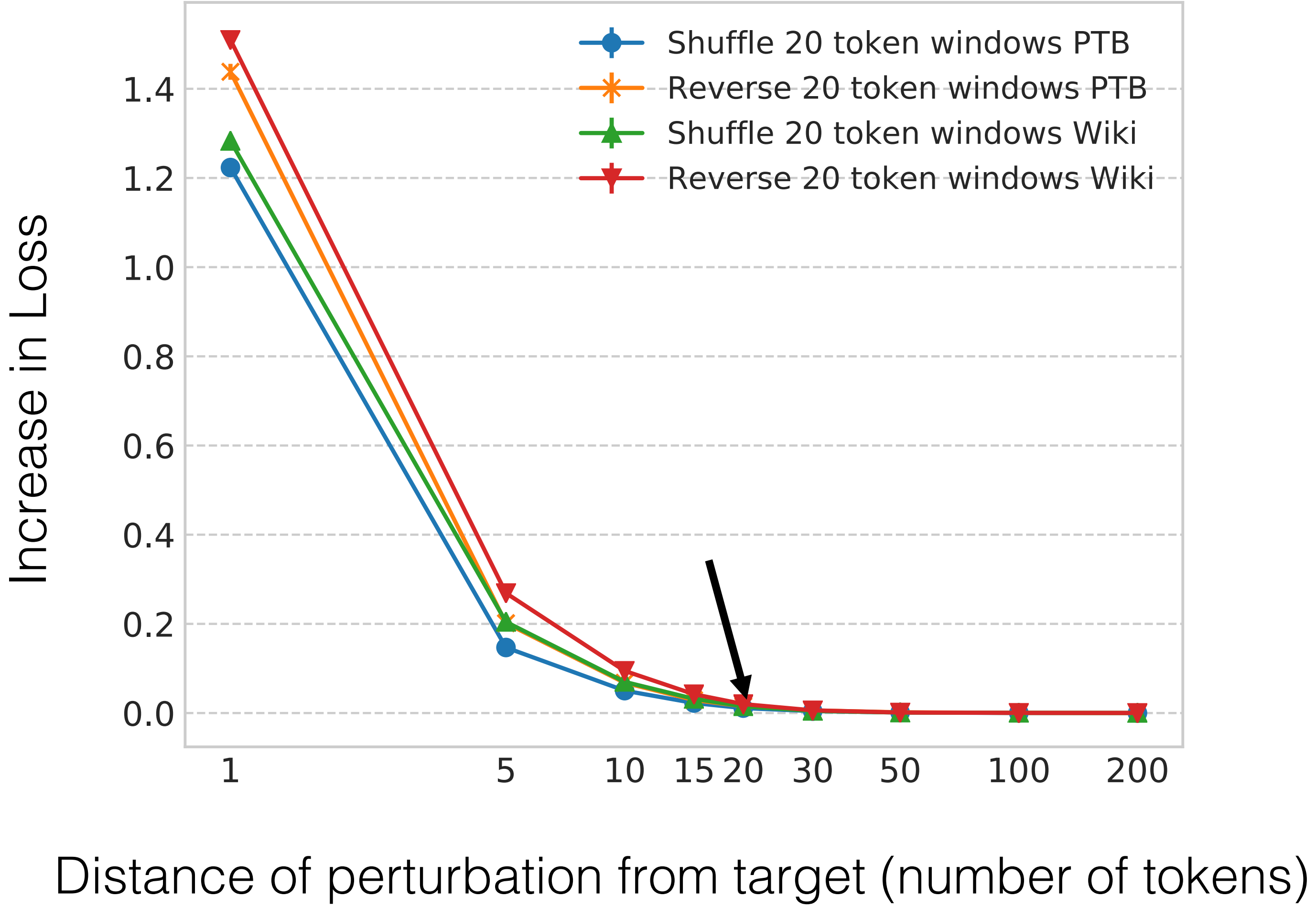
Global Word Order: Order within the entire context

investigate analytic by , model , use this using In sequences we of models language to A study . Context ablations language the assigns probabilities LSTM of words

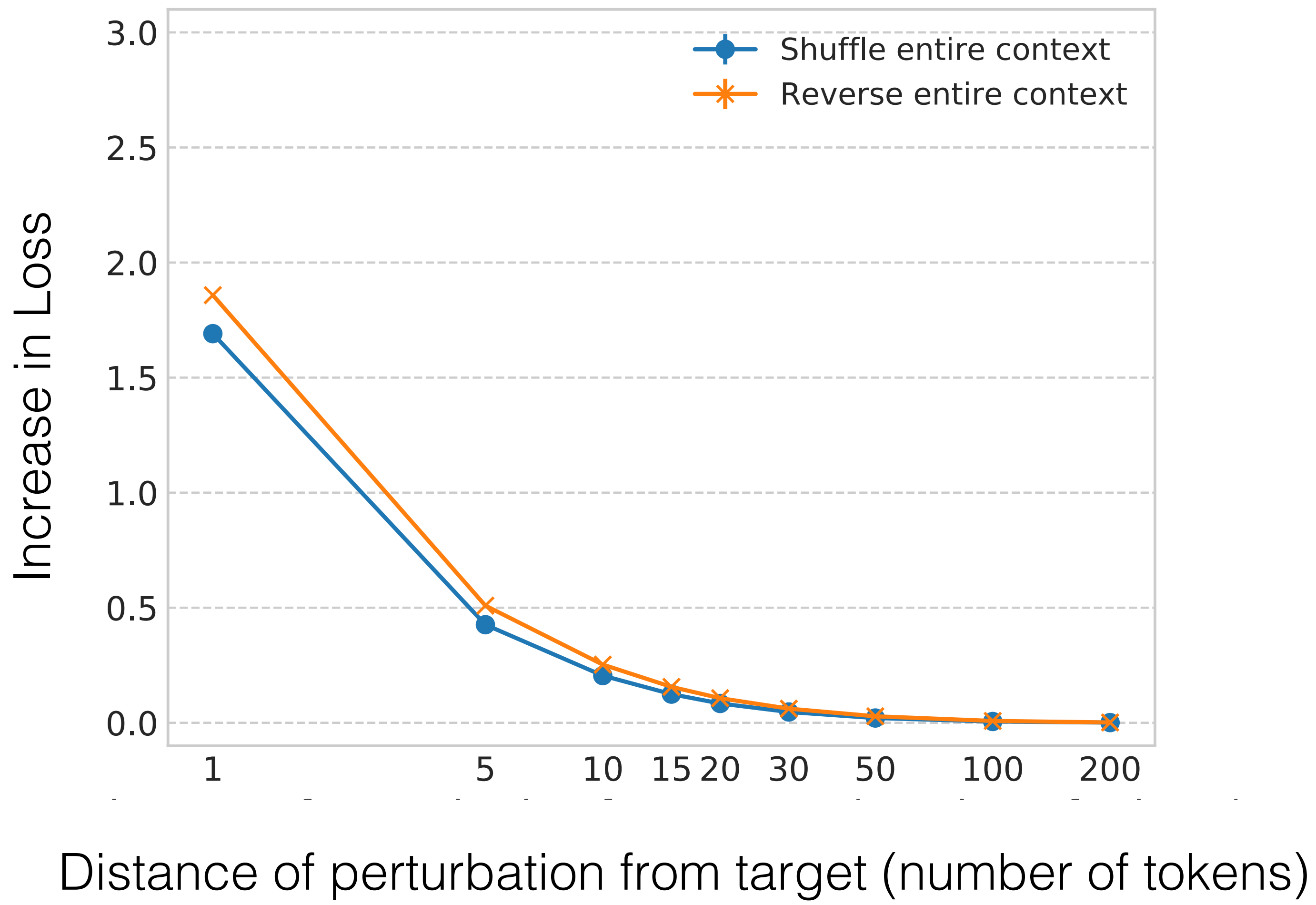
Local word order only matters for the first 20 tokens



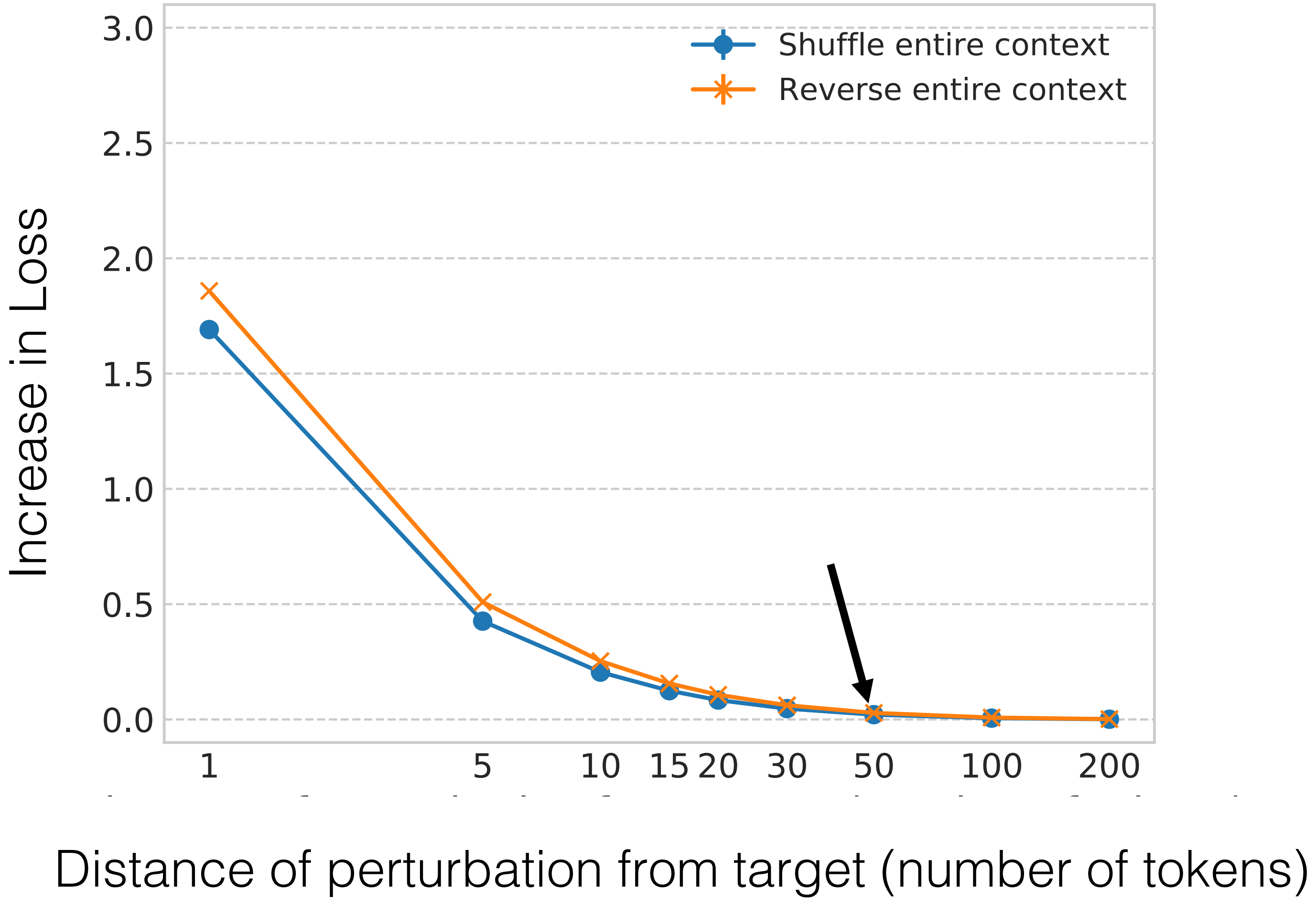
Local word order only matters for the first 20 tokens



Global word order only matters for the most recent 50 tokens



Global word order only matters for the most recent 50 tokens



Replace context with random train set sequence

In this analytic study , we investigate the use of context by LSTM language models ,

using ablations . A language model assigns probabilities to sequences of words

Replace context with random train set sequence

In this analytic study , we investigate the use of context by LSTM language models ,

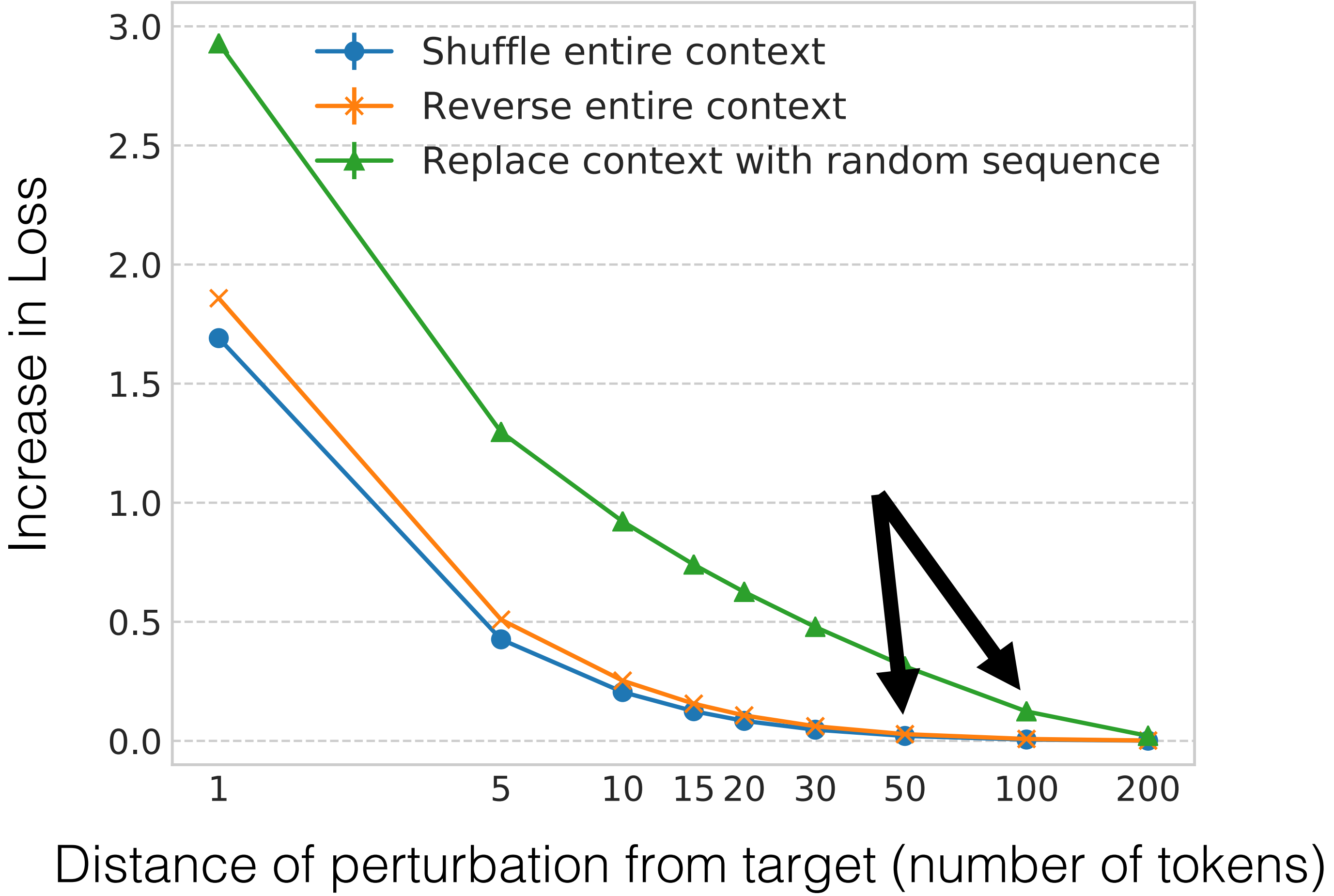
using ablations . A language model assigns probabilities to sequences of *words*

Replace context with random train set sequence

In this analytic study , we investigate the use of context by LSTM language models , using ablations . A language model assigns probabilities to sequences of <u>words</u>

Iron Man is a character in the Marvel universe . He joined forces with other Marvel characters to form the Avengers – Earth 's mightiest heroes of <u>words</u>

Global word order only matters for the most recent 50 tokens



Key Questions

- How much context is used by LSTM LMs?
  About 200 tokens.
  Agnostic to changes in hyperparameters.
  Context use is dynamic.

- Are nearby and long-range contexts represented differently?
  Word order matters nearby.
  Long-range context is modeled as a rough semantic field/topic.

- How do copy mechanisms help the model?
  By copying words from far away.

Key Questions

- How much context is used by LSTM LMs?

  About 200 tokens.
  Agnostic to changes in hyperparameters.
  Context use is dynamic.

- Are nearby and long-range contexts represented differently?

  Word order matters nearby.
  Long-range context is modeled as a rough semantic field/topic.

- How do copy mechanisms help the model?

  By copying words from far away.

Can LSTMs copy words without external copy mechanisms?

Attention
Cache

… on sequences of words most of the time … probabilities to sequences of <u>words</u>

Can LSTMs copy words without external copy mechanisms?

Attention
Cache

… on sequences of words most of the time … probabilities to sequences of words

Three classes of target words

1. Appear in their own nearby context (within 50 tokens).

… Langauge models operate on sequences of words most of the time . A language model assigns probabilities to sequences of words

Three classes of target words

1. Appear in their own nearby context (within 50 tokens).

2. Appear only in their long-range context (beyond 50 tokens).

… words … deep … hype … <token 51, token 50, token 49> … assigns probabilities to sequences of words

Three classes of target words

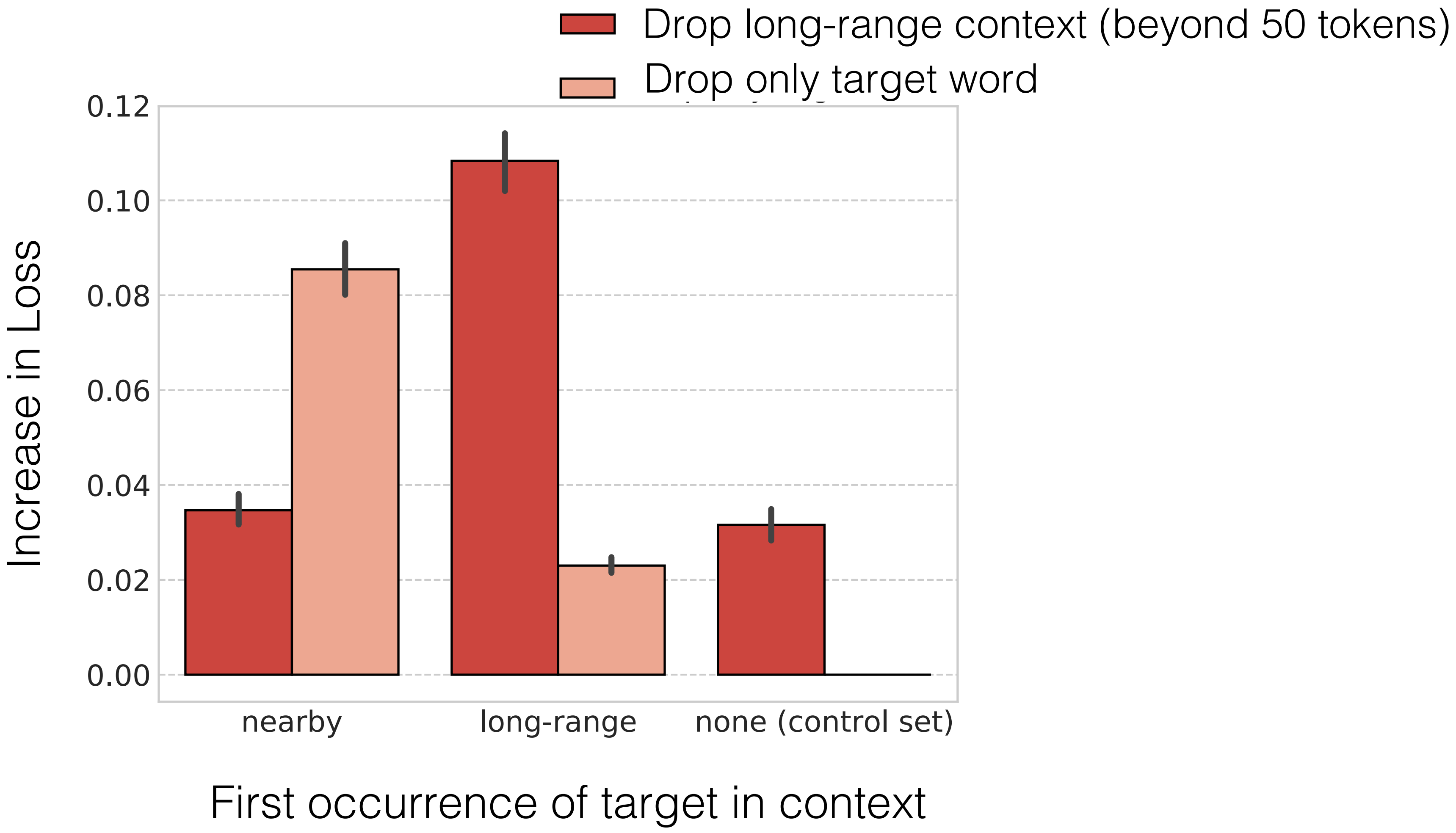
1. Appear in their own nearby context (within 50 tokens).
2. Appear only in their long-range context (beyond 50 tokens).
3. Never appear in their own context, ever (none).

Drop target words

… words …words … operate on sequences of words most of the time . A language model assigns probabilities to sequences of words

… ~~words~~ …~~words~~ … operate on sequences of ~~words~~ most of the time . A language model assigns probabilities to sequences of words

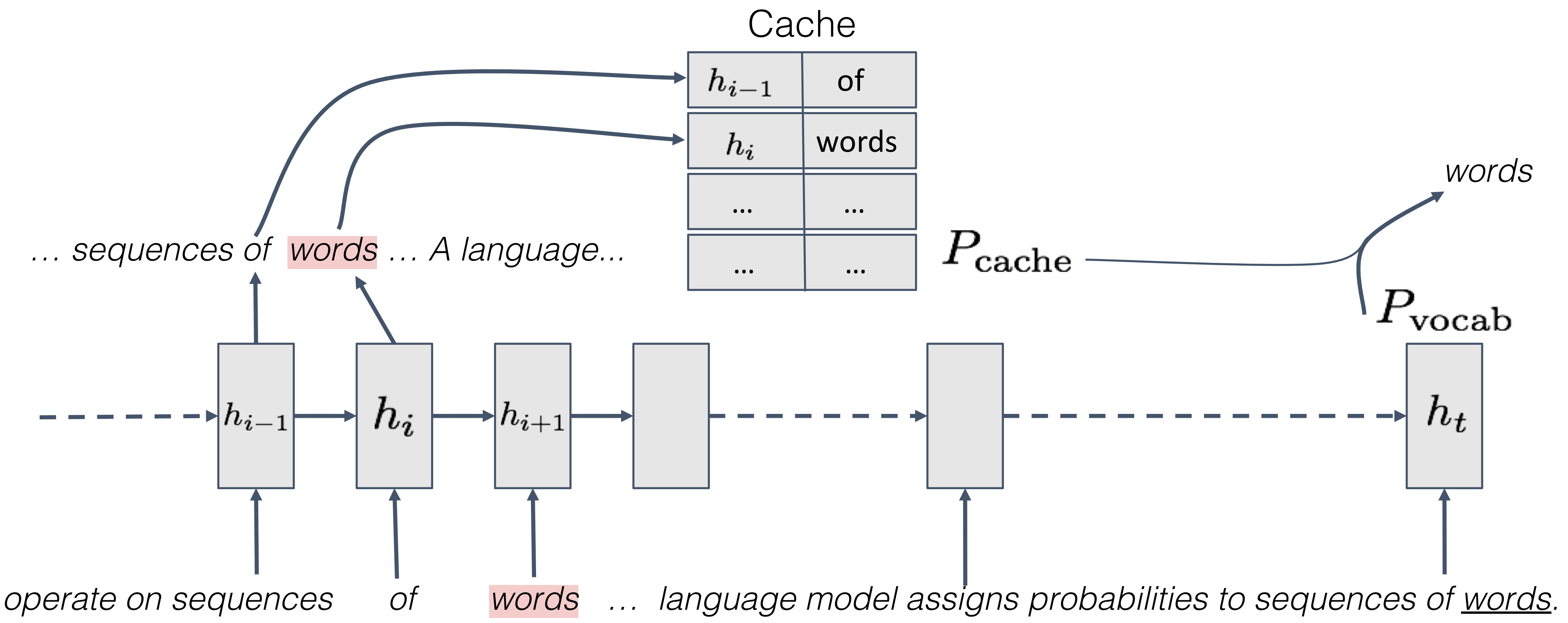
LSTM LMs can regenerate words seen in nearby context



How do external copy mechanisms help?

In this study, we consider the Neural Caching Model (Grave et al., 2017)

Neural Caching Model (Grave et al., 2017)

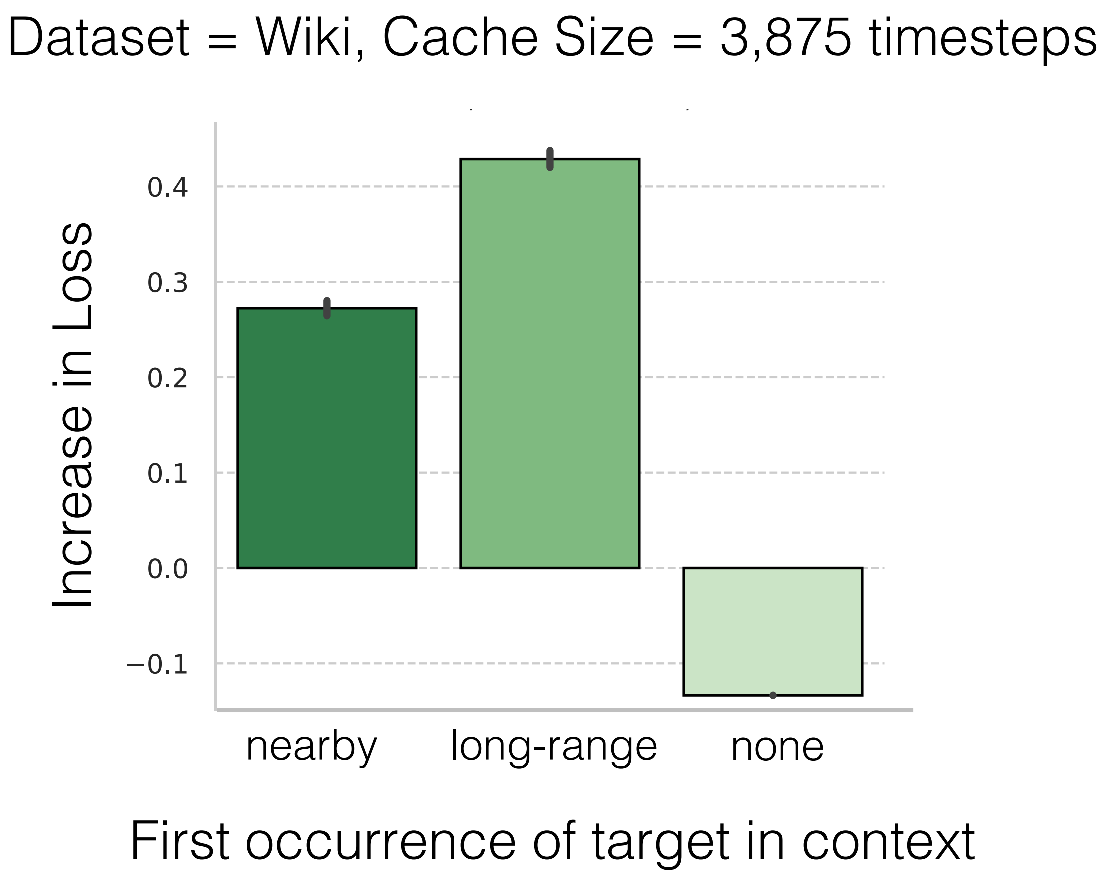


$$P_{\text{cache}}(w_t | w_{t-1}, \ldots, w_1; h_t, \ldots, h_1) \propto \sum_{i=1}^{t-1} \mathbb{1}[w_i = w_t] \exp(\theta h_i^T h_t)$$

Three classes of target words

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

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



Neural Cache Success and Failure Examples

Success:

La Fortuna , Mexico . UNK just off the coast of Mexico , the system interacted with land and began weakening . UNK later , convection rapidly diminished as dry air became entrained in the circulation . In response to quick degradation of the system 's structure , the NHC downgraded UNK to a tropical storm . Rapid weakening continued throughout the day and by the evening hours , the storm no longer had a defined circulation . Lacking an organized center and deep convection , the final advisory was issued on UNK . The storm 's remnants persisted for several more hours before dissipating roughly 175 mi (280 km) southwest of Cabo Corrientes , Mexico . </s> </s> = = Preparations and impact = = </s> </s> Following the classification of Tropical Depression Two @-@ E on June 19 , the Government of Mexico issued a tropical storm warning for coastal areas between UNK and Manzanillo . A hurricane watch was also put in place from UNK de UNK to Punta San UNK . Later that day , the tropical storm warning was upgraded to a hurricane warning and the watch was extended westward to La **Fortuna**

Failure:

) . Standing roughly 15 metres (49 ft) away , the cadres now raised their weapons . " You have taken our land , " one of them said . " Please don 't shoot us ! " one of the passengers cried , just before they were killed by a sustained burst of automatic gunfire . </s> Having collected water from the nearby village , UNK and his companions were almost back at the crash site when they heard the shots . UNK it was personal ammunition in the luggage exploding in the heat , they continued on their way , and called out to the other passengers , who they thought were still alive . This alerted the insurgents to the presence of more survivors ; one of the guerrillas told UNK 's group to " come here " . The insurgents then opened fire on their general location , prompting UNK and the others to flee . Hill and the UNK also ran ; they revealed their positions to the fighters in their UNK , but successfully hid themselves behind a ridge . After Hill and the others had hidden there for about two **hours**

Key Questions

- How much context is used by LSTM LMs?
  About 200 tokens.
  Agnostic to changes in hyperparameters.
  Context use is dynamic.

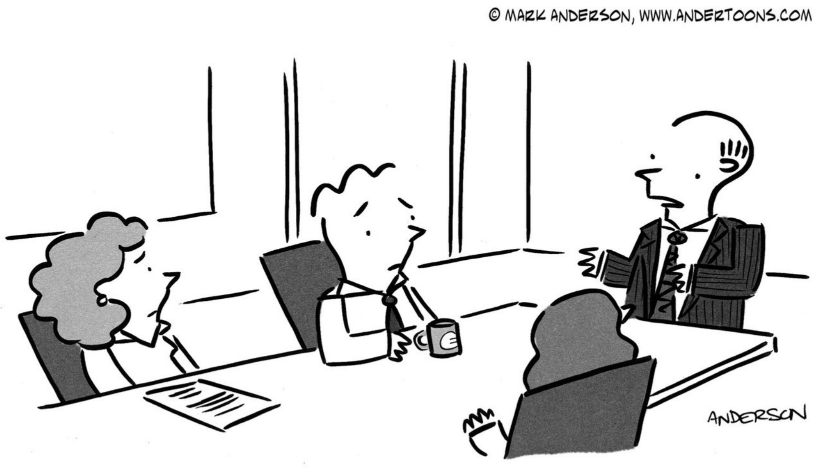
- Are nearby and long-range contexts represented differently?
  Word order matters nearby.
  Long-range context is modeled as a rough semantic field/topic.

- How do copy mechanisms help the model?
  LSTM LMs can regenerate words from nearby.
  Neural cache can copy from far away.

What's next?

- Improving existing models.
- Compare model classes on more than test set perplexities.
- Can we decouple the data from the models?
  - Experiment with a variety of model classes
  - Experiment on many different languages
- Theoretical justifications for LSTM behavior.

Thank You!



"To make a long story short, what it all boils down to in the final analysis is that what you should take away from this is…"

- How much context is used by LSTM LMs?
  About 200 tokens.
  Agnostic to changes in hyperparameters.
  Context use is dynamic.
- Are nearby and long-range contexts represented differently?
  Word order matters nearby.
  Long-range context is modeled as a rough semantic field/topic.
- How do copy mechanisms help the model?
  LSTM LMs can regenerate words from nearby.
  Neural cache can copy from far away.

Paper: https://nlp.stanford.edu/pubs/khandelwal2018lm.pdf
Code: https://github.com/urvashik/lm-context-analysis