LSTMs Exploit Linguistic Attributes of Data

Nelson F. Liu[♠♡], Omer Levy[♠], Roy Schwartz[♠♣], Chenhao Tan[♢], Noah A. Smith[♠♣]

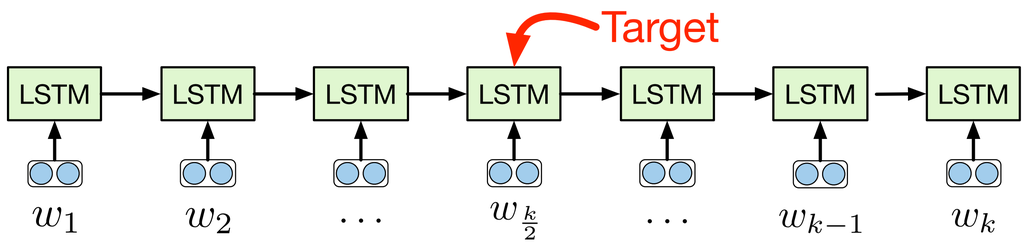
♠Paul G. Allen School of Computer Science & Engineering; ♡Department of Linguistics, University of Washington;
♣Allen Institute for Artificial Intelligence; ♢Department of Computer Science, University of Colorado Boulder

TL;DR

- **Data with linguistic attributes helps LSTMs learn a non-linguistic memorization task.**
- **To solve the task, LSTMs use individual neurons to count timesteps.**
- We hypothesize that LSTMs pick up on the patterns and structure in linguistic data and use them as additional noisy training signal.

Testbed Memorization Task

- Given a **constant-length** sequence of tokens, predict the identity of the middle token seen.



- This task is inherently **nonlinguistic**.

[g m d r p **j** w f h c]
[3 5 6 8 4 **0** 2 7 9 1]
[🌒 🍹 📽️ ↩️ ☁️ **⛰️** 📃 🏕️ 🐡 🚵]

Training Datasets with Various Linguistic Attributes

1. *Language* setting

Directly take sequences

[Pierre Vinken , 61]
[years old , will]
[join the board as]

2. *n-gram* setting

[Pierre Vinken] [, 61] [years old]
[, will] [join the] [board as] …

1. Chunk corpus into pieces of size n (n = 2 in this example)

2. Permute the chunks

3. Split into sequences of desired length (4 in this example)

[join the Pierre Vinken]
[years old board as]
[, 61 , will]

3. *Uniform* setting

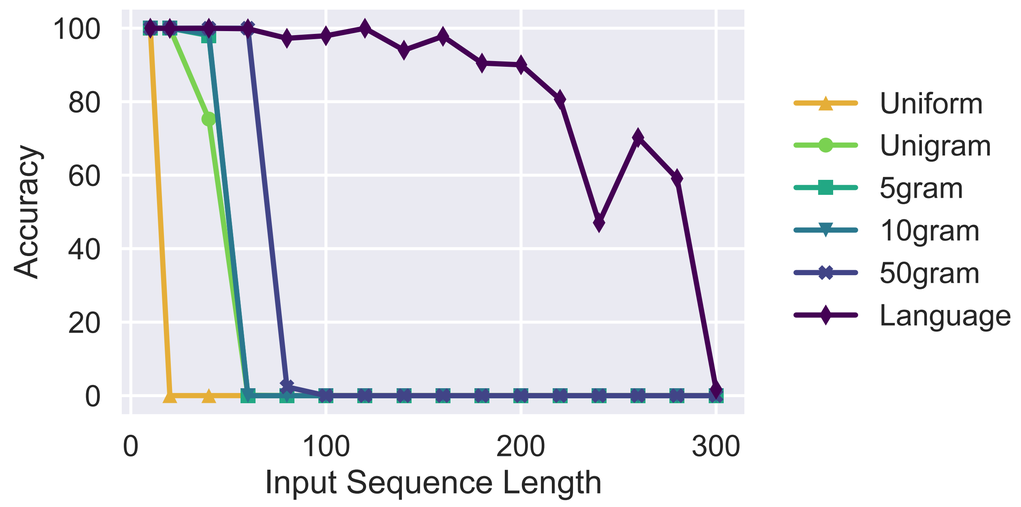
Randomly pick tokens from vocabulary

[volume vice , a]
[pilots or sign corp.]
[amid in <unk> banks]

Experiments

- Test data: uniform distribution over the 100 rarest words in the PTB.
- **Ensures that models truly generalize and are not just using training data-specific features.**

Models trained on data with linguistic features generalize better



What happens if we add more hidden units?

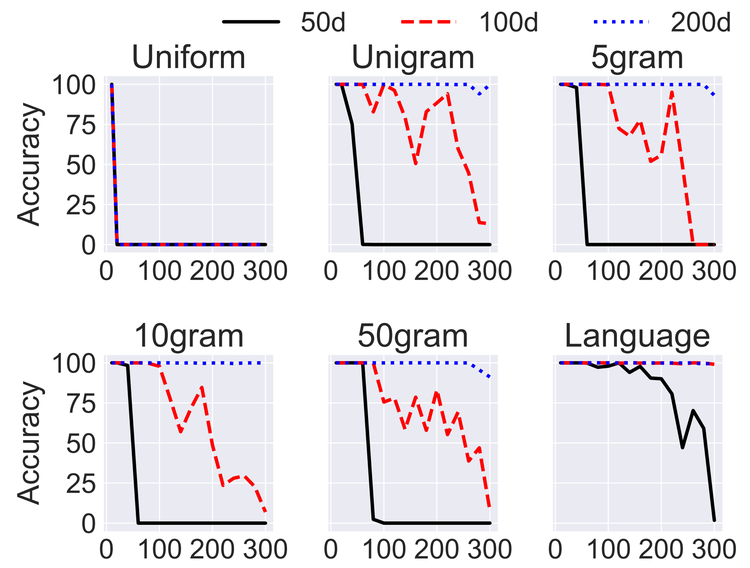
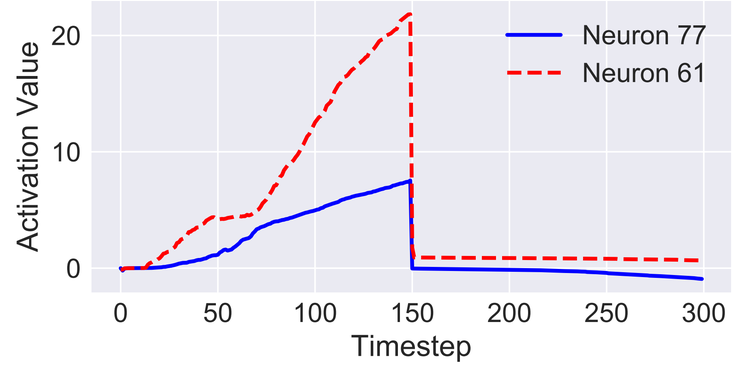


Figure 2: Models trained on **Uniform** data do poorly, even with more hidden units.

Further Analysis

- We further study an LSTM with 100 hidden units trained on **Language**, where train and test sequences are of length 300.

To solve the task, RNNs learn to count



The RNN exploits linguistic features to bootstrap itself early in training and learns to generalize later.

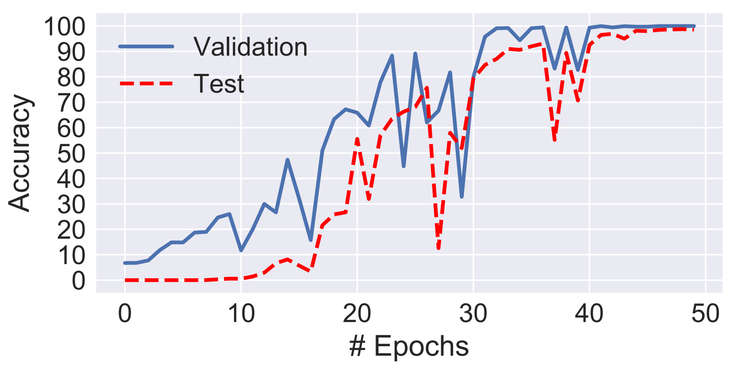


Figure 4: Validation set has the same distribution as train. We see validation accuracy improves faster than test at early epochs (exploiting linguistic features), but the two curves move in unison in later epochs (true memorization).

Code: git.io/lstms-exploit **Paper:** bit.ly/lstms-exploit

ACL 2018 Workshop on Representation Learning for NLP