LSTMs vs hierarchical structure in language

Marco Baroni

Facebook AI Research
In collaboration with

German Kruszewski  
**FAIR**

Yair Lakretz  
**NeuroSpin**

Dieuwke Hupkes  
**ILLC**
Motivation

• RNNs, LSTMs in particular, are amazing

• Train them on (variants of) language modeling, get improvements in tasks that required big pipeline of knowledge-based NLP tools

• What does this tell us, as theoretical linguists?
  • Nature vs nurture debate, etc.

• Questions I am currently interested in:
  • Are LSTMs really learning something about deeper, latent aspects of grammar, such as hierarchical structure?
  • If they do, how?
  • How does LSTM language processing compare to human language processing? [not discussed here]
Outline

• LSTMs are sensitive to hierarchical structure in language
  • Study 1: reading tea leaves in Principle Component space
  • Study 2: predicting syntactic tree depth with LSTMs

• How do LSTMs track long-distance linguistic information?
  • Study 3: meet the grandmother cells of long-distance number agreement
LSTM recap

- $C_t = f_tC_{t-1} + i_t\hat{C}_t$
- $h_t = o_t\tanh(C_t)$
- $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$
- $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
- $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
- $\hat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$

from Zheng et al, 2017
Study 1: LSTM sensitivity to constituent structure
Method

• French LSTM language model:
  • trained on 100M words from frWaC
  • 2 layers, 500 units per layer, dropout 0.5
  • Vocabulary size: 100k words

• Trained model applied to corpus of artificially constructed sentences with controlled syntactic structures

• Extract all word-by-word hidden vectors, compute their PCA

• Visualize trajectories of hidden vectors of same-structure sentences in first two PCs
The syntactic structures corpus

• 10k artificially generated French sentences

• Controlling for size of NP and VP, PoS of words at NP/VP boundaries (adjective, noun, ..., auxiliary, verb, ...), tense

• E.g.:
  • 2/6; N/Aux; P [cette cousin] [a consulté une lettre de boulanger]
  • 4/4; N/Aux; F [les noix du jardin] [vont nourrir ces frères]
  • 6/2; N/Aux; P [ces citrons du marché du patelin] [ont pourri]

• 66 different constructions, with about 160 instances of each
An example construction

La soeur timide va réparer un volet du libraire
The shy sister will repair a shutter of the bookseller

... ... ...

Un cousin nerveux va manger ces fruits du commerce
Des frères malades vont manger des pruneaux du verger
Ces cousines nerveuses ont bricolé les chaises du boulanger
DET ADJ NOUN AUX VERB DET NOUN PREP NOUN

NP VP NP
Hidden layer

The shy sister will repair a shutter of the bookseller
La soeur timide va réparer un volet du libraire

PC1 vs PC2
word embeddings
Word embeddings
Long NP, short VP
Negation
Study 2: Looking for the footprints of a tree
Looking for the footprints of a tree

Nelson et al. (2017)
Looking for the footprints of a tree

• Parsed BNC corpus with Stanford PCFG parser
• Extracted 18k sentences where depth was minimally correlated with position
• Pre-trained Gulordava language model:
  • trained on 90M words from Wikipedia
  • 2 layers, 650 units per layer, dropout 0.2
  • vocabulary size: 50k words
• (Lasso) regression to predict depth from hidden representations
## Results:

<table>
<thead>
<tr>
<th>Input Representation</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>word embeddings</td>
<td>0.27</td>
</tr>
<tr>
<td>bag of embeddings</td>
<td>0.45</td>
</tr>
<tr>
<td>$h_t$</td>
<td>0.66</td>
</tr>
<tr>
<td>$C_t$</td>
<td>0.70</td>
</tr>
</tbody>
</table>
You're talking about all sorts of geography that I didn't study at school.
Study 3:
How do LSTMs store long-term syntactic information?
The long-distance agreement task

• Linzen et al. (2016), Gulordava et al. (2018)

• 3,156 corpus-extracted examples in our version

• How do LSTMs carry morphosyntactactic features across the embedded constituents?
LSTM re-recap

- $C_t = f_t C_{t-1} + i_t \hat{C}_t$
- $h_t = o_t \tanh(C_t)$

- $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$
- $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
- $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
- $\hat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$
How might LSTMs carry morphosyntactic information through time?

\[ C_t = f_t C_{t-1} + i_t \hat{C}_t \]
Searching carry-through cells by ablation

• Look for units that impact performance on the Linzen data-set

single top layer cells

distributed patterns
<table>
<thead>
<tr>
<th>ablated unit(s)</th>
<th>accuracy</th>
<th>impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>94%</td>
<td>0%</td>
</tr>
<tr>
<td>775</td>
<td>85%</td>
<td>10%</td>
</tr>
<tr>
<td>769</td>
<td>90%</td>
<td>4%</td>
</tr>
<tr>
<td>775+769</td>
<td>66%</td>
<td>30%</td>
</tr>
<tr>
<td>any other unit</td>
<td>93%</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>majority baseline</td>
<td>63%</td>
<td>33%</td>
</tr>
</tbody>
</table>
Failure examples

Activation of unit 775 for incorrectly predicted target

- $h_t$
- $C_t$
- Prediction threshold
- Bias threshold

Words: often, the, only, players, on, either, side, of, the, ball, that, know, the, play, is, coming, are
Failure examples

Activation of unit 775 for incorrectly predicted target

- $h_t$
- $C_t$
- Prediction threshold
- Bias threshold
Tracking multiple dependencies: The case of non-subject relatives

• Subject relative:
  • The boy [that likes movies] is

• Non-subject relative:
  • The boy [that girls like] is

• 100 subject relatives, 14 non-subject relatives extracted from Linzen dataset with high-precision, low recall heuristics
Accuracy by sentence type

<table>
<thead>
<tr>
<th>sentence type</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>94%</td>
</tr>
<tr>
<td>subject relatives</td>
<td>96%</td>
</tr>
<tr>
<td>non-subject relatives</td>
<td>71%</td>
</tr>
<tr>
<td>majority baseline</td>
<td>63%</td>
</tr>
</tbody>
</table>
Generating controlled relative clauses
500 examples per structure

• The father that fears the girl definitely admires
• The father that fears the girls definitely admires
• The fathers that fear the girl definitely admire
• The fathers that fear the girls definitely admire

• The father that the girl fears definitely admires
• The father that the girls fear definitely admires
• The fathers that the girl fears definitely admire
• The fathers that the girls fear definitely admire
### Subject Relatives

The **father(s)** that **fear(s)** the **girl(s)** definitely **admire(s)**

<table>
<thead>
<tr>
<th>main noun</th>
<th>singular</th>
<th>plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>singular</td>
<td>ACC</td>
<td>ACC</td>
</tr>
<tr>
<td>plural</td>
<td>ACC</td>
<td>ACC</td>
</tr>
</tbody>
</table>

### Object Relatives

The **father(s)** that the **girl(s) fear(s)** definitely **admire(s)**

<table>
<thead>
<tr>
<th>main noun</th>
<th>singular</th>
<th>plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>singular</td>
<td>ACC</td>
<td>ACC</td>
</tr>
<tr>
<td>plural</td>
<td>ACC</td>
<td>ACC</td>
</tr>
</tbody>
</table>
subject relatives

The **father(s)** that **fear(s)** the **girl(s)** definitely **admire(s)**

<table>
<thead>
<tr>
<th></th>
<th>relative noun</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>main noun</td>
<td>singular</td>
<td>plural</td>
</tr>
<tr>
<td>singular</td>
<td>80%</td>
<td>87%</td>
<td>60%</td>
</tr>
<tr>
<td>plural</td>
<td>93%</td>
<td>100%</td>
<td>99%</td>
</tr>
</tbody>
</table>

object relatives

The **father(s)** that the **girl(s)** **fear(s)** definitely **admire(s)**

<table>
<thead>
<tr>
<th></th>
<th>relative noun</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>main noun</td>
<td>singular</td>
<td>plural</td>
</tr>
<tr>
<td>singular</td>
<td>46%</td>
<td>82%</td>
<td>27%</td>
</tr>
<tr>
<td>plural</td>
<td>89%</td>
<td>50%</td>
<td>92%</td>
</tr>
</tbody>
</table>
Inconclusions

• Language-modeling-trained LSTMs discover quite a bit about hierarchical structures in language

• Long-distance feature percolation is performed in a localist fashion by a small number of grandmother cells
  • How about structure tracking?

• Plenty of unsolved puzzles
  • Most interesting to me: does the LSTM keep track of multiple agreement patterns? if so, how?

• English sucks for this sort of studies