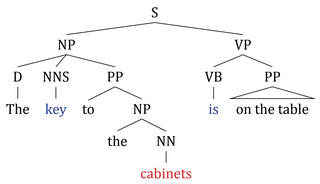
Agreement attraction errors in neural networks

Tal Linzen[1], Yoav Goldberg[2] and Emmanuel Dupoux[1]

[1]IJN & LSCP, CNRS, EHESS & ENS, PSL Research University [2]Bar Ilan University

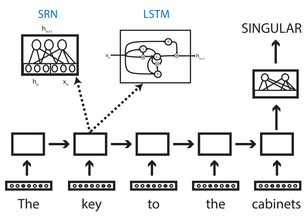
Introduction

- Recurrent neural networks (RNNs) have been shown to be effective in natural language processing tasks even though they are sequence models **without explicit structural representations**
- We use subject-verb agreement prediction to assess implicit structural learning in a sequential model (Elman, 1991)
- Our interest is in **learning from a natural corpus** rather than in the theoretical capabilities of RNNs
- Does the RNN make agreement errors?
- If so, are the errors similar to those that humans make?
- We focus on attractors that intervene in the linear order of words between the head of the subject and the verb



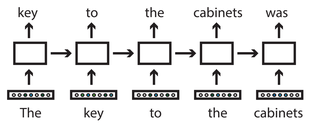
Models and training

Number prediction



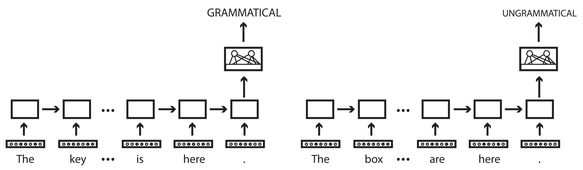
Word prediction

(language modeling)

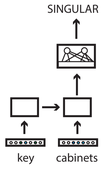


Evaluation: P(*is*) > P(*are*)?

Acceptability judgments

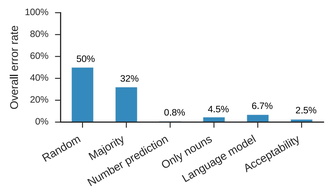


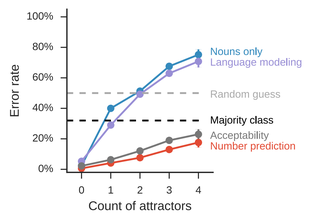
Nouns only

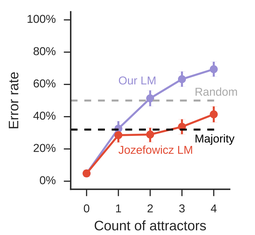


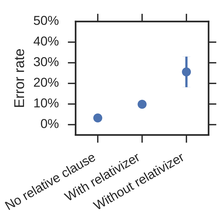
- 50 hidden units, 50-dimensional word representations
- Third-person present-tense subject-verb dependencies from the English Wikipedia: 121K in training, 1.21M in test
- Word prediction model trained with 20 different random initializations
- LM compared to LSTM with two 8192-unit layers (Jozefowicz et al., 2016)

Corpus error analysis





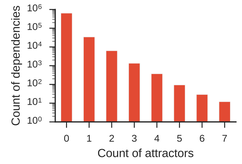


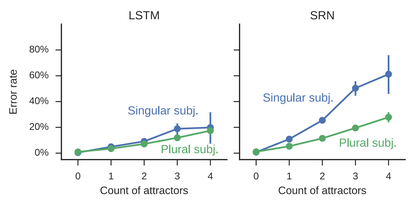


NO RELATIVE CLAUSE: The landmarks from this article...

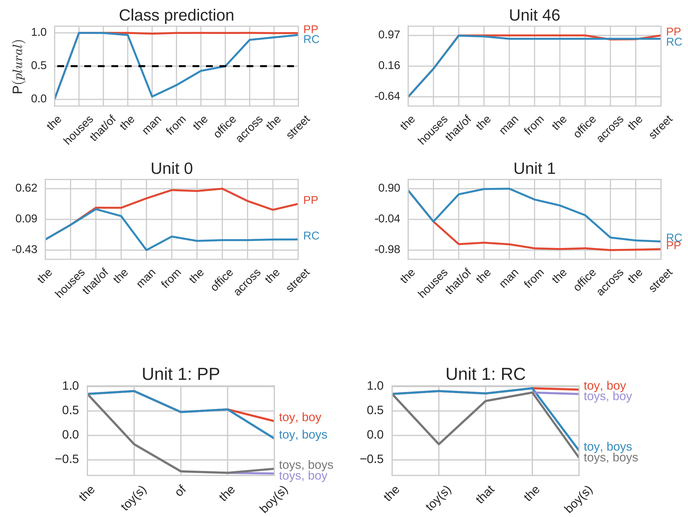
RC WITH RELATIVIZER: The landmarks that this article lists...

RC WITHOUT RELATIVIZER: The landmarks this article lists...





Leaky representation of structure



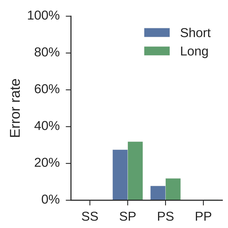
Comparison to human attraction errors

Bock and Miller (1991): Experiment 1

SHORT: The slogan(s) on the poster(s)...

LONG: The slogan(s) on the candidate's campaign poster(s)...

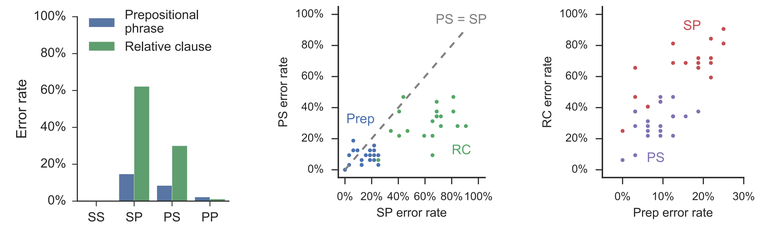
✓ *Match to human results*



Bock and Cutting (1992): Experiment 1

PREP: The demo tape(s) from the popular rock singer(s)...

RC: The demo tape(s) that promoted the popular rock singer(s)...



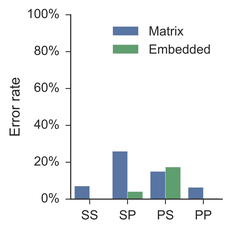
× *Humans made slightly **less** errors with RC than PREP modifiers*

Wagers, Lau and Phillips (2009): Experiment 2

EMBEDDED: The player(s) who the coach(es)...

MATRIX: The player(s) who the coach(es) like(s) the best...

✓ *Consistent with human self-paced reading results*



Conclusions

- We used an agreement prediction task to assess the emergence of structure in an RNN
- RNN / LSTM inductive biases are insufficient to develop structural representations from the word prediction signal alone
- With direct supervision, RNNs can learn an approximation of syntax that fails in difficult cases: they struggle with relative clauses (unlike humans)
- The singular/plural attraction error asymmetry occasionally emerges without an explicit assumption that plurals are marked