Psycholinguistics of Neural Language Models Summer 2023

Distinct patterns of syntactic agreement errors in recurrent networks and humans

Akshat Gupta, Tejaswi Choppa

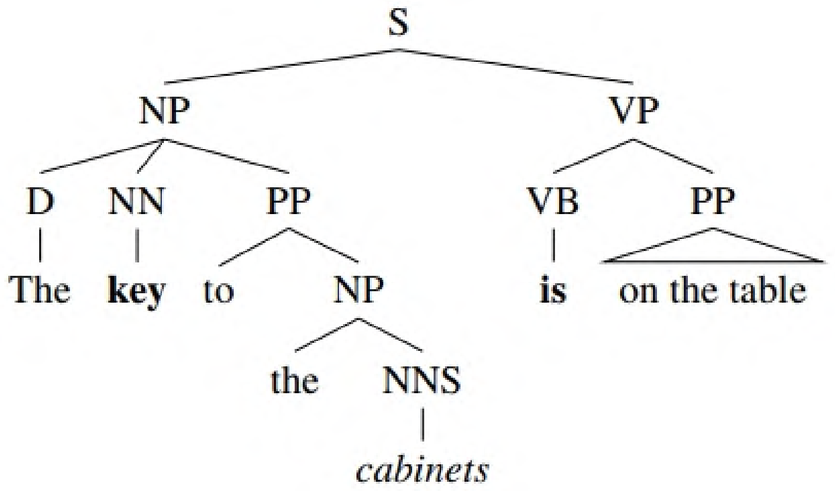
Idea

- The verb prediction task relies on the syntactic structure of the sentence
- But it is observed that RNNs perform comparably to humans even though they work sequentially
- The goal of this paper is to carry out a detailed comparison between the agreement errors made by humans and RNNs

Introduction

- Syntactic Dependencies of words could be expressed structurally
- The form of a verb often depends on whether the subject is singular or plural
- Many times the noun preceding the verb might not be a subject
- These explicit structural representations comprise the human syntactic knowledge but RNNs do not have this type of representation

"The *key* to the cabinets *is* on the table"



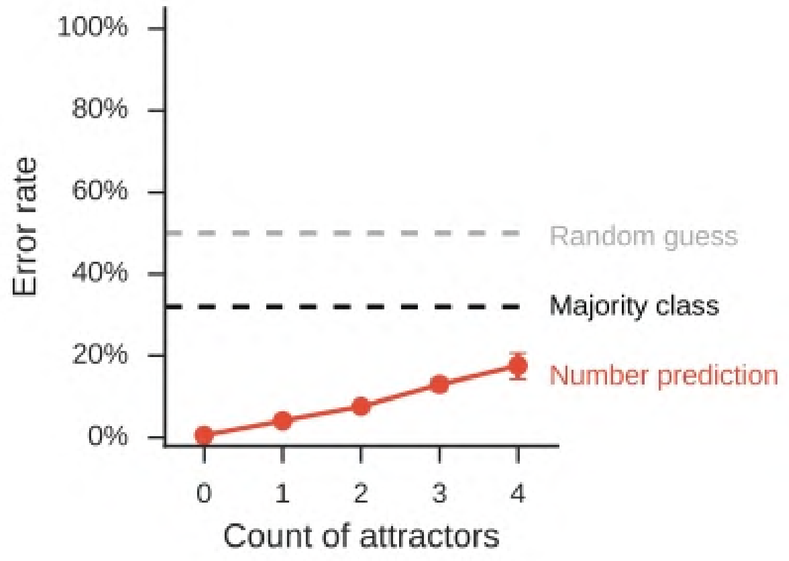
Introduction

- A study by Goldberg et al in 2016, trained RNN to predict whether the verb should be singular or plural based on words leading to the verb
- All preceding words till the verb called the **preamble**
- RNNs were trained on a large sample of sentences from Wikipedia

'Yet the *ratio* of men who survive to the women and children who survive *is*' [not clear in this story.]

Introduction

- RNN error rates on verb number prediction with the increasing number of attractors
- Still, RNNs performance degrades on sentences with attractors suggesting that syntactic processing in RNN is imperfect
- The majority class baseline always predict singular verb with 68% accuracy



Introduction

- They look into two factors:
  - The type of syntactic structure that contains the attractor
  - The structural position of the attractor

- For this, they look at two types of sentences:
  1. Prepositional phrase (PP):

     eg: The demo **tape** from the popular rock *singers*...

  2. Relative clause (RC):

     eg: The demo **tape** that promoted the rock *singers*...

Experiment 1

- The participants had to produce the sentence and decide the upcoming verb as 'is' or 'are' based on the preamble they heard

- The task was conducted under three settings:
  - **RSVP** → Rapid Serial Visual Presentation
  - **SPR** → Self-Paced Reading
  - Untimed Paradigm

Experiment 1: RSVP

- Words are displayed one by one in the center of the screen
- Each word was presented for 250ms, followed by a blank screen for 150 ms
- The participants were given 1500ms to choose between the verbs **is** and **are**

Experiment 1: SPR

- The sentences were revealed word by word
- Participants controlled the rate at which words were revealed
- As the word was revealed, the previous words were replaced by a string of dashes
- Participants were given 1500 ms to choose between **is** and **are** after the end of the sentence

Experiment 1: Untimed

- The full sentence and two response options were revealed at the same time
- Participants were given as much time as they wish to make the choice

- A total of 88 preambles with 32 critical items and 56 fillers were given under 8 conditions to each participant
- There were around 384 participants (128 per experiment)

Modifier type	Subject number	Local noun match	Preamble
PP	Singular	Match	The demo tape from the popular rock singer
PP	Singular	Mismatch	The demo tape from the popular rock singers
PP	Plural	Match	The demo tapes from the popular rock singers
PP	Plural	Mismatch	The demo tapes from the popular rock singer
RC	Singular	Match	The demo tape that promoted the rock singer
RC	Singular	Mismatch	The demo tape that promoted the rock singers
RC	Plural	Match	The demo tapes that promoted the rock singers
RC	Plural	Mismatch	The demo tapes that promoted the rock singer

Table 1: Materials of Experiment 1 (Bock & Cutting, 1992).

Experiment 1: Human findings

- The qualitative patterns from human experiments that we would like to compare to RNNs are:

  - **Attraction:** Errors are more likely in the presence of an attractor
  - **Number asymmetry:** Errors are more likely when the subject is singular and the attractor is plural than the other way around
  - **Relative clause advantage:** Attractor in prepositional phrases leads to more mistakes than in relative clauses.

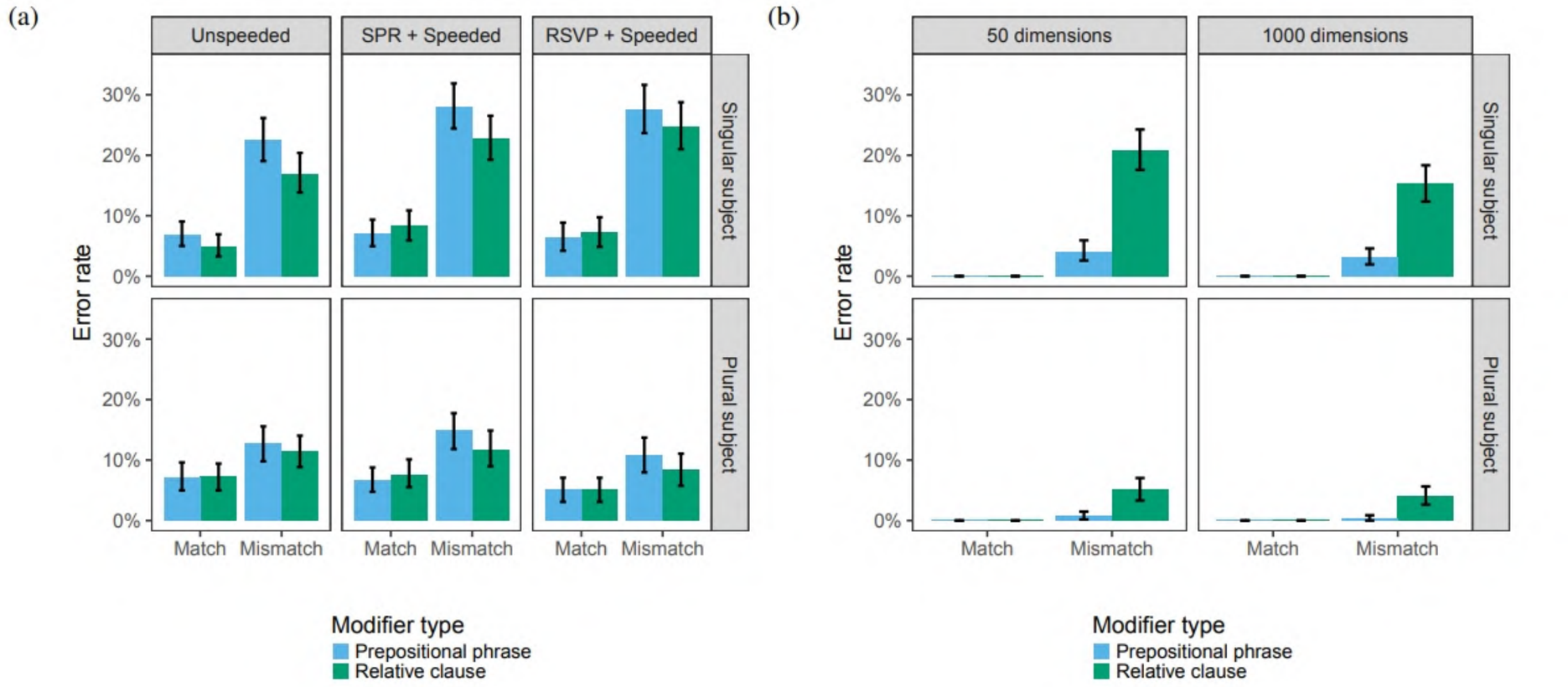
Experiment 1: Dataset

- The RNN was trained in a supervised way to find the number of the upcoming verb
- The model was trained on a corpus of English sentences extracted from the Wikipedia
- Sentences in which there were no subject-verb agreements were excluded
- One verb in a sentence was randomly selected and by deleting the portion after it preambles were created
- Around 1.27 million preambles as training and 142k as validation test set was used

Experiment 1: Model Training

- Each word of the preamble was encoded as word embedding
- These word embeddings were fed into RNN with a single layer of LSTM units
- The model did not have access to the characters that made up the word
- The number of the verb was predicted from the final state of the RNN
- Smaller models were trained, in which the recurrent layer had 50 units
- Larger models had 1000 units
- Word representations were 50-dimensional in both cases

Experiment 1: Results



(a) Human agreement errors in experiment 1

(b) RNN prediction errors for the same

Experiment 1: Results

- Five of the 32 items showed error rates exceeding 20% across conditions, including the Match conditions
- The unusual number of errors in these preambles was due to the presence of low frequency words
- The model did not make any errors in preambles in which the subject and the local noun had same number
- Two aspects of the networks' error patterns are consistent with the human data:
  - Agreement errors were more common when the local noun did not match the number of the subject
  - Errors in attraction occurred more often when the subject was singular and the noun was plural
- Overall performance of models was good with 30%error rate in 50 dimensions and 24% in 1000 dimensions

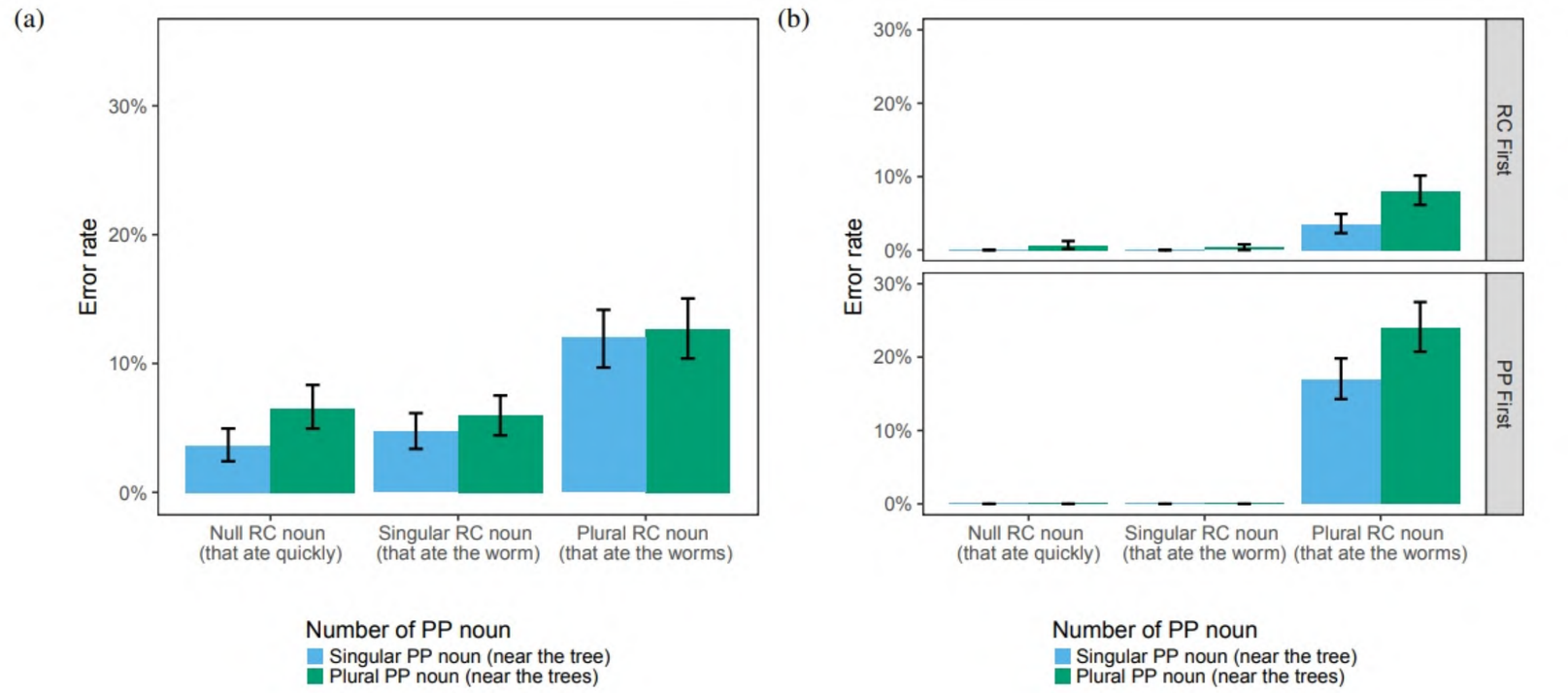
Experiment 2: Dataset

- RNNS makes more errors when there are two attractions
- Previous work by Franck et. al shows that a word closer to the subject was a more effective attractor than a word closer to the verb
- A relative Clause (RC) is used with a Prepositional phrase(PP) embedded in it
- Eg. The bird that ate the worm(s) near the tree(s)
- Because of this, we can vary the number of intervening nouns by replacing first nouns with adverbs
- Phrase(the worm) with an adverb(quickly)
- 36 critical items in six conditions and 76 filler items were constructed

Experiment 2: Model Training

- Model architecture is the same as in experiment 1
- Agreement errors were overall less frequent than in experiment 1.
- Even with plural attractors, the average error rate was 12.7% compared to 22.6% in singular mismatch RC
- One reason could be the word used in Experiment 2 was more frequent than in Experiment 1
- The presence of a plural attractor at the beginning of RC is the main cause of the error
- The number of second nouns did not affect the error rate
- There were very few attraction errors when there was exactly one attractor and it was inside PP
- This confirms that RNNs were able to ignore nouns embedded in a PP
- Errors increased when the noun directly following the beginning of RC was an attractor

Experiment 2: Results



(a) Human agreement errors in experiment 2

(b) RNN prediction errors for the same material

Experiment 2: Results

- There was a cumulative attraction effect with a higher error rate when both nouns were plural
- To test reliance on short RC heuristic, the expectation is that this reversal should increase the interference from the relative clause heuristic
- The reversed sentences had more than double the error rate of the original ones
- And with Humans, the attractor inside PP was successfully ignored, even if it was close to the subject

Short RC heuristic

Short: The lion that the tigers (ate)
Medium: The lion that the hungry tigers (ate)
Long: The lion that the extremely hungry tigers (ate)

- The hypothesis is that RNNs were relying on the heuristic that RCs tend to be short.
- Errors in these sentences are consistent with the RC length heuristic
- In the RC-external prediction point, interference from the attractor decreases as the RC becomes longer
- In the RC-internal condition, the error rate increases due to the revert of RNN to the main subject to make its prediction

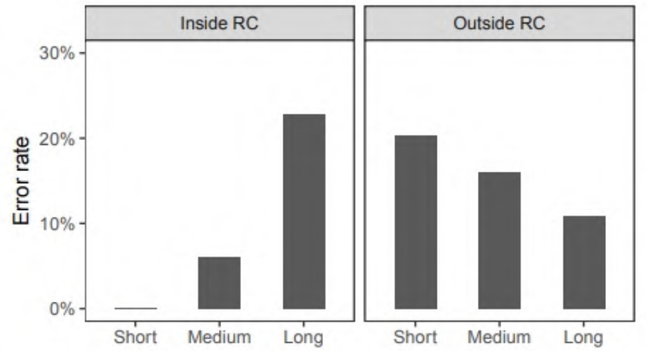


Figure 4: Effect of relative clause length on error rate in predicting the number of the embedded verb (inside RC) and main verb (outside RC).

Conclusion

- RNN tend to make errors in sentences with relative clauses(RC) whereas,
- Human shows a small difference on RC
- Errors in RNN were effected by the proximity of the attractor to the verb rather than to the subject as in humans
- The syntactic representation used by RNNs differs from humans
- Attraction errors were more likely when the subject was singular and the attractor plural than the other way around

Our Project Proposal

- We propose to extend this experiment for a variety of architectures in increasing order of complexity of models
- We want to measure how the understanding of syntactic structure progresses with complexity
- We want to test LSTM, LSTM with attention, and Transformer and compare the performance
- Find out possible reasons for the change in performance of different models
- Compare results with gold data to see how closely the model is able to understand the syntactic structure

Our Project Proposal

- Dataset
  - Dataset of wiki Vocab: https://github.com/TalLinzen/rnn_agreement
  - Human Data:https://github.com/jhupsycholing/RNNvsHumanSyntax
- Tools
  - Python will be our primary coding language
  - TensorFlow for deep neural networks, and other supported libraries for preprocessing
  - VScode for coding, and GitHub for code management
- We will report all experiments using wandb or mlflow so that all experiments remain transparent

Our Project Proposal

- **Research Question:** Effect of Complexity of Model on a syntactic understanding of sentences in neural networks
- **Results:** Concrete results in terms of error rate, accuracy, precision, recall, and f1-score
- Compare different architectures on different evaluation criteria which will establish the result in numbers about how good or bad the model understands sentences for syntax

References

- Distinct patterns of syntactic agreement errors in recurrent networks and humansTal Linzen, Brian Leonard
- Syntactic Structure from Deep LearningTal Linzen, Marco Baroni
- Dataset of wiki Vocab: https://github.com/TalLinzen/rnn_agreement
- Human Data:https://github.com/jhupsycholing/RNNvsHumanSyntax

Thank you

Questions?