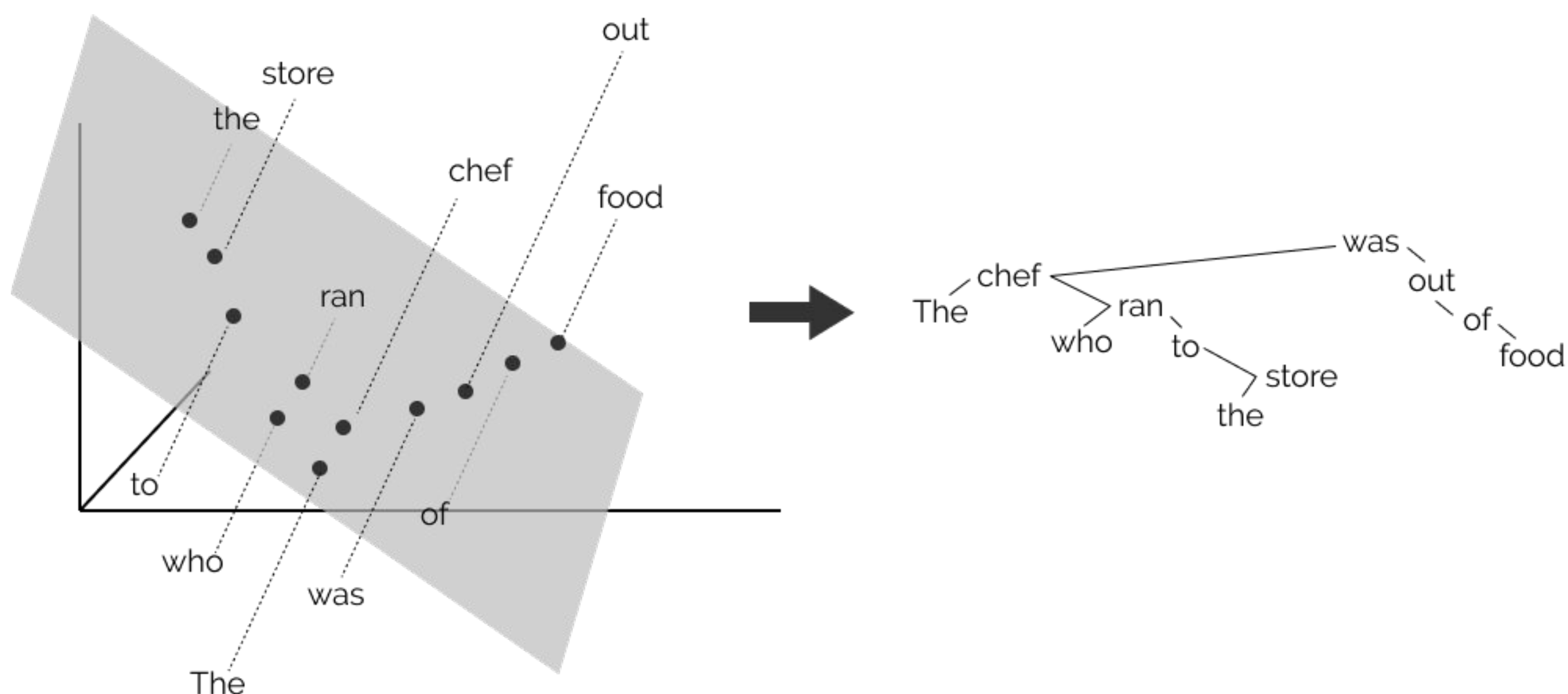


Recurrence vs Attention: Probing Linguistic Properties

Aman Hussain and Omar Elbaghdadi
University of Amsterdam

Introduction

- Attention-based models currently dominate recurrent models, but why?
- **Research question:** What are the differences in the representations learnt by attention-based models and recurrence-based models?
- Probes: help explore linguistic phenomena encoded in hidden representations
- We study syntax: POS Tags and parse tree embeddings



Methodology

- Attention model: DistilGPT-2
- Recurrent model: LSTM
- Probes *predict* a feature from representations. More predictive representations better
- *Control tasks:* control for probe learning task from supervision
- New success metric *selectivity:* difference between task accuracy and control accuracy.
- Use *structural probes* to check for existence of whole syntax trees in representations.

Results

- LSTM and DistilGPT-2 good at POS tag classification
- LSTM has **higher selectivity** at POS tag classification than DistilGPT-2
- Both models reconstruct parse tree quite well, but LSTM slightly better.

POS Tag Classification Results

	Recurrence	Attention
Test Accuracy	89.17%	90.58%
Control Accuracy	62.09%	70.08%
Selectivity	27.08%	20.50%

Discussion

- Higher selectivity: LSTM learns to encode POS tag information better
- Transformer doesn't learn more informative representation for this task
- Structural probing experiment does not have a control task
- LSTM are trained with much less data than Transformers

Structural Probe Results

	Recurrence	Attention
Distance UUAS	67.39%	65.44%

