Probing Autoregressive Language Models

Michael Neely & Vanessa Botha

Natural Language Processing 2, University of Amsterdam

University of Amsterdam

Problem Statement

We investigate popular **autoregressive language models** for clues on how they process linguistic phenomena using **diagnostic classifiers** (Hupkes, Veldhoen, and W. Zuidema 2017).

Research Questions:

1. What **linguistic** and **structural** properties are encoded in the representations of autoregressive language models?
2. Are these properties localized to certain hidden layers?
3. Can we find more expressive probes using control tasks (Hewitt and Liang 2019)?

Selectivity

- Popular probes are over-parameterized and can **memorize linguistic tasks**.
- We can check if a model is over-parameterized by using a **control task**, which associates input types with random outputs.
- We can only derive valid conclusions from **selective** models: ones that achieve high accuracy on a given task and low accuracy on the control task.

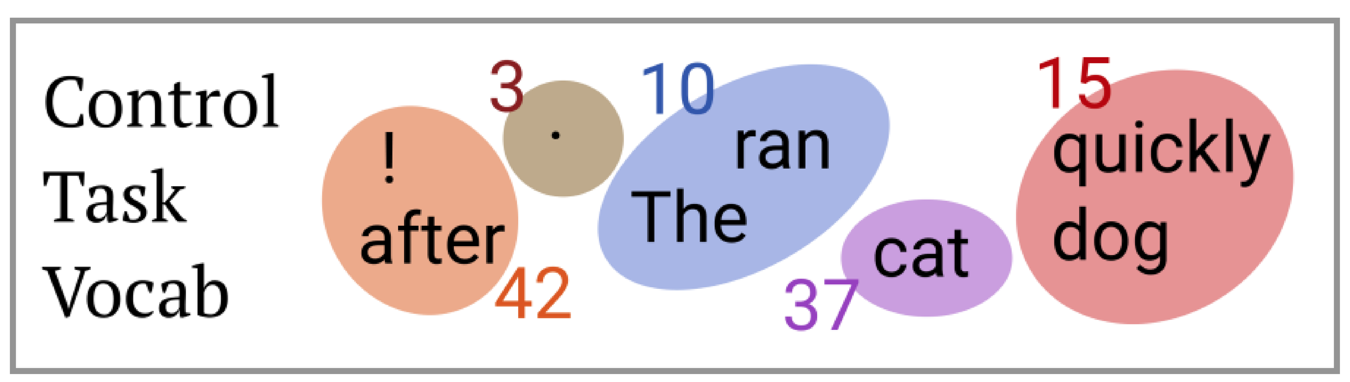
Selectivity = Performance on Task - Performance on Control Task

A good selectivity score is greater than a 20% different (Hewitt and Liang 2019).

Probing Linguistic Properties

POS-Tagging Task: We train a simple **Linear Classifier** f that maps a model's representation h to a corresponding POS tag t: $f(h) \to t$.

POS-Tagging Control Task: We define a random behavior for each word type in the vocabulary.

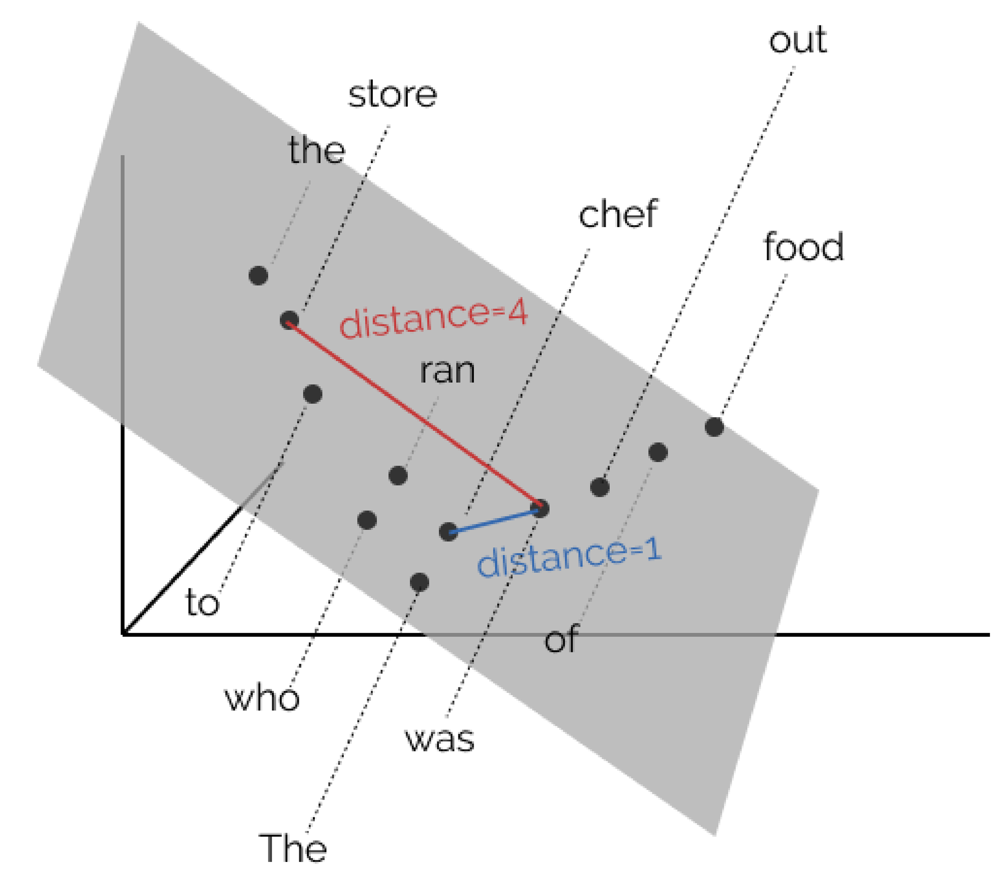


Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	**10**	**37**	**10**	**15**	**3**
Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	**10**	**15**	**10**	**42**	**42**

Probing Structural Properties

Tree Distance Task: We train a simple linear model to recreate the tree distance between all pairs of words (w_i, w_j) in a sentence by learning the matrix $\mathbf{B}$ using the hidden state $\mathbf{h}$ such that the symmetric, positive semi-definite edge matrix $\mathbf{A} = \mathbf{B}^\top\mathbf{B} = (\mathbf{Bh})^\top(\mathbf{Bh})$.

Tree Distance Control Task: We define a random distance between every word pair (w_i, w_j) in the vocabulary.



Experiments

Dataset: Universal Dependencies English Web Treebank (Silveira et al. 2014).

Models:

- The LSTM of Gulordava et al. (2018) (**recurrent**)
- Distilled GPT2 (Sanh et al. 2019) (**attention-based**)
- XLNet (Yang et al. 2019) (**attention-based**)

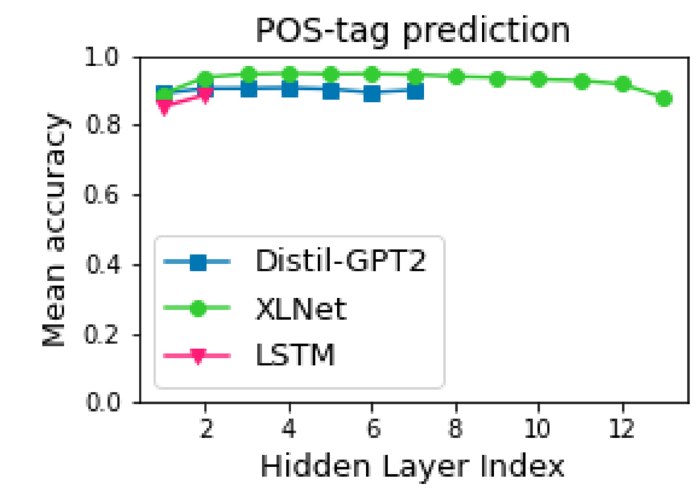
Evaluation Metrics:

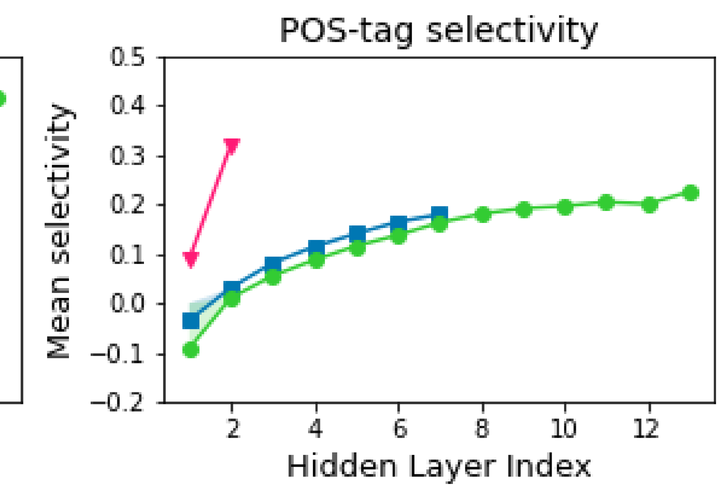
- **POS-Tagging Task**: Token Accuracy
- **Tree Distance Task**: Undirected Unlabeled Attachment Score (UUAS)

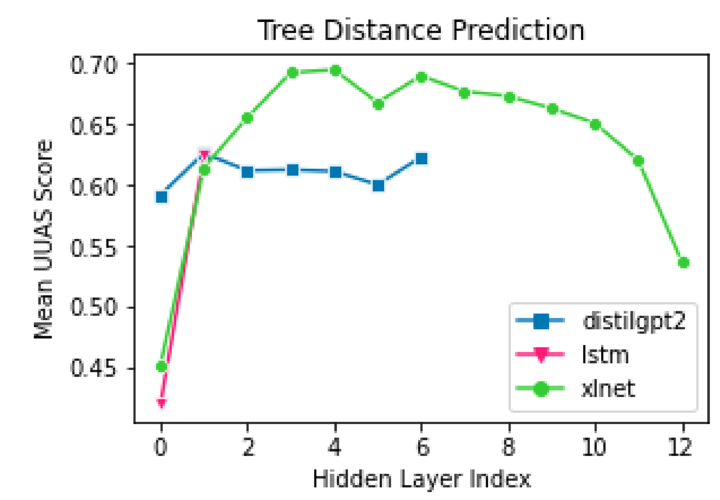
Results

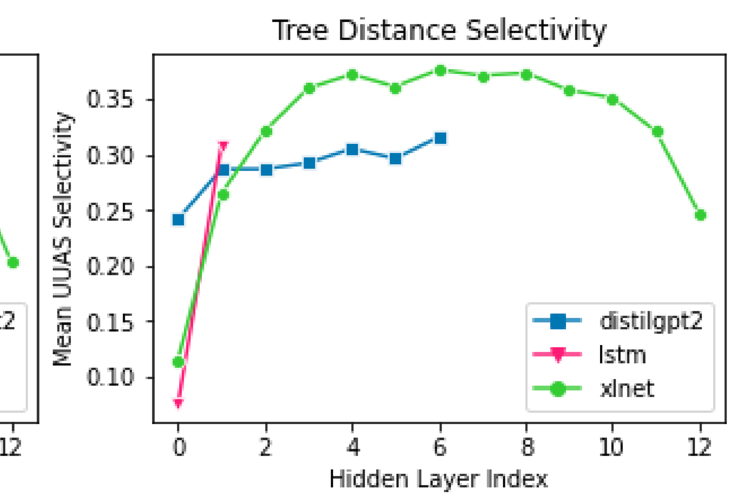
Task	Language Model	Avg. Score*	Avg. Selectivity*
POS-Tagging	Gulordava LSTM	0.8871	0.3201
	Distilled GPT2	0.9016	0.1797
	XLNet	0.8803	0.2254
Tree Distance	Gulordava LSTM	0.6246	0.307
	Distilled GPT2	0.6229	0.3154
	XLNet	0.5373	0.2462

*Results reported from last hidden layer.









Discussion

- Linguistic information about POS-tag information encoded within the Transformers cannot easily be extracted by a linear transformation. Would more complex probes still remain selective?
- The quality of structural information varies between layers, while the POS-tag information is clearly embedded in each layer.
- No selective probes are found when correcting for the performance ceiling on the structural control task.

Future Work: Can tree distance structural probes recover parse trees from sequence-to-sequence models trained on recursive artificial languages like those of Veldhoen, Hupkes, and W. H. Zuidema (2016) and Hupkes, Dankers, et al. (2019)?

References

Gulordava, Kristina et al. (2018). *Colorless green recurrent networks dream hierarchically.* arXiv: 1803.11138 [cs.CL].

Hewitt, John and Percy Liang (2019). *Designing and Interpreting Probes with Control Tasks.* arXiv: 1909.03368 [cs.CL].

Hupkes, Dieuwke, Verna Dankers, et al. (2019). *Compositionality decomposed: how do neural networks generalise?* arXiv: 1908.08351 [cs.CL].

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema (2017). *Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure.* arXiv: 1711.10203 [cs.CL].

Sanh, Victor et al. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* arXiv: 1910.01108 [cs.CL].

Silveira, Natalia et al. (May 2014). "A Gold Standard Dependency Corpus for English". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland: European Language Resources Association (ELRA).

Veldhoen, Sara, Dieuwke Hupkes, and Willem H. Zuidema (2016). "Diagnostic Classifiers Revealing how Neural Networks Process Hierarchical Structure". In: *CoCo@NIPS.*

Yang, Zhilin et al. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding.* arXiv: 1906.08237 [cs.CL].