Uni-Directional Context is Insufficient for Hierarchical Language Understanding

Michael Neely
University of Amsterdam
Student number: 12547190
michael.neely@student.uva.nl

Vanessa Botha
University of Amsterdam
Student number: 10754954
vanessa.botha@student.uva.nl

Abstract

We investigate the extent to which popular recurrent and attention-based neural models trained with an auto-regressive language modeling objective can represent the hierarchical nature of language by probing linguistic and structural properties using diagnostic classifiers. We tune our classifiers to be selective, for which we introduce a novel tree distance control task and the generalized selectivity metric. Our results suggest that access to bidirectional context improves the quality of extracted parse tree geometry.

1 Introduction

The success of deep learning models on zero-shot tasks when pretrained with a language modeling objective implies that these models understand language (Radford et al., 2019). However, investigating the encoded representations of a model provides a more reliable signal of the presence of linguistic phenomena than performance on downstream Natural Language Processing (NLP) tasks.

In probing tasks, supervised *diagnostic classifiers*[1] (Hupkes et al., 2018) trained on top of these representations can yield constructive evidence for the existence of a particular linguistic property, at least in a manner extractable by the classifier. Valid conclusions are limited to *selective* diagnostic classifiers: those that achieve laudable task scores even when accounting for memorization capacity as measured by a *control task* (Hewitt and Liang, 2019).

In this paper, we train selective diagnostic classifiers to examine whether recurrent models and attention-based transformers models encode linguistic and structural properties differently. Recurrent models have a chain-like nature and process sentences word by word while maintaining a hidden state that summarizes past inputs. In contrast, transformer architectures process all words in the sentence at once, using self-attention to learn dependencies between words (Vaswani et al., 2017). We introduce a novel control task for our structural probe and propose an alternative selectivity metric, which we call *generalized selectivity*, which is a more reliable indicator of a probe's ability to extract meaningful information from encoded representations. Our experiments show that the choice of language modeling objective, not neural architecture, is the most significant determining factor of probe selectivity and performance.

[1] Also referred to as *probes* by (Hewitt and Manning, 2019). We use the terms interchangeably.

2 Related Work

Hupkes et al. (2018) formally introduce diagnostic classification as a method to test a network's internal representations of input features and its symbolic processing strategies. More recent work probes contextual token embeddings produced by pretrained encoders for linguistic and structural information. At the sub-sentential level, contextual embeddings can capture part-of-speech (Belinkov et al., 2017b), sentence length (Adi et al., 2016), and morphological (Belinkov et al., 2017a) information. Tenney et al. (2019) propose a series of edge probing tasks designed to detect evidence of syntactic, semantic, and varied distance phenomena. They find a large improvement in syntactic tasks compared to static word type vectors (Mikolov et al., 2013).

Seeking more conclusive proof, Hewitt and Manning (2019) propose a structural probe to test encoded representations for consistently embedded syntax trees. They find strong evidence of this behavior with a targeted case study of BERT (Devlin et al., 2018) and ELMo (Peters et al., 2018). Hewitt and Liang (2019) introduce control tasks in a follow-up paper, echoing the concerns between probe complexity and the validity of the inferred features first noted by Hupkes et al. (2018). They demonstrate the memorization capacity of popular probe architectures in POS-tagging and edge

dependency prediction tasks and propose the selectivity metric to provide added context to probe results.

3 Approach

We focus on three models trained with an autoregressive (AR) language modeling objective. Unlike auto-encoding (AE), AR language modeling captures long-range dependency between tokens and matches the sequential ordering of language.

3.1 Recurrent models

LSTM We use the LSTM model trained on the English Wikipedia by Gulordava et al. 2018 for a downstream long-distance number agreement task.

3.2 Attention-based models

Distilled GPT-2 This is the distilled variant (Sanh et al., 2019) of the large transformer GPT-2 (Radford et al., 2019) that was pre-trained on text from eight million websites. GPT-2 is unidirectional and encodes absolute word positions in the embeddings.

XLNet XLNet (Yang et al., 2019) uses permutation language modeling, — a generalization of AR — and is, therefore, able to capture bidirectional context. It re-uses hidden-states computed over previous sequences, while encoding the relative word positions, to capture longer-term dependency.

3.3 Probing linguistic properties

POS tag prediction We train a linear classifier, referred to as a linguistic probe, to predict Part-of-Speech (POS) tags from the language model's representations of individual words. Succesful classification may mean that the target language model encodes such token-level linguistic properties.

Control task We use the control task introduced by (Hewitt and Liang, 2019). Let $\mathcal{Y}$ be the set of all POS tags in a parsed corpus's training set and v a word type in the vocabulary V. For this task, the control behavior $C(v)$ defines a deterministic mapping from a token w_i with word type v to a POS tag $y_i \in \mathcal{Y}$, sampled independently for each v according to the empirical distribution of the POS tags in the training set.

3.4 Probing linguistic structure

Tree Distance prediction We test for the existence of embedded syntax trees using the tree distance structural probe of Hewitt and Manning (2019). Let $\mathcal{M}$ represent a language model with hidden dimensionality z that receives an input sequence of n words $\mathbf{w}^{\ell}_{1:n} \in \mathbb{R}^n$ and calculates an output sequence of encoded vector representations $\mathbf{h}^{\ell}_{1:n}$ at a particular hidden layer, where ℓ serves as a sentence identifier. There exists a linear projection $\mathbf{B} \in \mathbb{R}^{z \times k}$ of rank k such that

$$d_B(\mathbf{h}^{\ell}_i, \mathbf{h}^{\ell}_j)^2 = \sum_k (\mathbf{B}(\mathbf{h}^{\ell}_i - \mathbf{h}^{\ell}_j))^{\mathrm{T}}(\mathbf{B}(\mathbf{h}^{\ell}_i - \mathbf{h}^{\ell}_j))$$

represents the squared tree distance between the word pair (w_i, w_j) in sentence ℓ. The probe's parameters are the elements of B which are optimized through gradient descent to recreate the true tree distance between all pairs of words in all sentences $\mathbf{T}^{\ell}$ in the training set of a parsed corpus:

$$\min_B \sum_{\ell} \frac{1}{|s^{\ell}|^2} \sum_{i,j} |d_{T^{\ell}}(w^{\ell}_i, w^{\ell}_j) - d_B(\mathbf{h}^{\ell}_i, \mathbf{h}^{\ell}_j)^2|$$

Control task Let $\mathcal{Y}$ be the set of all tree distances in a parsed corpus's training set and v a word type in the vocabulary V. Then the control behavior $C(v_i, v_j)$ is defined as a deterministic mapping of an ordered pair of word types (v_i, v_j) to a tree distance $y_i \in \mathcal{Y}$, where $C(v_i, v_j) = 0$ if $i = j$, sampled independently for each (v_i, v_j) according to the empirical distribution of tree distances in the training set. Thus, the control tree distance matrix for any given sentence ℓ remains positive and symmetric and therefore yields a valid minimum spanning tree.

3.5 Generalized Selectivity

The performance ceiling of a control task, as defined by Hewitt and Liang (2019), is the fraction of words (or word pairs) that occur in the training set plus biased chance accuracy on all other tokens. Thus, the selectivity s of a probe depends heavily on the percentage of out-of-vocabulary (OOV) occurrences. If a probe's performance p approaches the control task's ceiling c_{control}, this indicates overparameterization. We define *generalized selectivity* as an accurate measure of a probe's selectivity after accounting for possible memorization instances:

$$\mathrm{s}_{\text{generalized}} = \mathrm{p}_{\text{task}} - \frac{\mathrm{p}_{\text{control}}}{\mathrm{c}_{\text{control}}} \quad (1)$$

4 Experiments and Results

4.1 Experiments

We use the Universal Dependencies English Web Treebank (Silveira et al., 2014) with the provided

train/test/validation splits. We extract the Universal POS tags for the gold labels of the linguistic task. For the tree distance task, we generate the true distances matrices $\mathbf{T}^{\ell}$ using the annotated token trees. For the classifier input data, we feed in the treebank sentences and extract hidden representations at each hidden layer of each model.

For the LSTM, we use the pretrained English model weights and vocabulary of Gulordava et al. (2018). For the transformer models, we use the pretrained `distilgpt2` and `xlnet-base-cased` weights and tokenizers from the Huggingface transformer library (Wolf et al., 2019). To account for the byte pair encoding used in transformer tokenizers, we calculate a word's representation by averaging over the representations of its subwords (Hewitt and Manning, 2019). The number of hidden layers and their representation dimensionality are specified in Appendix A Table 2. We include the embedding layers of each model, indexed as hidden layer 0. Due to a PyTorch limitation, we do not include the first hidden layer of the LSTM.

We tune the linguistic probe to maximize generalized selectivity by artificially constraining the number of training sentences since Hewitt and Manning (2019) report this form of complexity control to improve selectivity. Out of a total of 12543 sentences, we train on $\{10000, 1000, 100\}$, which roughly corresponds to the recommended 100%, 10%, 1% subsets (Zhang and Bowman, 2018). For the structural probes we vary the rank k of $B \in \mathbb{R}^{z \times k}$ in $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$. We train the linguistic and structural probes for a maximum of 300 and 25 epochs, respectively, using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 and a batch size of 64. The learning rate is reduced by half when the validation loss fails to decrease after a single epoch, and we use early stopping with a resetting patience value of 5 epochs. For reproducibility, we train our probes using three fixed seeds.

We measure the performance of the linguistic probe using POS tag accuracy. For the structural probe, we compute the undirected, unlabeled attachment score (UUAS) Hewitt and Manning (2019). We report generalized selectivity scores for both probes, but note that the performance ceiling of the linguistic probe is precisely 1.0, since all OOV words are mapped to a fixed, randomly sampled POS-tag. Independent-samples T-tests are used to test for significance ($\alpha = 0.05$).

4.2 Results

Table 1 details the task scores of linguistic and structural probes on the encoded representations extracted from the final hidden layer of each model. The decoder of a language model uses these representations to predict the probability distribution of the next token, so it is natural to expect the final hidden layer to contain the densest concentration of linguistic and structural information and, therefore, lead to high generalized selectivity scores.

Task	Language Model	Score	Generalized Selectivity
POS tag	LSTM	0.8871 ± 0.0000 *	0.2522 ± 0.0047
	DistilGPT2	0.9016 ± 0.0001 *	0.1831 ± 0.0081
	XLNet	0.8801 ± 0.0002 *	0.2261 ± 0.0039
Tree distance	LSTM	0.6253 ± 0.0007 †	0.2127 ± 0.0002
	DistilGPT2	0.6437 ± 0.0009 †	0.2358 ± 0.0041
	XLNet	0.5389 ± 0.0008 †	0.1385 ± 0.0096

Table 1: The accuracy(*) and UUAS (†) obtained on the last hidden layer of each model. For all metrics, we report the average score over three runs $\pm$ the standard deviation. All results were significant $p < .05$.

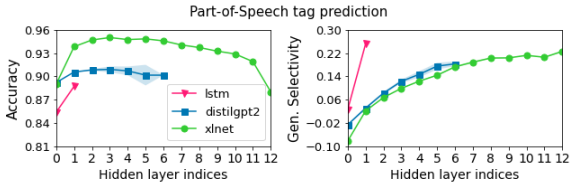

Figure 1: POS tag accuracy and generalized selectivity attained by linguistic probes across the language model hidden layers. Scores are averaged across three runs, with the shaded regions denoting the standard deviation.

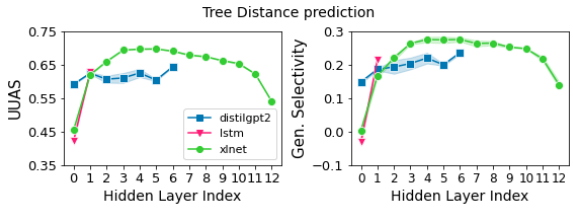

Figure 2: UUAS and generalized selectivity attained by structural probes across the language model hidden layers. Scores are averaged across three runs, with the shaded regions denoting the standard deviation.

While Figure 2 confirms this notion holds for the smaller LSTM and DistilGPT2 models, it also reveals that parse tree geometry is maximally discoverable in the intermediate layers of XLNet. Structural probes trained on the final hidden layer of XLNet achieve low performance and generalized selectivity similar to those trained on the earlier

hidden layers. In contrast, the generalized selectivity of the linguistic probe consistently increases over the hidden layers (see Figure 1). This result suggests that information about POS-tags is best encoded in the final hidden layer. A noticeable difference in syntactic information between layers can also be observed in Figure 2, which is consistent with the results of Hewitt and Manning (2019) and Tenney et al. (2019). Figures 7 and 8 in Appendix C show the important of tuning the structural probe, as increasing the rank k of the linear transformation B beyond 32 results in no additional UUAS performance gain and harms generalized selectivity.

Figure 3 shows that while performance deteriorates for all models as sentence length increases, the LSTM and DistilGPT2 display remarkably similar behavior. The difference between the scores reported by these two models is not significant $p = 0.8409$), unlike the differences between the LSTM and XLNet ($p = 0.0003$) and DistilGPT2 and XLNet ($p = .0004$).

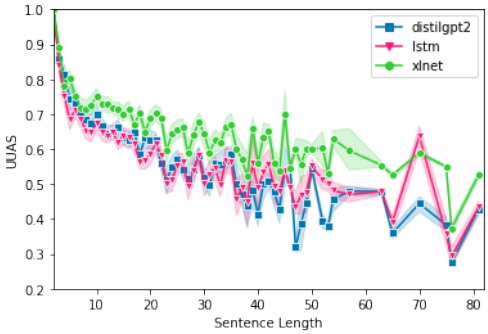

Figure 3: Average UUAS per sentence length for the most selective probes chosen by their generalized selectivity scores.

5 Discussion

In our study, we aimed to examine the difference between recurrent and attention-based architectures. We hypothesize that the choice of language modeling training objective, not the neural architecture, influences the localization and robustness of encoded syntax. Our maximum reported UUAS and POS-tag accuracy on the hidden layers of DistilGPT2 and the LSTM are substantially lower than those reported by Hewitt and Manning (2019) on ELMo and BERT and (Hewitt and Liang, 2019) on ELMo. Both of these models are pretrained with a classic AR objective, which encodes a unidirectional context. BERT's AE objective, while still problematic (see (Yang et al., 2019)), utilizes bidirectional context, and ELMO's constructed word embeddings take the entire sentence into account. Richer context is often required to resolve syntactic and lexical ambiguity introduced by phenomena such as the garden path effect and polysemy. Thus, the superior performance of the XLNet-base model in our experiments is likely due to its generalized bidirectional AR pretraining objective and segment recurrence mechanism and relative encoding scheme features. We suspect the XLNet-large model will achieve higher scores than its base version, but we leave that to future work.

Our probe's high generalized selectivity scores provide strong evidence for our conclusions. However, diagnostic classification only provides a limited form of constructive evidence. As John Hewitt points out, probes do not answer the question of *how* neural models achieve their behavior, only "does my model encode X phenomenon in a manner extractable by my probe?"[2]. Perhaps the strongest criticism of diagnostic classifiers comes from Saphra and Lopez (2019), who argue that probes may memorize the association between an embedding and its most frequently associated output category. Control tasks help address this argument by placing a diagnostic classifier's performance in the context of its potential memorization capacity. The generalized selectivity metric introduced in this paper extends this notion even further. However, diagnostic classification is architecture-dependent: we can only claim a feature exists within an encoded representation if it can be accurately decoded. Since there is no guarantee these features can be recovered from a simple linear transformation, it is worth considering an alternative probing method like Singular Vector Canonical Correlation Analysis (SVCCA) (Saphra and Lopez, 2019), which evaluates the similarity between the hidden representations of the language model of interest and those captured by an independently-trained tagger.

As another use for diagnostic classifiers, we would like to investigate whether structural probes can recover parse trees from models trained on recursive artificial languages like those of Hupkes et al. (2018) and Hupkes et al. (2019). Successful models that can generalize to longer sequences may internally construct parse trees and predict the correct output through recursive sub-tree squashing.

[2] https://nlp.stanford.edu/~johnhew/structural-probe.html

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. Compositionality decomposed: how do neural networks generalise?

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with. *Proceedings of the 2019 Conference of the North*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Kelly W. Zhang and Samuel R. Bowman. 2018. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis.

A Language Models

Language Model	# Hidden layers	Rep dim
Gulordava LSTM	2	650
distilled GPT-2	6	768
XLNet	12	1024

Table 2: Shown are the number of hidden layers and the dimensionality of the hidden represenations for each model.

B Additional Linguistic Probe Results

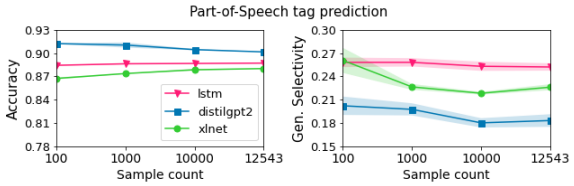

Figure 4: Shown are the POS tag accuracy and the generalized selectivity for different number of sentences in the training set.

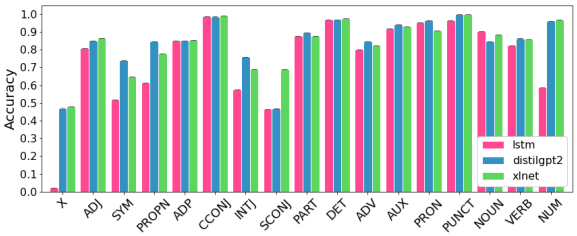

Figure 5: Shown are the accuracies of our best linguistic probe on different types of pos tags when trained on the embedding layer of the language model,

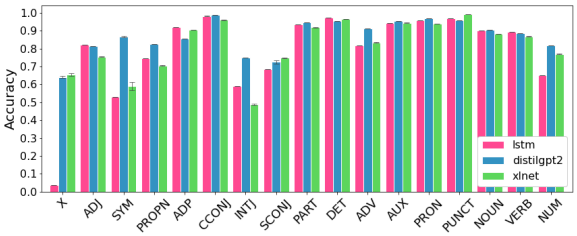

Figure 6: Shown are the accuracies of our best linguistic probe on different types of pos tags when trained on the last hidden layer of the language model.

C Additional Structural Probe Results

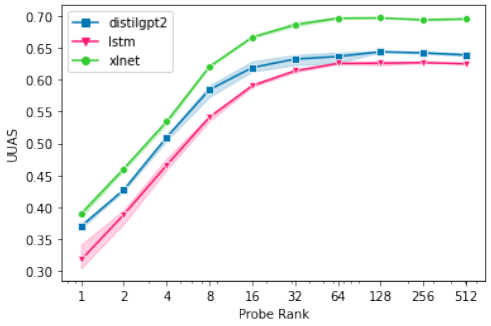

Figure 7: The maximum UUAS of structural probes across all hidden layers as the rank k of the linear transformation B increases. Scores are averaged across three runs, with the shaded regions denoting the standard deviation.

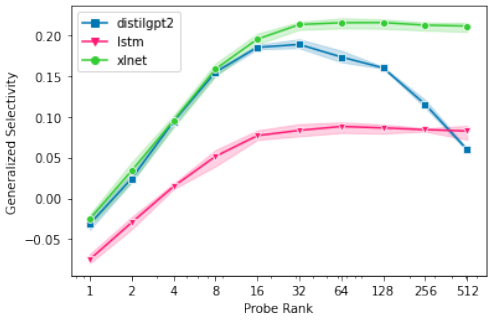

Figure 8: The average generalized selectivity of structural probes across all hidden layers as the rank k of the linear transformation B increases. Scores are averaged across three runs, with the shaded regions denoting the standard deviation.