Reading Assignment 2
Hamza Rashid, 260971031
COMP 550, Fall 2024

Probing a language model for its syntactic knowledge is important for evaluating its ability to encode the grammatical structures of the language at hand, providing concrete architectural insights that can guide troubleshooting and performance improvements in a variety of tasks. The paper, *"A Structural Probe for Finding Syntax in Word Representations"*, by John Hewitt and Christopher D. Manning, proposed a structural probe that tests a language model's ability to embed syntactic information. In particular, the probe was trained to evaluate whether syntax trees are embedded in a linear transformation of a neural network's word representation space. The (pre-trained, english) language models in question were ELMo (embeddings from language model, a BiLSTM) and BERT (bidirectional encoder representations from transformers), the latter utilizing an architecture not covered in class. The authors used the Wall Street Journal section of the Penn Treebank corpus to test each model's ability to capture the Stanford Dependencies formalism. They utilized supervised learning on the dataset to train the probe, namely the matrices that approximate the linear transformations. To guide the research, the language models were compared against baselines that were expected to encode features useful for training a parser, but not be capable of parsing themselves. While the probe found strong evidence that ELMo and BERT encode the desired syntactic knowledge, it is limited by its strict formulation and reliance on annotated data.

The probe was trained to approximate linear (structure-preserving, matrix-form) maps from a language model's contextualized word representation space to the spaces given by two metrics on the parse trees: a squared L2 distance encoding word distances in the tree, and a squared L2 norm encoding depth, both measured in edges. Distances between pairs of words are important for capturing hierarchical behavior, such as subject-verb agreement, while the depth norm captures edge directions and imposes a total order on the words in a sentence. These intuitions are complementary in encoding the structural properties of the parse tree. The authors used mini-batch gradient descent to minimize L1 loss of the post-tranformation squared distances and squared norms over all word-pairs and individual words in a sentence, respectively. While this approach is minimal and provides concrete insights, its bias towards the supervised data makes it prone to filtering out other, possibly correct, syntactic details that are encoded by the language models. An unsupervised method can be useful for modelling such syntactic nuances that are not present in the dataset. However, in this alternative, it is difficult to make concrete hypotheses and observations about how information is encoded by the language models.

The authors probed three representation models: ELMo, BERT-base, and BERT-large. These were compared with simpler baseline models: LINEAR (from the assumption that English parse trees form a left-to-right chain), ELMo0 (character level word embeddings with no contextual information), DECAY0 (assigns each word a weighted average of all ELMo0 embeddings in the sentence), and PROJ0 (contextualizes the ELMo0 embeddings with a randomly initialized BiLSTM layer). The models were evaluated on how well the predicted distances reconstructed the true parse trees and correlated with the parse trees' distance metrics. Given an input sentence, the predicted distances were used to construct the minimum spanning (predicted) tree, which was evaluated on its Undirected Unlabeled Attachment Score – the percentage of undirected edges placed correctly against the true tree. Furthermore, the authors averaged the Spearman correlation between the true and predicted distances for sentences of a fixed length, and macro-averaged the results across all lengths from 5 to 50. Similarly, each model was evaluated on its ability to recreate the order of words specified by their depth in the parse tree, and the Spearman correlations between the true and predicted depth orderings were averaged as before. A closely related evaluation metric, "root%", looked at models' accuracy in identifying the root of the sentence as the shallowest.

Shown in Table 1 of the paper, PROJ0 delivered the best results among the baselines (but still worse than ELMo and BERT), likely owing to its contextual representations. The BERT models consistently outperformed ELMo, with BERT-large's 16^{th} hidden layer, the best overall, achieving scores in the high 80's in the distance and depth correlations, and 90% accuracy in root word predictions. The best hidden layer of BERT-base performed no more than 3% worse than the best layers of BERT-large across all evaluation metrics, grounding the observation that the effective rank of linear transformation required was low compared to each layer's dimensionality, as was shown in figure 5.

Overall, the probe provided strong evidence for the existence of syntax-encoding vector structures in ELMo and BERT. The probe, while simple and effective, has no specific parsing ability, and is prone to missing syntactic details that are not present in the PTB WSJ dataset or cannot be retained in their strict metric formulations. On the other hand, it is difficult to obtain concrete insights in an unsupervised approach. Such tradeoffs between probe complexity, probe task, and hypotheses tested are a common struggle probing literature.

References

Hewitt, J., & Manning, C. D. (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)* (pp. 4129–4138). Association for Computational Linguistics.