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

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 (neural-network based) language model's ability to embed syntactic knowledge in its word representation space. Concretely, the authors test for syntactic knowledge in the form of parse trees, and the probe is structural in the sense that it approximates structure-preserving (linear) maps between a language model's contextualized word representation space and the spaces given by two metrics on the parse trees: one that encodes the distance between words in the tree, and one that encodes depth. The (english) language models in question are 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 use the Wall Street Journal section of the Penn Treebank corpus to test each model's ability to capture the Stanford Dependencies formalism. They utilize supervised learning – gradient descent with the PTB WSJ dataset – to train the probes, namely the matrices that approximate the linear transformations. To aid analysis, 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 proposed method effectively determines that ELMo and BERT capture syntactic knowledge, it is limited by its strict formulation and reliance on supervised learning.

The proposed metrics consist of a squared L2 distance that encodes the distance between words in the parse tree, and one in which squared L2 norm encodes depth in the parse tree, where distance and depth are 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. The authors utilized mini-batch gradient descent on the corresponding matrices to minimize L1 loss of the predicted (post-tranformation) squared distance over all word-pairs in a sentence, or squared norm for each word. While this approach is minimal and effective, the probes are vulnerable to bias towards the supervised data, potentially missing out on structural properties that are captured by the language models, but not captured by the true parse trees or the proposed metrics. In which case, an unsupervised approach can be useful for modelling such syntactic nuances that are not present in the dataset. However, in unsupervised learning, 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. They also probed baseline models for more granular insights: 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). In addition to the L1 loss on word distances, the authors reconstructed the tree given by the predicted distances (via the Minimum Spanning Tree algorithm) and evaluated its Undirected Unlabeled Attachment Score – the percentage of undirected edges placed correctly against the true tree. For distance correlation, they computed the Spearman correlation between true and predicted distances for each word in each sentence, and averaged them between all sentences of a fixed length (from 5 to 50). Similarly, they evaluated the models on their ability to recreate the order of words specified by their depth in the parse tree, computing the Spearman correlation between the true and predicted depth orderings, and averaging across sentence lengths as before. A closely related evaluation metric "root%" looked at models' accuracy in identifying the root of the sentence as the shallowest.

As depicted in Table 1 of the paper, ELMo0 and DECAY0 failed to consistently outperform the LINEAR model. This is likely owing to a lack of context embedding, as PROJ0 delivered the best results among the baseline models – most notably acheiving a score of 73% in its Spearman correlation averaged across the fixed lengths, only 10% less than ELMo1 (first hidden layer of ELMo, and the most performant according to Figure 1). The BERT models consistently outperformed ELMo, with BERT-large's 16th hidden layer achieving scores in the high 80's in the distance and depth correlations and 90% accuracy in root word predictions. The seventh hidden layer of BERT-base performed at most 2.1% worse than layers 15 and 16 of BERT-large across all metrics, grounding the observation that the effective rank of linear transformation required was low compared to each layer's dimensionality in all models.

conclusion

Strengths and Limitations

The approach is innovative in its simplicity and direct measurement of syntax structure, supporting evidence of hierarchical information in pretrained embeddings. The probe's reliance on linear transformations makes it computationally efficient and interpretable. However, limitations include reliance on supervised data, which may bias results, and the probe's inability to capture syntactic nuances that require more complex transformations. Additionally, understanding the probe's practical applications beyond syntactic verification could be expanded upon.

Points of Uncertainty

Some concepts in the probe's design and its assumptions about the geometry of vector space remain unclear. For example, how the squared L2 distances perfectly encode parse tree structures or why a linear transformation is assumed optimal for all syntactic relations. Further clarification on alternative geometries for embedding syntactic knowledge might be beneficial.

Probing Approach: Supervised vs. Unsupervised

The paper employs a supervised probing approach, where parse distances are learned using labeled syntactic data. While supervision enables direct and interpretable syntactic measurements, it limits generalizability across languages and treebank styles. An unsupervised approach could mitigate reliance on labeled data, potentially revealing latent syntactic structures across models, but might struggle to match supervised accuracy.

Importance of Syntactic Probing

Probing models for syntactic knowledge is essential both practically, as it guides model improvements and applications in syntax-sensitive tasks, and theoretically, as it reveals linguistic patterns in language embeddings. Establishing syntax presence in embeddings provides insights into the representation power of pretrained models, crucial for understanding their success in tasks like machine translation and parsing.

References