1 Description of the reading material

The paper "A Structural Probe for Finding Syntax in Word Representations" introduces two innovative probing methods—the structural probe and the depth probe—to examine whether deep neural language models implicitly capture syntactic structures within their word embeddings, specifically testing models like ELMo and BERT.

► **1. Probe 1: Structural Probe.** This probe evaluates whether syntactic relationships between words are encoded in a neural model's word representations by finding a transformation where squared L2 distances between word vectors correspond to syntactic distances in a parse tree. It checks if word pairs with shorter distances in representation space have fewer syntactic edges, revealing hierarchical syntactic structures. The probe ensures non-negativity and symmetry in the distance metric, which is essential for consistent representation, and captures global syntactic relationships, reflecting each word's position relative to others. This method allows for a low-rank, compact syntax embedding within the model's representation space.

► **2. Probe 2: Depth Probe.** The depth probe focuses on another syntactic property—the depth of words within a parse tree. It examines whether the parse depth (i.e., the number of edges from each word to the root) is encoded within the word embeddings by assessing the squared L2 norm of word vectors. This probe is designed to identify transformations that align the word representation space with parse depth, thus establishing if hierarchical depth information, such as the root-word relationship, is implicitly captured by the model.

► **3. Strength and limitation.** A **strength** of these probing methods is their straightforward and interpretable design, which focuses on global syntactic relationships rather than just direct dependencies, effectively highlighting the hierarchical structure of syntax within the model's space. The study finds that syntactic information resides in a low-dimensional, compact subspace, suggesting efficient representation use within the neural network, and ensures non-negativity and symmetry in the distance metric for consistent syntactic relations. However, there are **limitations**. The reliance on linear transformations may limit the capture of intricate language relationships that nonlinear methods might better reveal. **This point brings confusion to me, as the initial assumption of linearity seems incompatible with the inherent nonlinearity of neural networks, raising the question of whether linear transformation is indeed suitable for this task.** Additionally, the finding that squared distance outperforms regular distance leaves some unresolved theoretical questions. The paper also doesn't explore other NN factors, such as gradients or model architecture, due to space constraints.

2 Discussion

► The proposed probing approach in this paper is **supervised**, as it relies on labeled syntactic data to learn a transformation that aligns the word representations with syntactic distances and depths in parse trees. A supervised approach offers the advantage of precision, as it enables the probe to target specific syntactic structures with high accuracy, providing clearer insights into how well the model encodes syntax. However, this dependency on labeled data can limit generalizability, as it may be costly or infeasible to obtain sufficient labeled syntactic data in various languages or domains. In contrast, an unsupervised approach would allow broader applicability across datasets and languages without relying on annotations but may lack the specificity needed to reveal detailed syntactic relationships, making it harder to draw precise conclusions about the model's syntactic knowledge.

► Probing a model at the first place is crucial for both practical and theoretical reasons. **Practically**, understanding whether a model captures syntactic structure can improve its deployment in language technology applications, as syntactically interpretable models tend to perform better in tasks like translation and question answering, where grasping sentence structure is essential. Knowing a model's syntactic capabilities can also help to select or refine models for tasks that require nuanced understanding of language, leading to more accurate and reliable language systems. **Theoretically**, by examining if and how models encode syntax, researchers can evaluate whether neural networks capture similar structures to those in human linguistic theory, such as hierarchical relationships in sentences, or rely on more shallow patterns. This insight contributes to the scientific understanding of language processing in artificial systems and informs computational linguistics.