Tommi Buder-Gröndahl
(University of Helsinki)

Linguistic representations in large language models: Some foundational problems

Due to the black-box status of large language models (LLMs), increasing their explainability on a human-readable level has become a central goal. As part of this endeavour, numerous studies have aimed to interpret LLM-internal structures (known as embeddings) via familiar linguistic formalisms, such as phrase-structures or dependency graphs.

The most prominent methodology here is called probing, and involves mapping embeddings to pre-defined linguistic target labels. Moreover, the experimental literature has regularly endorsed the idea that LLMs internally represent such linguistic structures. However, it is unclear how this claim should be interpreted in the first place: some readings of "linguistic representation" would make assigning them to LLMs trivially true, while others would make it trivially false. Finding an appropriate middle-ground is necessary for making the claim informative to begin with, but turns out to be more difficult than expected. In this presentation, I give an overview of how the problem plays out in the research program of interpreting LLMs, and link it to related debates concerning the status of linguistic representations in cognitive science, linguistics, and the philosophy of language.