Disentanglement of syntactic structures in pre-trained language models
Learning transferable latent structures

Syntactic structures from PLMs

Pre-trained language models (PLMs) like BERT[1] and DistilBERT[2] learn contextual word embeddings by passing tokens through multiple transformer blocks. Their main goal is to encode semantics into the final representation, but the hidden layers have been shown to represent syntactic features [3].

We disentangle constituents from the hidden token representation, by projecting them to a latent space. In this space, the distance between tokens follow the cluster structure in parse trees.

Agglomerative clustering loss function

\[
L_C(x_1, x_2, \ldots, x_n, c) = \frac{\max\{d(x_i, x_j) | x_i, x_j \in C\}}{\min\{d(x_i, x_j) | x_i \in C, x_j \not\in C\}}
\]

Validation

We trained the latent space using Ontonotes. Then for validation we project sentences from different NER datasets into the latent space, run hierarchical agglomerative clustering on the latent space vectors and measure recall for multi-word named entities. The model is compared against Berkeley Neural Parser[4].