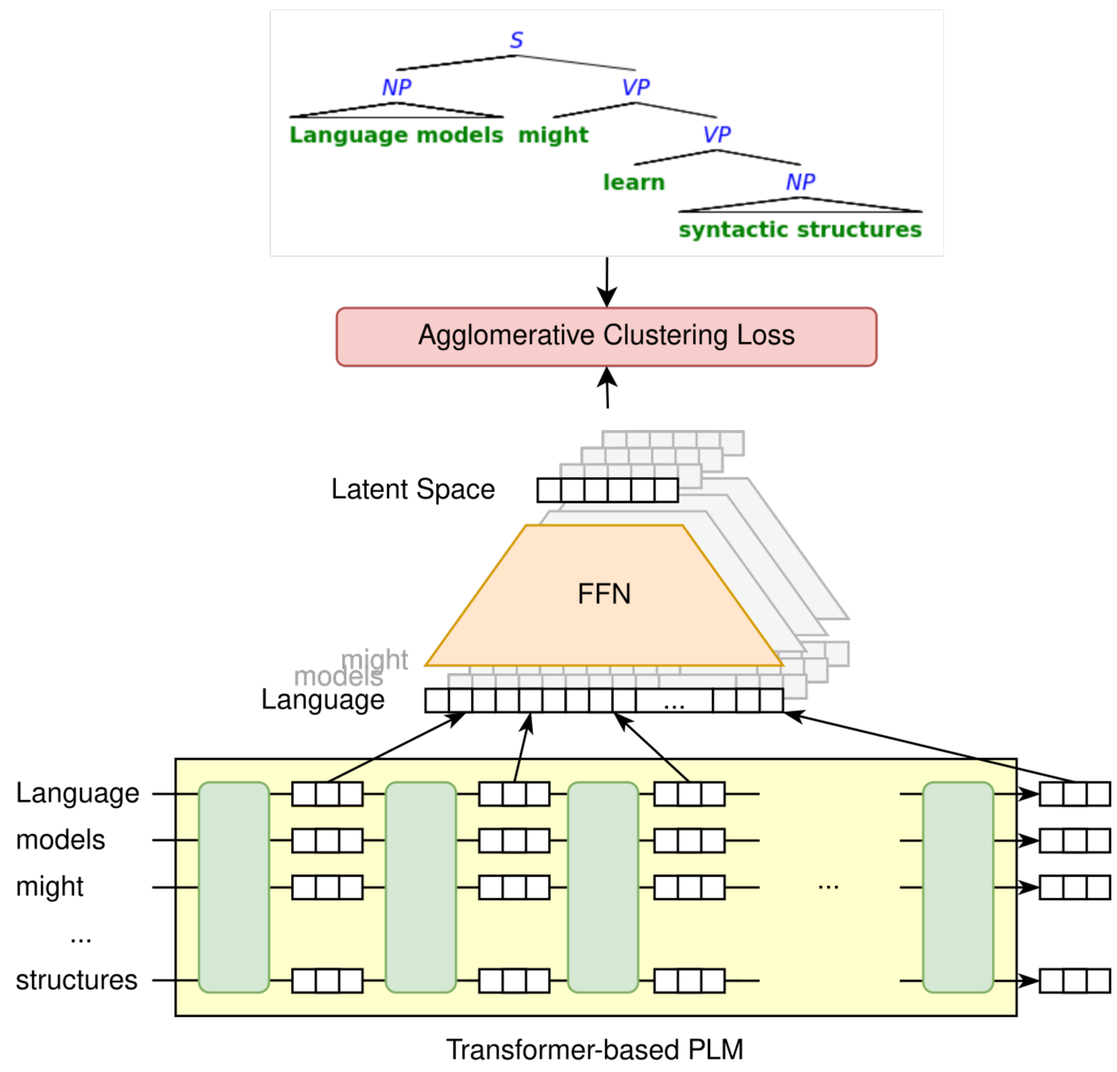
Disentanglement of syntactic structures in pre-trained language models

Learning transferable latent structures

Syntactic structures from PLMs

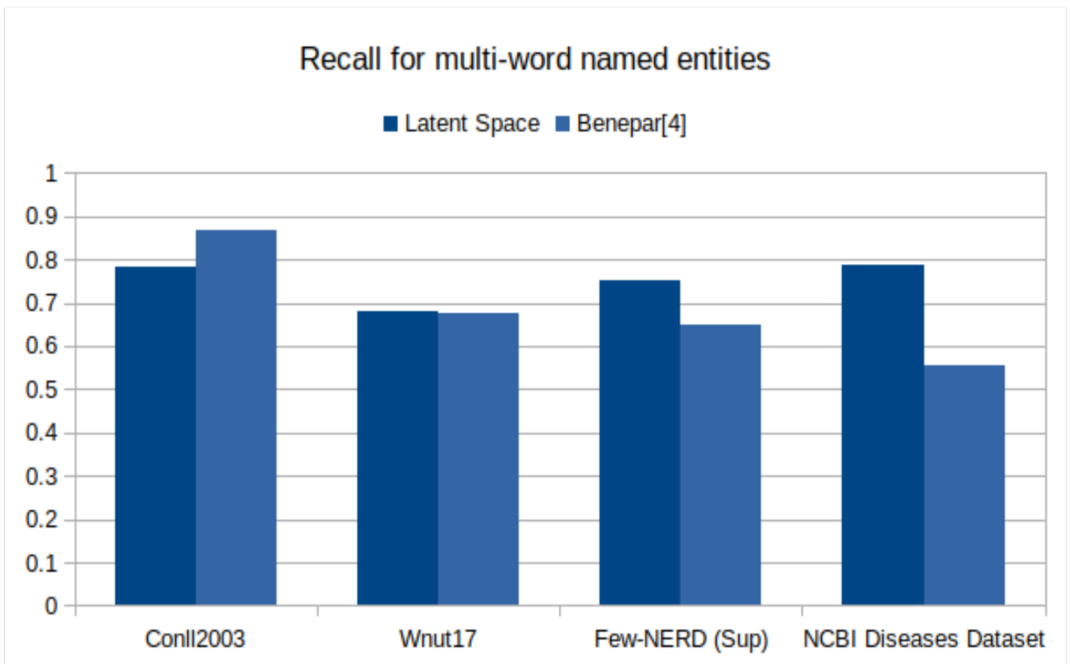
Pre-trained language models (PLMs) like BERT[1] and DistillBERT[2] learn contextual word embeddings by passing tokens through multiple transformer blocks. Their main goal is to encode semantics into the final representation, but the hidden layers have been shown to represent syntactic features [3].

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

$$L_C(x_1, x_2, \ldots, x_n, c) = \frac{max\left(\{d(x_i, x_j) | x_i, x_j \in C\}\right)}{min\left(\{d(x_i, x_j) | x_i \in C, x_j \notin C\}\right)}$$



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]

[1] Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”. NAACL (2019)
[2] Sanh, Victor et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” ArXiv abs/1910.01108 (2019)
[3] Hewitt, John and Christopher D. Manning. “A Structural Probe for Finding Syntax in Word Representations.” NAACL (2019).
[4] Kitaev, Nikita and Dan Klein. “Constituency Parsing with a Self-Attentive Encoder.” ACL (2018).

HPI Research School "Data Science and Engineering"
Alejandro Sierra-Múnera

E-Mail: alejandro.sierra@hpi.de
Tel.: 033155093401
Web: https://hpi.de/en/research-schools/hpi-dse/mitglieder/research-pages/alejandro-sierra-munera.html
August-Bebel-Straße 88 (Building F, Campus II) 14482 Potsdam

Hasso Plattner Institut