Universal Dependencies According to BERT: Both More Specific and More General
Tomasz Limisiewicz, David Mareček, Rudolf Rosa

Goal
We introduce a head ensemble method, combining multiple attention heads which capture the same dependency relation label.

Dependency Accuracy
\[ \text{DepAcc}_{i,i} = \frac{|\{(i,j) \in E_{i,i} : j = \arg \max A[i]\}|}{|E_{i,i}|} \]
- \(E_{i,i}\) - all directed dependency edges
- \(A[i]\) - \(i^{th}\) row of the attention matrix

Ensembles Overlap

Extracted trees are directed and labeled.

Key Findings
1. Using head ensembles instead of single heads improves:
   a. Average DepAcc: 67.8% → 74.1%
   b. UAS: 37.2% → 52.0%
   c. LAS: N/A 21.7%
2. We have observed many-to-many relationship between heads and syntactic functions.
3. The method is effective for 9 typologically diverse languages.

Dependency Tree Extraction
- Trees are extracted from averaged head ensembles by an MST algorithm. Similar approach to (Raganato and Tiedemann, 2018)
- Extracted trees are directed and labeled.