Investigating UD Treebanks via Dataset Difficulty Measures

UPPSALA UNIVERSITET

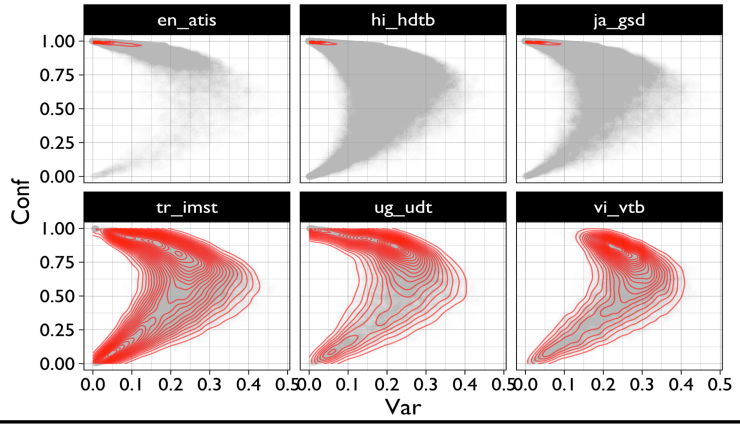
Artur Kulmizev Joakim Nivre

Dataset Cartography

Which treebanks appear hard or easy to parse, given a model's confidence throughout training, and variability therein?

$$\textsc{Conf}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta_e}(y^*|\mathbf{x})$$

$$\textsc{Var}_i = \sqrt{\frac{\sum_{e=1}^{E}(p_{\theta_e}(y^*|\mathbf{x}) - \textsc{Conf}_i)^2}{E}}$$



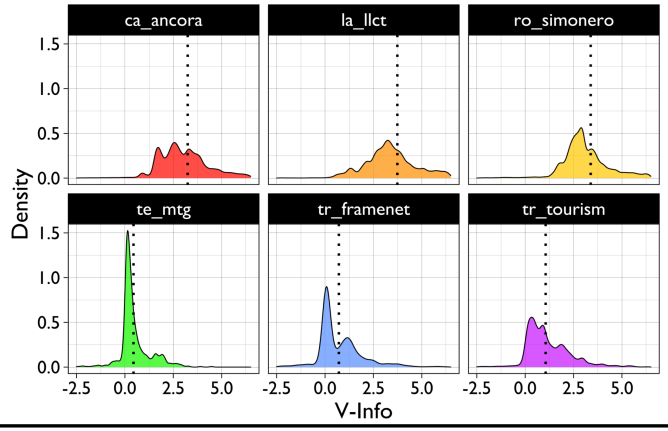
V-Information

Which treebanks contain the most (or least) information that is actually usable by a parser, with respect to a naive baseline?

$$H_{\mathcal{V}}(Y|X) = \inf_{f\in\mathcal{V}}[-\log f_\theta[X](Y)]$$

$$H_{\mathcal{V}}(Y) = \inf_{f\in\mathcal{V}}[-\log f_\theta[\varnothing](Y)]$$

$$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|X)$$



Minimum Description Length

Which treebanks are the most (or least) sample efficient, i.e. most easily fit by a parser, irrespective of training set size?

$$L^{\text{online}}(y_{1:n}|x_{1:n}) = \sum_{s=0}^{S-1}\sum_{n=t_s}^{t_{s+1}} -\log_2 p_{f_s}(y_n|x_n)$$