Investigating UD Treebanks via Dataset Difficulty Measures

Artur Kulmizev **Joakim Nivre**
Department of Linguistics and Philology
Uppsala University
{artur.kulmizev,joakim.nivre}@lingfil.uu.se

Relevant UniDive working groups: WG1, WG3

1 Introduction

Datasets have long played a crucial role in dictating the pace of progress in NLP. Their function, for most tasks, is largely two-fold: 1) to collect data points (and their corresponding gold-standard labels) on which statistical models can be trained, and 2) to serve as benchmarks though which various models can be evaluated and compared. In recent years, much research has been devoted towards developing new datasets, tasks, and benchmarks for NLP — so as to articulate the distinguishing aspects of a bevy of new neural models. Syntactic parsing has remained an active area of research in this regard, and Universal Dependencies (UD) (Nivre et al., 2016, 2020) has emerged as a crucial initiative within NLP, offering a set of cross-lingually consistent annotation principles that have since been adapted to 217 treebanks that span 122 languages and 18 domains (version 2.9).

Though UD and other initiatives have aided in driving recent advances in NLP, overall progress has typically been measured via aggregate accuracy metrics, which provide little more than a bird's eye view into the data. In the era of deep learning, where popular models are notoriously opaque, it has thus proven vital to study the contents of datasets and identify aspects that may misrepresent model performance. In this vein, numerous studies have shown that the crowd-funded nature of some popular NLP datasets makes them prone to annotation artefacts that are readily exploitable by neural models as heuristics (Kaushik and Lipton, 2018; Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019). With such insights in mind, researchers have shifted their focus towards the *datasets* instead of the models, proposing general methods for exploring the former so as to better understand the performance of the latter. Such approaches have drawn from, e.g., information theory (Perez et al., 2021; Ethayarajh et al., 2022), item response theory (Rodriguez et al., 2021; Vania et al., 2021), and model training dynamics (Swayamdipta et al., 2020). This work, however, has predominately focused on classification tasks and has proven difficult to extend to other classes of problems, such as the structured prediction tasks of UD.

In this paper, we perform an analysis of 88 Universal Dependencies (UD) treebanks through the perspective of a popular parsing architecture — namely that of Dozat and Manning (2016). As opposed to much previous work, which prioritizes metrics like LAS in order to build accurate parsers, we aim instead to better understand the underlying data, as well as how our parser interfaces with it. To do so, we extend recently proposed dataset analysis methods based on model training dynamics (Swayamdipta et al., 2020), $\mathcal{V}$-information (Xu et al., 2020; Ethayarajh et al., 2022), and minimum description length (Blier and Ollivier, 2018; Voita and Titov, 2020; Perez et al., 2021) to the dependency parsing scenario. In working with each method, we formalize the following set of research questions:

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

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

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

2 Dataset Cartography

Dataset cartography (DC) consists of two complimentary measures: Confidence (CONF) and Variability (VAR). CONF refers to the average probability assigned to a token w_i by a model M (e.g. a parser) after training for E epochs. VAR is corresponding standard deviation of this value. If CONF is high and VAR is low, w_i is considered "easy to learn". Conversely, if both values are low, then

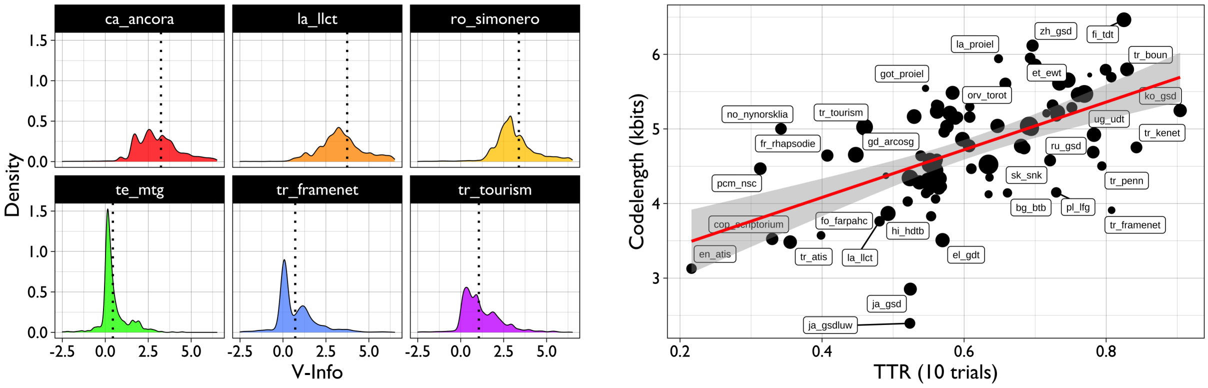

Figure 1: Left: Arc-level PVI density for Top-3 and Bottom-3 V-INFO treebanks, across arcs (labels omitted for space). Right: Block-wise codelength (in bits) for Top-3 and Bottom-3 MDL treebanks, across arcs.

w_i is considered "hard to learn"; such cases often correspond to annotation errors and various other artefacts (Swayamdipta et al., 2020). We find that DC, when applied to UD, is capable of painting a nuanced picture of how *easy* or *hard* treebanks might be to parse. In zooming in on the three "easiest" treebanks — English Atis, Hindi HDTB, and Japanese GSD — we observe that the majority of arcs contained therein are trivially fit by the parser. Conversely, the "hardest" treebanks — Turkish IMST, Uyghur UDT, and Vietnamese VTB — contain much variation, with Turkish exhibiting a particularly high density of "hard to learn" arcs.

3 $\mathcal{V}$-information

$\mathcal{V}$-information (V-INFO) is an information-theoretic measure introduced by Xu et al. (2020), which quantifies the amount of "usable" information that can be extracted by M from a dataset D. V-INFO presupposes the use of two models: an M trained on D in a straightforward fashion, and a "baseline" model M' trained on a corrupted version version of D (e.g. all input tokens replaced with _). Given M and M', V-INFO is calculated by taking the sum of differences between negative log probabilities yielded by M' and M for each token $w_i \in D$ (the validation partition thereof). High V-INFO values indicate that D contains much information that cannot be inferred by naive baselines, while low values indicate that M would not fare worse than picking a class at random. We demonstrate that V-INFO can reveal nuanced treebank characteristics when applied to UD. For example, Turkish Tourism yields a low V-INFO score due to its limited genre, short sentence length, and persistent placement of various words at fixed positions. In contrast, treebanks like Latin LLCT and Romanian SiMoNERo return high V-INFO scores, indicating that they possess a varied distribution of structures and vocabulary usage.

4 Minimum Description Length

Minimum Description Length (MDL) (Rissanen, 1978) is an information-theoretic measure that captures how well M can compress D. MDL can be estimated via online coding (Rissanen, 1984; Blier and Ollivier, 2018): a technique which splits D into S blocks and measures the fit of M on each successive block. Intuitively, MDL expresses the ability of M to generalize with respect to D: models that learn efficiently from limited instances will yield shorter codelengths (lower MDL). In our experiments, English Atis and Japanese GSD return the lowest MDL overall, indicating that they are the most sample efficient. We attribute this to the former's limited genre and the latter's tokenization scheme. In general, we show that MDL correlates strongly with morphological complexity metrics across treebanks, indicating that it is influenced by typological factors, vocabulary usage, and token frequency.

Acknowledgements

This is an extended abstract of work to appear at EACL 2023. We would like to thank Anders Søgaard for providing useful feedback on an early version of this work, as well as Sasha Berdicevskis for lending insight about computational measures of morphological complexity. We acknowledge the computational resources provided by CSC in Helsinki through NeIC-NLPL (`www.nlpl.eu`).

References

Léonard Blier and Yann Ollivier. 2018. The description length of deep learning models. *arXiv preprint arXiv:1802.07044*.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. pages 5988–6008.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. *arXiv preprint arXiv:2103.03872*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Jorma Rissanen. 1984. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information theory*, 30(4):629–636.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158, Online. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*.