Après Toi: Scoring Systems based on Dataset Votes

Yuval Pinter
Department of Computer Science
Ben-Gurion University of the Negev
Beer Sheva, Israel
uvp@cs.bgu.ac.il

Relevant UniDive working groups: WG4

1 Introduction

In order to promote universality and diversity in development of NLP systems, some attention should be directed towards defining evaluation metrics for multilingual tasks that reward systems for their success in handling different languages. In current practice, the most common metric is a simple average of performance over all datasets, which may undermine the goal of diversity in several ways. First, the inherent difference in what constitutes a good score can vary widely between languages, making a given difference in score between systems either very meaningful or not. For example, when dependency parsing is evaluated using labeled attachment score (LAS), languages with strict word order might be easily parsed to the 0.9 level with errors being sparse but crucial, whereas languages with free word order might be challenging to parse past 0.6, with handling of individual problems resulting in large gains. Second, the method of score aggregation by averaging gives no incentive for system developers to improve performance through work on a diverse set of languages; since a given point improvement on any language is worth the same in the overall score, it is likely the motivation is in fact to continue working within the "comfort zone" of the system so far.

We propose **voting-based** score aggregation in multiple datasets. The principle is simple: each dataset d is given a budget of points it allocates to systems $s \in \mathcal{S}$ based on their rank on that dataset, $r_d(s)$, where the rules of point allocation are the same for all datasets and are a monotonic decreasing function of the rank, $p(r)$. As long as there are no ties, this results in each dataset allocating the exact same total number of points, satisfying a condition of equity among datasets.[1] Finally, systems are ranked based on the total number of points allotted to them, $P_{\mathcal{D}}(s) = \sum_{d \in \mathcal{D}} p(r_d(s))$. This shift in focus from raw numbers to ranks answers both of the shortcomings identified above: credit is gained by surpassing other systems rather than by constant score increases, meaning that large gains in points are made through improving past the levels that challenge many other systems, leading to more meaningful practical breakthroughs; and the "lowest-hanging fruit" for improving a system is not by staying in its comfort zone, but rather by shifting the development effort towards datasets and languages that suffer to that point from underattention.

2 Pilot Study—Universal Dependencies

The *multilingual parsing* task took place in two editions in 2017 (Zeman et al., 2017) and 2018 (Zeman et al., 2018) as part of the CoNLL conference, studying the ability of models to produce dependency parse trees in the UD framework for multiple languages. We chose it as suitable to examine our proposed aggregation method due to the large numbers of both treebanks (81 and 82, respectively, representing a set of languages that is diverse by many criteria) and participating systems (33 and 26, respectively), with all systems reporting results on all treebanks. In this study, we focus on the LAS metric, which was highlighted by the task organizers and is agreed to be the main metric of interest.[2]

We implement three versions of voting aggregation: **Rank Complement (RC)** simply subtracts the rank of the system from the overall system count ($+1$); **Top Ten (TT)** subtracts the rank for the top ten performing systems from 11 and assigns the rest zero points; **Eurovision (ESC)** scoring modifies TT by assigning the top two systems 12 and 10 points respectively, thus upweighing the importance of outperforming all systems on an individual dataset.

[1] Ties may also be accounted for while preserving this principle. For the pilot study below, tied systems were given points according to the best rank in the range. This did not happen frequently and only affected low-ranking systems.

[2] Other metrics include UAS and POS tagging accuracy, but also upstream tasks like tokenization and word segmentation where many systems opted for a baseline solution, resulting in mostly uninteresting rankings.

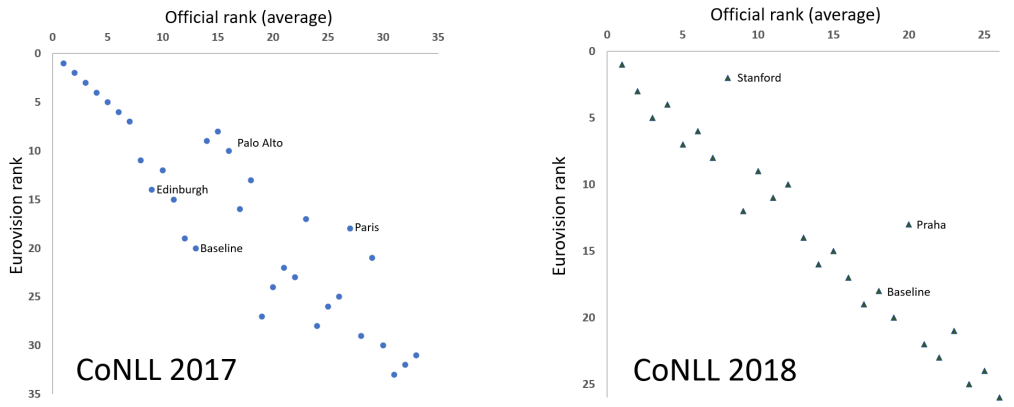

Figure 1: Comparison of system ranks in official scoring and ESC scoring in CoNLL the 2017–8 shared tasks.

2.1 Results

Plots demonstrating the difference in overall system ranking between the averaging metric (x-axis) and the ESC metric (y-axis) in the 2017 (left) and 2018 (right) tasks are presented in Figure 1. The TT metric produced nearly identical results to ESC, and RC's were somewhere in between it and averaging. The scoring method produces interestingly different rankings in both cases, and the differences themselves seem to exhibit different patterns.

The top seven systems of the 2017 tasks are surprisingly robust to the change in aggregation, indicating that systems were calibrated for winning by improving over all datasets in a balanced manner. Meanwhile, the rest of the systems evidently followed different strategies, leading to large gains in the averaged metric for systems such as Edinburgh's, while more diversity-friendly ones like Paris's (de La Clergerie et al., 2017) may have been unfairly treated by the race for average. The baseline UDPipe system also proved inadequately balanced, ranking low on many datasets but keeping a high average.

In 2018, however, the results follow a different pattern. The Stanford system (Qi et al., 2018) vaulted from a mediocre average-based ranking to second place thanks to its top performance on many of the datasets. Praha's system followed the same trend further down the pack, but all other systems oscillated within three places of their ranks.

Another advantage of the voting system is that it produces score lists with much larger margins between systems, helping readability and interpretation. In the average-ranked list for 2018, five top-ten system pairs are within 0.1 point of each other, and #1 is five points above #10. In the ESC table, the smallest margin is nine ranking points, and the margin between #1 and #10 is 572 points.

3 Related Work

Several alternative treatments of system evaluation have been proposed over the years, and we intend to compare their outputs to our systems' as well. One particular example is the **paired evaluation** framework, such as the one suggested by Peyrard et al. (2021), whose empirical tests suggest a change in SoTA for important tasks. While also a ranking-centric measure, we note that both its implementation and application are substantially more complex than our proposal.

4 Planned Work

For the next steps in the experimentation of voting-based aggregation strategies, we aim to collect data from many other multilingual shared tasks such as SemEval (e.g., Barnes et al., 2022) and Sigmorphon (e.g., Batsuren et al., 2022), and perform similar analyses on their results; to extend the set of voting methods examined and produce their scores as well (this can include methods not solely dependent on rank but still distributing an equal budget per dataset); to investigate the notion of statistical significance when considering rank-based measures; to produce a human-based meta-evaluation protocol for multi-dataset system ranking and running it against the various methods; and finally, to contact developers of systems particularly excelling in voting-based aggregation and solicit them for possible reasons for this success, collating their responses for the use of the UniDive action in its publications as recommendations for practitioners.

References

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Éric de La Clergerie, Benoît Sagot, and Djamé Seddah. 2017. The ParisNLP entry at the ConLL UD shared task 2017: A tale of a #ParsingTragedy. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 243–252, Vancouver, Canada. Association for Computational Linguistics.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. 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 2301–2315, Online. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.