WG4

2nd General Meeting
University of Naples "L'Orientale Naples, Italy,
8-9 February 2024
https://unidive.lisn.upsaclay.fr/

Après Toi: Scoring Systems based on Dataset Votes

Yuval Pinter
Ben-Gurion University
uvp@cs.bgu.ac.il

The Problem

- When running a multi-dataset competition between systems, we use the **averaging** aggregate metric for deciding the winner
- On one hand, this ensures that small datasets, typically representing low-resource languages, are viewed as equivalent to large datasets
- On the other hand, language differences may lead to lower points of saturation for some, leading to focus work on "easy" languages (or on one's "comfort zone")

System 1 wins on the average metric, but is it really the best?

	Dataset A	Dataset B	Dataset C
System 1	90	53	79
System 2	80	56	82
System 3	70	57	83
System 4	60	58	84

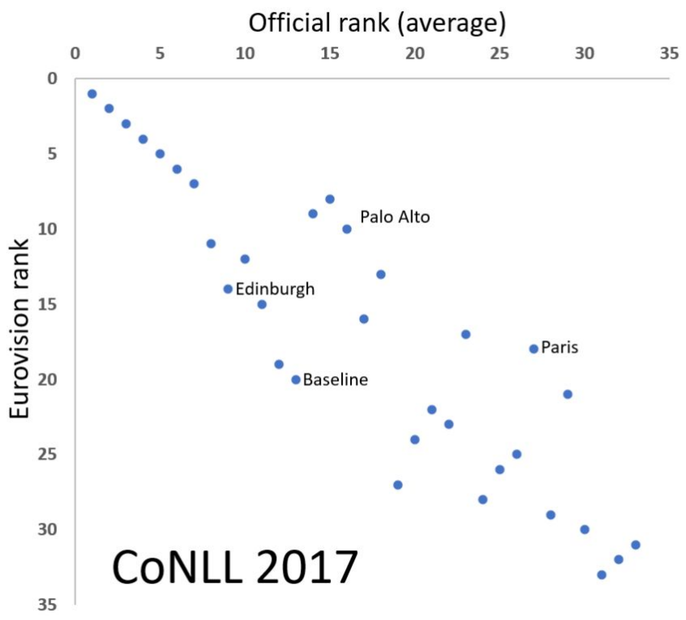
The Alternative(s)

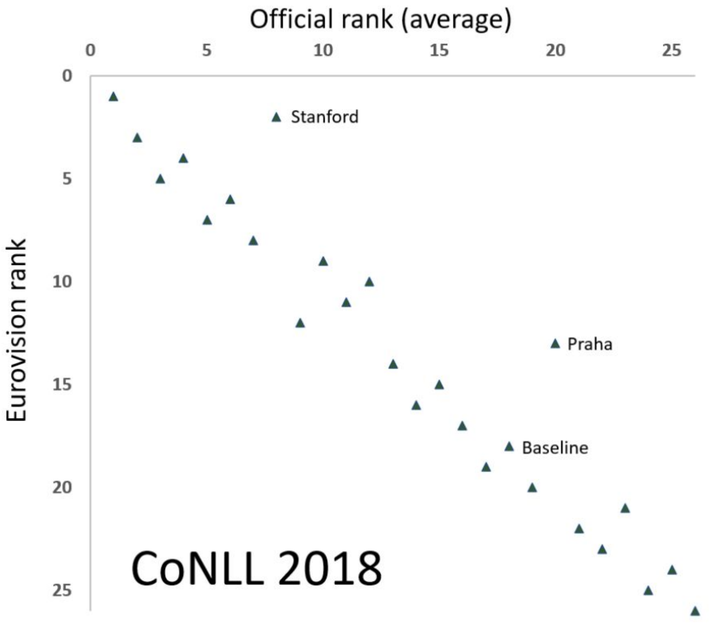
Voting-based scoring

- Each dataset has a budget of "votes" which it distributes among the systems
- In Eurovision-style voting (ESC):
  - Top system gets 12 points
  - #2 gets 10
  - #3 gets 8, and #4-#10 get one point less each
- Final scores are the accumulations from the datasets

Pilot Study

- CoNLL 2017-2018 multilingual parsing challenges
  - 81--82 treebanks
  - 26--33 participants
- Metric: LAS
- Comparison: average vs. ESC ranks
  - Pearson's correlations: 0.90, 0.95
- Not seen here: point ranges, which are much more human-friendly in the ESC setup





Future Directions

- Adding **more voting systems**
  - Paired evaluation (e.g. Elo)
- Adding **more benchmarks and competitions**
  - SIGMORPHON
  - SemEval
  - …
- Incorporating **significance testing**
- Correlating with **Human** notions of winning

Join me!
uvp@cs.bgu.ac.il

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
UniDive
Funded by the European Union