UD Shared Task Inside Out

Proposal of a UniDive WG3-organized evaluation campaign by Dan Zeman (2023-09-08, Istanbul)

There have been several shared tasks with at least partial focus on morphosyntactic parsing of low-resource languages. Some participating systems were able to use multi-lingual or cross-lingual models to parse languages for which there is no (or very little) labeled training data. Typically, the performance of such models is not great; if the winning system has over 60% LAS, it may be impressive in the context of nonexistent training data, but the produced analysis is still too bad to be actually useful for anything. Therefore I think we should use the shared task to help us improve the data situation for under-resourced languages.

I envision focus on traditional basic UD annotation (segmentation, UPOS, features, lemmas, relations), although a similar approach could be applied to other annotation layers. Unlike in "normal" shared tasks, the goal would not be to identify the tool that can predict annotations closest to a gold standard we provide. The real goal would be to *obtain* the gold standard data.

Instead of preparing the gold standard before the task, we would only provide raw text data and identify at least one person per language who commits to review the submitted annotations during the evaluation phase. In accord with the goals of UniDive, we would try to provide languages that have little or no annotated data in UD so far, and preferably from families and genera that are underrepresented. In the initial stage of the task, the community would be encouraged to add more languages (including the commitment to review submissions).

We would also have to sketch the backbone of the language-specific UD guidelines for the language in question. If nothing else, then the language will have to be registered in the UD infrastructure and provided with lists of auxiliaries, relations, permitted UPOS-feature-value combinations etc. so that the UD validator can check data in the language. Moreover, the person(s) who will do the evaluation may want to figure out guidelines for some specific constructions. But this does not have to be perfect and detailed before the task.

From the point of view of resources used by the participants, this would be the most open shared task ever. People could use large language models, polyglot models, cross-lingual models, even models trained on data that are not publicly available, if their terms of use do not prevent the resulting annotations from being freely available. People could employ rule-based heuristics or even annotate the data manually (but the participant must not be the only reviewer of the given language, so that we have at least two people looking at each sentence). The credit is not for the best parsing/tagging algorithm but for the contribution to putting another language on the NLP map. Participants are welcome to provide annotations for all languages in the task, or just for a subset, maybe only for one language.

The evaluation is obviously the tricky part here, and the evaluation phase will have to be significantly longer than is typical for NLP shared tasks. I assume that the reviewers would post-edit the submitted annotations, then the standard UD `eval.py` script (in the UD tools repo, originally from the CoNLL 2017 and 2018 tasks) would be used to compare the submitted data with the post-edited data. The submissions would be required to pass standard UD validation before being reviewed. The most difficult part is to figure out how to invest the reviewer's time so that we get a

“large” and good enough dataset, but still pay a fair amount of attention to each submission. A fully automatic LLM-based polyglot model could annotate millions of words in dozens of languages, but we can hardly verify them all. Assuming that we have only one reviewer for language L, and N_L submissions for that language, we could have the reviewer look at M sentences of each submission (random sentences? same for all submissions, or different?), where M depends on how large N_L is.

After this initial stage, “bad” submissions would be identified and disregarded, while one or more “good” submissions would be reviewed further. If there are multiple “good” submissions, we could use some tools to highlight their differences and possibly take advantage of them.

It is not clear how exactly we should measure that a language is undeveloped enough to be selected for this task. It may not be crucial, as we can do this even for English if there are people willing to review English. But (as per Joakim’s idea) the credit for the participating systems should be inversely proportional to the amount of data that existed for the language before. (Plus possibly some fancy tweaking: How much of it is in large language models? Is there anything for closely related languages?) And a system that helps with multiple low-resource languages is better than a system that only helps with one.

Besides classical system description papers in the shared task proceedings, it would make sense to write language description papers, with the reviewer and the submitters being co-authors.

Additional remarks from the discussion in Istanbul

- If possible, the raw data on input should be mixed genres (cf. WG2.2).
  - Spoken data?
  - Social networks?
  - More parallel corpora?
- We may try to combine outputs of multiple systems.
- Serge: We may want a quality estimation task similar to what people do at WMT. The participating system estimates what is good/bad in a parser’s output, without seeing gold standard data.
- Verginica: Do a task that will improve annotation of existing treebanks and make them closer to each other.
- Verginica: A task that combines UD and PARSEME annotation.