Enough is Enough! A Case Study on the Effect of Data Size for Evaluation Using Universal Dependencies

Rob van der Goot, Zoey Liu, Max Müller-Eberstein

Motivation

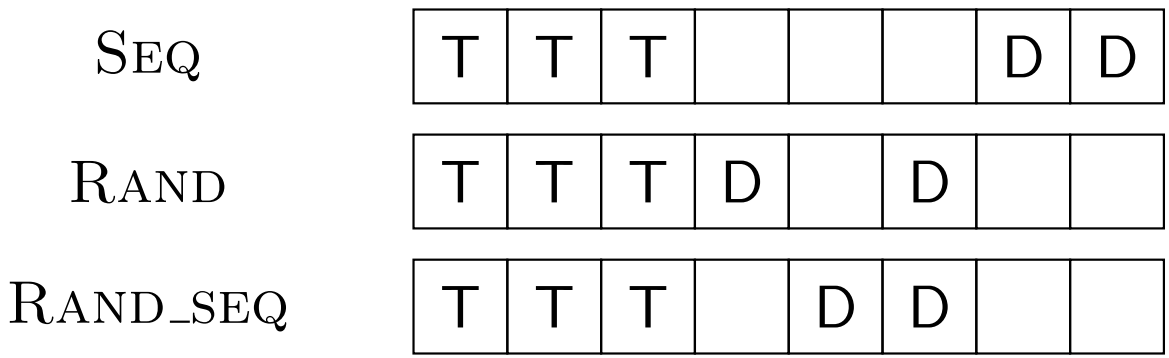
Too large evaluation datasets:

- less training data
- less coverage
- more costs

Setup

- Compare 33 multi-lingual language models for parsing
- Use 50k for training, rest for obtaining the "gold ranking"
- Get ranking for smaller sizes, use Kendall's Tau for comparing them

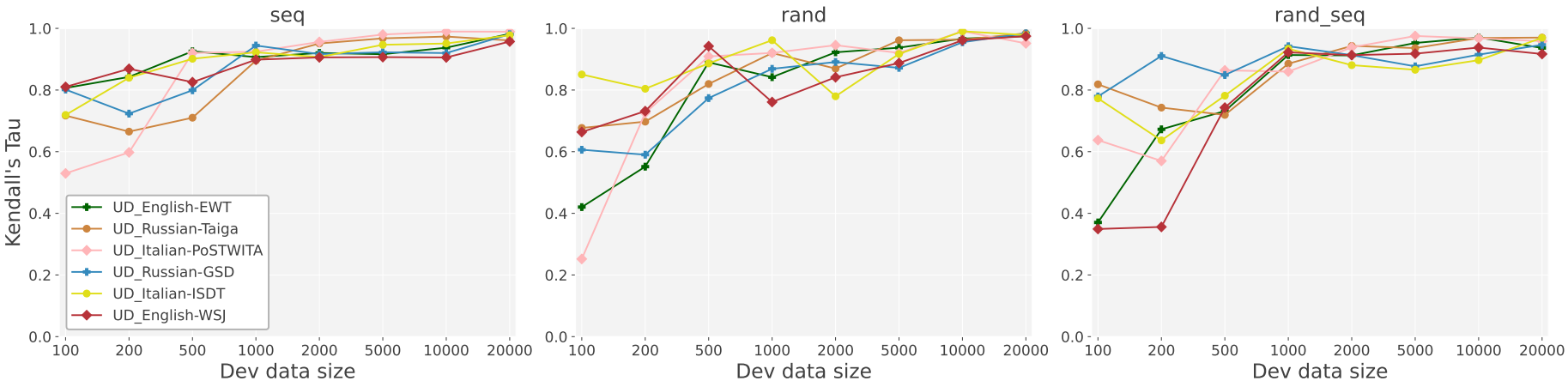
Splitting strategies



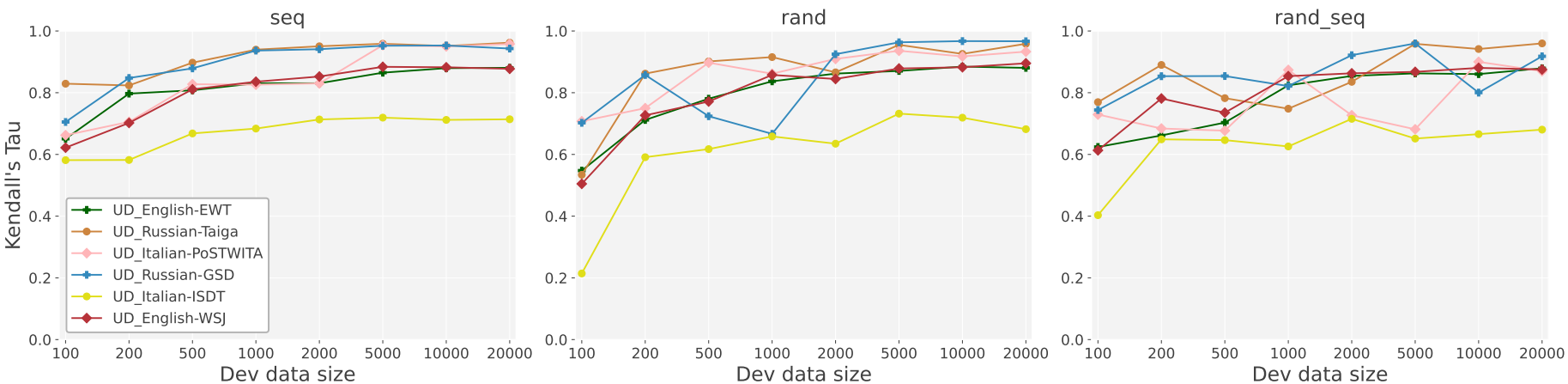
Data

Domain/Language	English	Italian	Russian
News	WSJ	ISDT	GSD
Web	EWT	PoSTWITA	Taiga

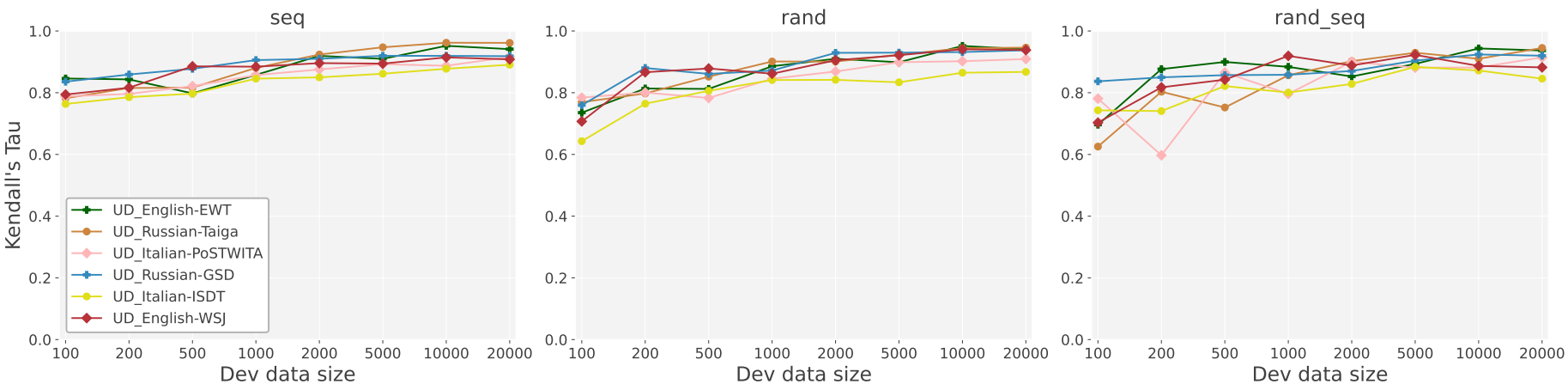
Results


seq
rand
rand_seq
Kendall's Tau
Dev data size
1.0
0.8
0.6
0.4
0.2
0.0
100
200
500
1000
2000
5000
10000
20000
UD_English-EWT
UD_Russian-Taiga
UD_Italian-PoSTWITA
UD_Russian-GSD
UD_Italian-ISDT
UD_English-WSJ

Results Cross-domain



Results Cross-lingual



Thanks!

Code and predictions are available:
`https://bitbucket.org/robvanderg/data_size/`