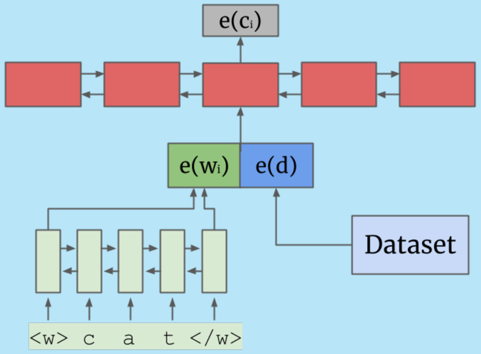
On the Effectiveness of Dataset Embeddings in Mono-lingual, Multi-lingual and Zero-shot Conditions

IT UNIVERSITY OF COPENHAGEN

Rob van der Goot, Ahmet Üstün and Barbara Plank

university of groningen

Models



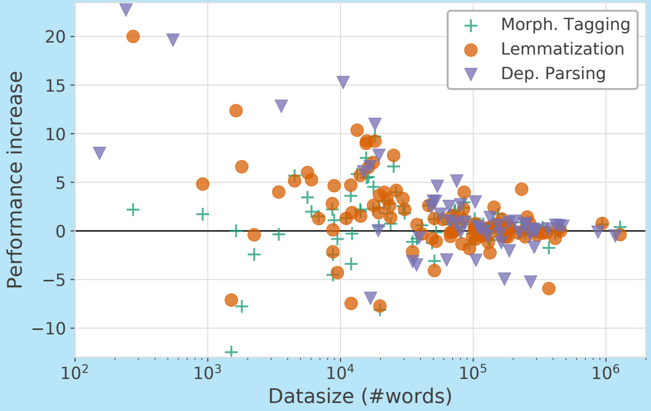
base Stacked BLSTM parser (Smith et. al, 2018) and morphological analyzer (Üstün et. al, 2019)

concat Concatenation of all datasets in cluster

gold Gold dataset embeddings

pred Use SVM to predict datasource to embed

Analysis



In-distribution Results

Filtering	#src	Morphological Tagging (F1) base	concat	gold	pred	Lemmatization (Accuracy) base	concat	gold	pred	#src	Dependency Parsing (LAS) base	concat	gold	pred
All	104	92.04	91.43	92.75	91.85	91.10	91.02	92.55	91.41	58	72.92	74.07	75.53	74.52
Single-lang	59	94.14	93.94	95.84	94.13	93.66	93.83	95.73	93.84	10	80.48	79.84	82.74	80.29
Multi-lang	45	89.30	88.14	88.69	88.88	87.75	87.33	88.38	88.22	48	71.35	72.87	74.03	73.32

Out-of-distribution results

	#src	concat	pred
All	53	53.80	**53.87**
∃ same-lang	35	66.35	**66.62**
∄ same-lang	18	**29.39**	29.06

Conclusions

- Dataset embeddings most useful for single language clusters on in-distribution data
- Predicted dataset embeddings result in slightly lower performance increase
- On out-of-distribution performance increase vanishes

  Source code is available at:

  `https://bitbucket.org/robvanderg/dataembs/src`