Mila

RLLab
McGill University

Vārta: A Large-Scale Headline-Generation Dataset for Indic Languages

Rahul Aralikatte*, Ziling Cheng*,
Sumanth Doddapaneni, Jackie Chi Kit Cheung

- **Vārta:** a large-scale high-quality corpus containing 41.8M news articles for 14 Indic languages and English.
- **Headline generation on Vārta:** challenging even for state-of-the-art text generation models.
- **Vārta as a pretraining corpus:** models show strong performance on Indic NLU and NLG benchmarks.