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

Rahul Aralikatte[1,2*], Ziling Cheng[1,2*], Sumanth Doddapaneni[3,4], Jackie Chi Kit Cheung[1,2,5]

[1]Mila– Quebec Artificial Intelligence Institute [2]McGill University [3]IIT Madras [4]AI4Bharat [5]Canada CIFAR AI Chair

Motivation

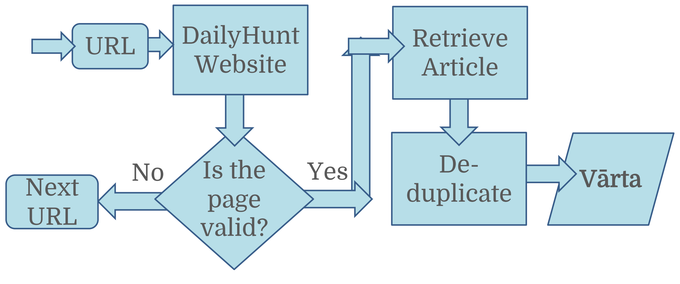
- Indic languages in models: low-resource and under-represented
- Quality data: crucial for downstream task performance
- Family-specific pretraining: enhanced language transfer

Main Contributions

Vārta, a **diverse, challenging, large-scale, multilingual, high-quality** headline generation dataset for **14** Indic languages

- For multilingual summarization research
- For strong language models pre-training

Data Collection & Processing



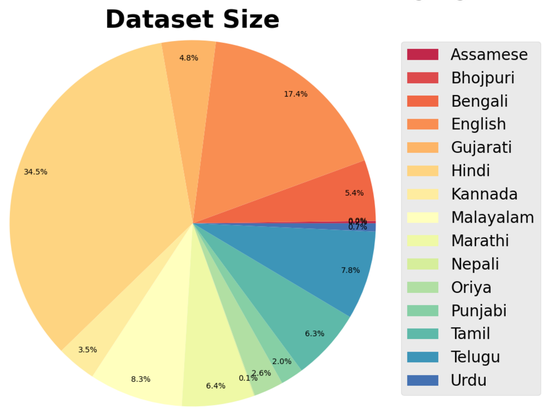
Data Statistics

- Comparison with Other Datasets

	MLSum	XLSum	Indic-Headline	Vārta
# of langs.	5	44	11	15
# of Indic langs.	0	9	11	14
Headline present	✗	✓	✓	✓
Summary present	✓	✓	✗	✗
Size of Indic parts	0	165K	1.3M	34.5M
Total size	1.5M	1M	1.3M	41.8M

Table 1: Comparison of existing multilingual headline generation and summarization datasets with Vārta.

- Data Size Distribution Across Languages



- Statistics & Properties of Vārta

Vārta	Ratio of novel n-grams (%)				CR	|article|		|headline|	
	1-gram	2-gram	3-gram	4-gram	(%)	Tok.	Sent.	Tok.	Sent.
Avg.	29.13	62.96	77.63	83.32	5.72	302.43	16.99	12.22	1.13

Table 2: Average statistics on Vārta. CR stands for Compression Ratio.

Vārta	Dataset Size	Vocab Size	Domain Count
Total	41.8M	121.4M	1773

Table 3: Other statistics on Vārta.

- Ratio of novel n-grams: high degree of abstraction, low lexical overlap
- Headline size: extreme summarization
- CR: concise headlines
- Vocabulary size, domain count: diversity
- Dataset size: large-scale

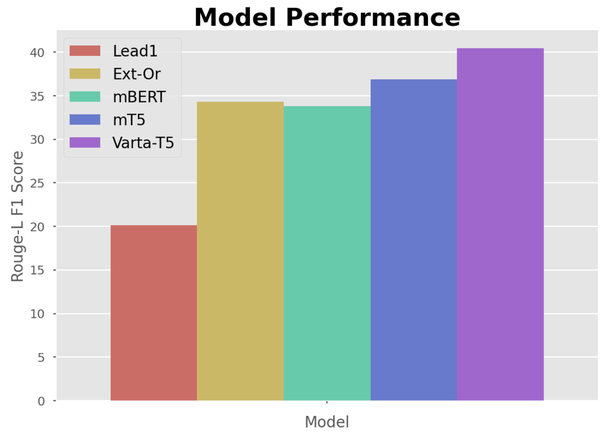
Dataset Example

Headline: IPL 2021 Auction: These players may not be bought by any team!

Article: The players' auction will be held in Chennai on February 18, before the Indian Premier League's 14th season is held. [...] The team has decided to release Jadhav, who has long been associated with the Chennai team. It seems unlikely that the 35-year-old player will be interested in this auction.

Experiments & Results

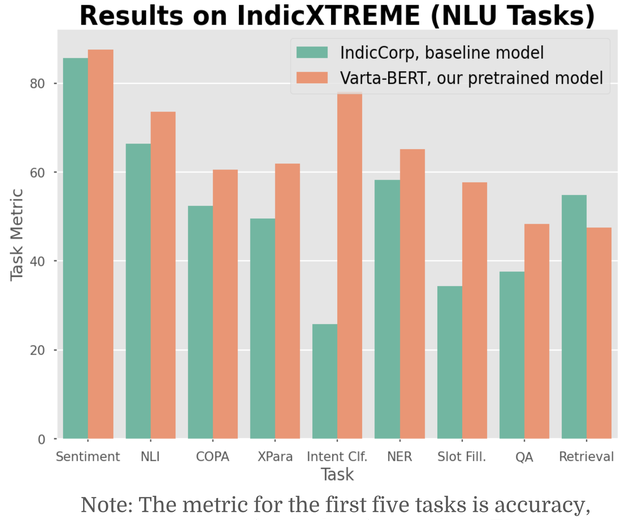
1. Headline Generation on Vārta



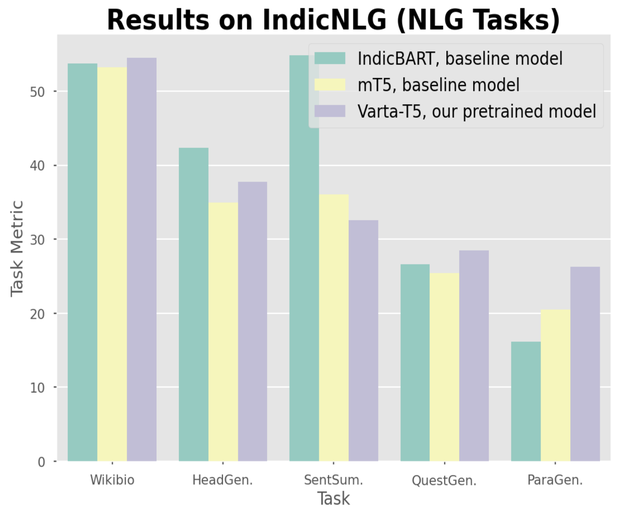
- Multilingual training: positive transfer among related languages
- Vārta: challenging even for state-of-the-art text generation models.

2. Vārta as a Pretraining Corpus

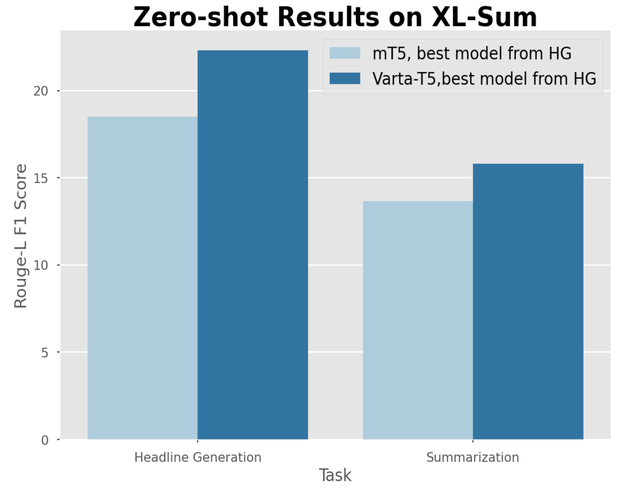
- Vārta-BERT and Vārta-T5: strong baselines on both NLU and NLG benchmarks
- Vārta: a pretraining corpus to train strong NLU and NLG models.



Note: The metric for the first five tasks is accuracy, while the metric for the last four tasks is F1 score.



Note: The metric for the first four tasks is Rouge-L F1 score while the metric for the last task is BLEU score.



References

1. Tahmid Hasan et al. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.
2. Sumanth Doddapaneni et al. 2022. Indicxtreme: A multi-task benchmark for evaluating indic languages.
3. Aman Kumar et al. 2022a. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages.