PMIndiaSum: Multilingual and Cross-lingual Headline Summarization for Languages in India

Ashok Urlana[1,*] Pinzhen Chen[2,*] Zheng Zhao[2],
Shay B. Cohen[2] Manish Shrivastava[1] Barry Haddow[2]

[1]IIIT-Hyderabad [2]University of Edinburgh [*]Equal contribution

ashok.urlana@research.iiit.ac.in, pinzhen.chen@ed.ac.uk, zheng.zhao@ed.ac.uk

Main Contributions

A **multilingual** and **massively parallel** summarization data for languages in India:

1. **14 Languages, 4 Families:** Dravidian, Indo-Aryan, Indo-European, Tibeto-Burman
2. **196 summarization directions** for monolingual, cross-lingual, and multilingual
3. **Open-source** at `hf.co/PMIndiaData` under CC BY 4.0

Benchmark experiments show that:

1. Both IndicBART and mBART are great for monolingual and cross-lingual
2. mBART is better for multilingual
3. More target-side data = better multilingual performance
4. IndicBART and mBART work better for Indo-Aryan languages than others.

Preparation

PMIndiaSum document-summary pairs are sourced from the Prime Minister of India website [1], where many articles are available in multiple languages. We used the raw data from the PMIndia parallel dataset [2] and crawled more. Figure 1 shows the data origins.

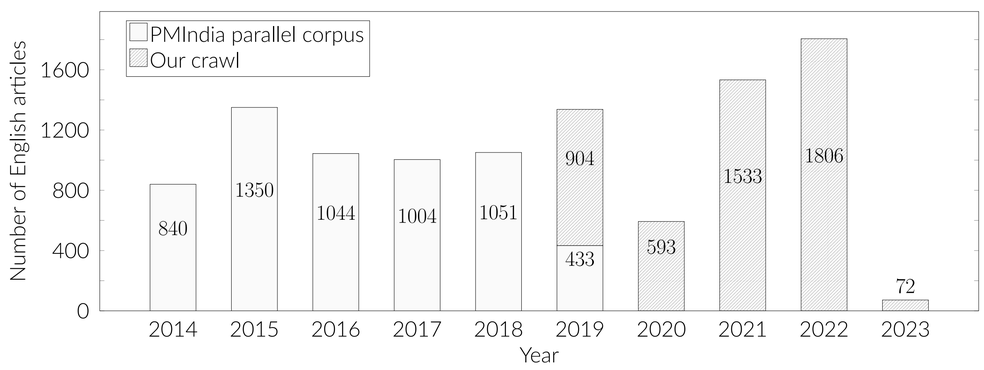

Figure 1. PMIndiaSum acquisition statistics for English, where 56% are from PMIndia parallel corpus and 44% are newly crawled by us.

We cleaned data pairs using **rule-based processing:**

1. **Language mismatch:** text not in desired Unicode range
2. **Duplicates or empty**
3. **Prefix:** a document starts with the summary
4. **Length:** document < 2 sentences, or summary < 10 tokens

PMIndiaSum Inspection

Token-based statistics:

- **Vocabulary size**: 2,000 to 8,000 for each language
- **Average Length:** document = 27 sentence or 518 tokens, summary = 12 tokens
- **Density:** low overlap between a document and the summary
- **Novelty:** unique uni/bi-grams in summaries > 90%
- **Redundancy:** low information repetition in summaries

Multilingualism and parallelism:

- Raw articles are written in multiple languages as shown in Figure 2.
- High cross-lingual LaBSE scores, for the same article but different languages, between summaries = 0.86 and between documents = 0.88

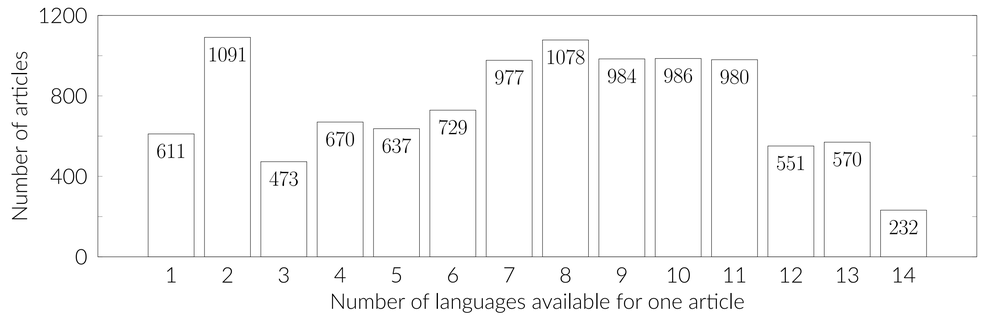

Figure 2. Degree of article parallelism in PMIndiaSum.

Benchmark Experiments

	monolingual	cross-lingual	multilingual
Extractive: lead	✓		
Extractive: oracle	✓		
Summarize-then-translate		✓	
Translate-then-summarize		✓	
Fine-tuning: full	✓	✓	✓
Fine-tuning: zero-shot		✓	
LLM prompting	✓		

Table 1. Techniques we benchmarked on our proposed PMIndiaSum.

Only selected language directions are shown due to limited space. Please check our paper for more results.

Monolingual:

1. Extractive oracle > lead, implying abstractive summaries
2. mBART > IndicBART, but mBART supports fewer languages
3. **Room for work**: LLM prompting quality is subpar

	Lead			Oracle			IndicBART			mBART		
	R-2	R-L	BLEU	R-2	R-L	BLEU	R-2	R-L	BLEU	R-2	R-L	BLEU
hi	44.8	58.2	30.1	49.1	62.5	33.2	53.1	66.9	46.0	**55.9**	**69.4**	**48.6**
ml	23.8	37.1	12.7	**32.7**	45.8	**15.4**	30.3	**47.5**	15.3	30.2	47.4	14.6
mni	38.3	50.5	26.4	**42.0**	54.2	26.0	38.7	53.0	32.0	41.3	**56.4**	**35.0**
te	31.2	41.0	18.0	**34.4**	**45.2**	**19.5**	16.3	32.7	9.9	16.0	33.4	9.8

Table 2. Monolingual: separate models for each language.

Crosslingual:

1. Summarize-then-translate > fine-tuning with a substantial gap
2. **Room for work**: end-to-end cross-lingual summarization

	Summarize-then-translate						Fine-tuning					
	IndicBART			mBART			IndicBART			mBART		
	R-2	R-L	BLEU	R-2	R-L	BLEU	R-2	R-L	BLEU	R-2	R-L	BLEU
ml-mni	**14.9**	**24.7**	**12.5**	13.5	23.7	9.6	8.2	18.9	4.8	7.6	18.9	2.5
mr-bn	13.9	32.2	8.7	**14.1**	**33.0**	**9.5**	12.9	30.4	7.2	12.8	31.5	7.0
en-mni	**24.0**	**35.7**	**18.2**	23.7	35.3	18.0	7.8	18.4	3.6	5.9	14.9	1.5

Table 3. Cross-lingual: separate models for each language direction.

Mulitlingual:

1. IndicBART produces sensible results for a few directions
2. mBART performs remarkably better
3. **Room for work**: multilingual still < monolingual for both PLMs

Human Annotations on Model Errors

Different models and different language directions lead to different errors

1. Monolingual models suffer from omission and redundancy
2. Cross-lingual models suffer from factual mistakes
3. Multilingual IndicBART suffers from language mismatch
4. **Room for work**: only 53% monolingual and 30% cross-lingual are correct

	hi-hi				**en-hi**			
	Monolingual		Multilingual		Cross-lingual		Multilingual	
	IndicBART	mBART	IndicBART	mBART	IndicBART	mBART	IndicBART	mBART
Comprehensibility	0	0	9	0	1	1	**39**	0
Grammar & Fluency	2	1	0	1	0	1	0	1
Factuality	4	4	3	3	**22**	**26**	3	7
Omission	**23**	**22**	12	19	11	12	1	15
Redundancy	**13**	**12**	**11**	**11**	9	5	1	3
No error	16	20	17	24	13	10	6	20

Table 4. Error analysis on different models and different language scenarios.

References

[1] https://www.pmindia.gov.in/en/

[2] PMIndia—A Collection of Parallel Corpora of Languages of India. Barry Haddow and Faheem Kirefu. 2020. arXiv preprint.

Acknowledgement

This work received funding from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant numbers 10052546 and 10039436] and UKRI Centre for Doctoral Training in Natural Language Processing (EP/S022481/1).

EMNLP 2023 · International Institute of Information Technology Hyderabad · The University of Edinburgh informatics · UK Research and Innovation