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

Ashok Urlana1,*  Pinzhen Chen2,*  Zheng Zhao2, Barry Haddow2
1IIIT-Hyderabad  2University of Edinburgh
*Equal contribution
ashok.urlna@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
<table>
<thead>
<tr>
<th></th>
<th>monolingual</th>
<th>cross-lingual</th>
<th>multilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extractive</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Extractive oracle</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Summarize-then-translate</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Translate-then-summarize</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Fine-tuning: full</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Fine-tuning: zero-shot</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>LLM prompting</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

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

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

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

Multilingual:
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

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

References

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).

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.

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

Table 2. Monolingual: separate models for each language direction.

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

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

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

Figure 2. Degree of article parallelism in PMIndiaSum.