MILDSum: A Novel Benchmark Dataset for Multilingual Summarization of Indian Legal Case Judgments

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, Saptarshi Ghosh
Indian Institute of Technology Kharagpur, India

Motivation

- Legal case judgments are lengthy, unstructured, and intricate - summarization is a practically important task.
- Most judgments in India are in complex English, whereas a **large fraction of India's population lacks a strong command in English**.
- There is a pressing need to provide **summaries of legal case judgments in Indian languages, to ensure equal access to justice**.

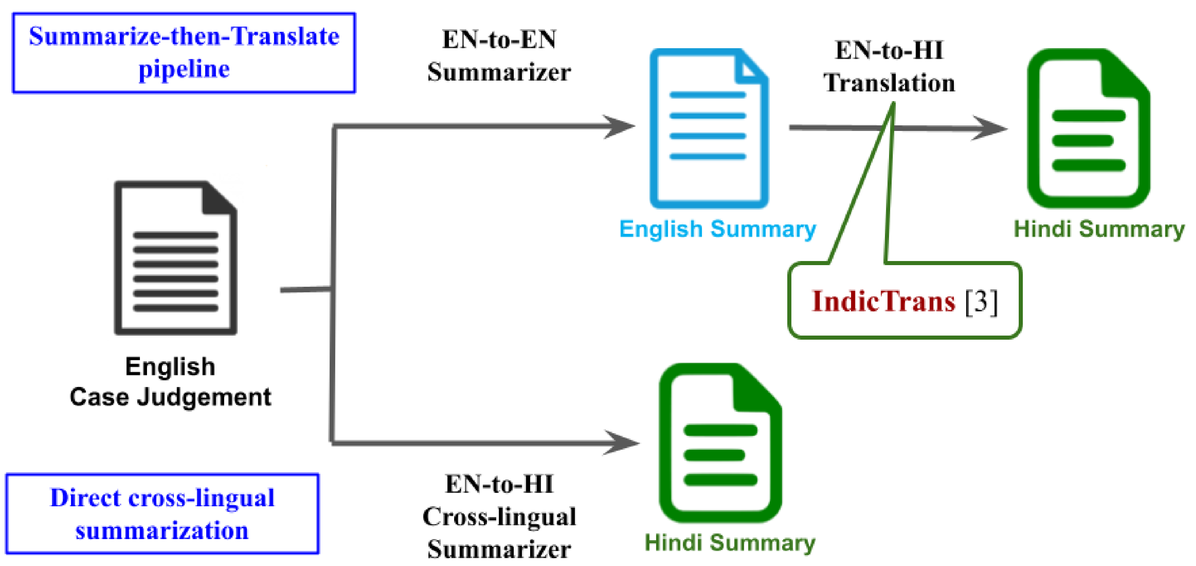
Our Contributions

- **MILDSum** (**M**ultilingual **I**ndian **L**egal **D**ocument **Sum**marization), the first dataset for cross-lingual summarization in the Indian legal domain.
- We benchmark the performance of a wide range of summarizers over MILDSum, considering two broad approaches – **Summarize-then-Translate pipeline**, and **Direct Cross-lingual Summarization**.

Key Dataset Properties

- **3,122 case judgments** from multiple High Courts and the Supreme Court of India in English and **abstractive summaries in both English & Hindi**.
- Target summaries collected from a reputed website **LiveLaw** (https://www.livelaw.in/), which publishes concise articles as a summary of judgments in both English and Hindi.
- Summaries are a mixture of **verbatim quotes** from the source judgments and simplified **abstractive paragraphs** drafted by Law practitioners.
- ***Coverage***: **0.90**, ***Density***: **24.42**, ***Compression Ratio***: **1:6**
- Train-Val-Test split ratio: **70:15:15**

Two approaches for cross-lingual summarization



Example Data Point (EN judgment, EN article, HI article)

CRM(NDPS) 1779 OF 2023

In Re: An application for Bail under Section 439 of the Code of Criminal Procedure, 1973 filed in connection with English Bazar Police Station Case No. 546 of 2023 dated 07.04.2023 under Sections 21(c)/25/27A/29 of the NDPS Act, 1985 pending in the Court of the learned Additional District & Sessions Judge, 3rd Court, Special Court at Malda being NDPS Case No. 51 of 2023.

And

In the matter of: Sanjay Kashyap & Anr.

. .. . . . .Petitioners.

Mr. Purbayan Chakraborty
Mr. Swastika Chowdhury
for the Petitioners.

Mr. Swapan Banerjee
Mr. Suman De
for the State.

1. Learned counsel for the petitioners argues that there are several contraventions of the NDPS Act and as such, the rigour of Section 37 of the NDPS Act is not applicable to the present case.

Calcutta High Court Grants Bail To NDPS Accused, Says No Reason To Obligate Them To Rigors Of S.37 In Absence Of Forensic Report

A Calcutta High Court vacation bench comprising **Justices Sabyasachi Bhattacharya and Partha Sarathi Chatterjee** have granted bail to the petitioners who had been accused under the NDPS Act, upon noting that the rigors of Section 37 of the NDPS Act which requires a Court to only grant bail if there are reasonable grounds to believe the accused are not guilty and will not reoffend, would not apply to the present case.

कलकत्ता हाईकोर्ट ने एनडीपीएस अभियुक्तों को जमानत दी, कहा कि फॉरेंसिक रिपोर्ट के अभाव में उन्हें धारा **37** की कठोरता के लिए बाध्य करने का कोई कारण नहीं

कलकत्ता हाईकोर्ट की एक अवकाश पीठ ने, जिसमें **जस्टिस सब्यसाची भट्टाचार्य और जस्टिस पार्थ सारथी चटर्जी** शामिल थे, एनडीपीएस अधिनियम में आरोपी याचिकाकर्ताओं को यह देखते हुए जमानत दे दी कि एनडीपीएस अधिनियम की धारा 37 की कठोरता कि अदालत केवल तभी जमानत दे सकती है, जब यह मानने के उचित आधार हो कि अभियुक्त दोषी नहीं है और दोबारा अपराध नहीं करेगा, यह वर्तमान मामले पर लागू नहीं होगा।

Summarization Results

Model	English Summary		Hindi Summary	
	ROUGE-L	BERTScore	ROUGE-L	BERTScore
Pipeline approach (Translation via IndicTrans)				
SummaRuNNer [1]	**30.34**	84.13	**24.55**	74.30
LexRank	29.24	83.97	23.73	74.36
Legal-Pegasus	28.36	**84.14**	22.58	74.62
LongT5	23.47	83.90	20.34	**75.70**
Cross-lingual summarization (without translation)				
CrossSum [2]	–	–	15.86	70.55
CrossSum-finetuned	–	–	20.68	75.05

Key Findings

- Translation introduces errors; **need better Machine Translation models in the legal domain.**
- CrossSum-finetuned largely outperforms off-the-shelf CrossSum; **demonstrates the utility of our MILDSum corpus in enhancing cross-lingual summarization.**
- Pipeline approach performs better than direct cross-lingual summarization despite the errors in the Translation stage; **need for better cross-lingual summarizers in the legal domain.**

Paper

Contact Information

References

[1] Nallapati et al.; Summarunner: A recurrent neural network based sequence model for extractive summarization of documents; AAAI 2017
[2] Bhattacharjee et al.; CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs; ACL 2023
[3] Ramesh et al.; Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages; TACL 2022