Unsupervised Abstractive Summarization of Bengali Text Documents

Radia Rayan Chowdhury*, Mir Tafseer Nayeem*,
Tahsin Tasnim Mim, Md. Saifur Rahman Chowdhury, Taufiqul Jannat
Ahsanullah University of Science & Technology
radiarayan.rrc@gmail.com, mir.nayeem@alumni.uleth.ca,
tahsintasnimmim@gmail.com, saif.chowdhury1997@gmail.com, taufiquljannat@gmail.com
*Equal contribution, listed by alphebetical order

EACL 2021

Text Summarizer

- Compression of large document
- Represents the most important or relevant information within the original content

Research Goal

- Unsupervised Text Summarizer for single document of low-resource language Bengali: 7th most spoken language in the world with 250 million native speakers

Our Contributions

- **BenSumm Model** : This model is the very first unsupervised model to generate abstractive summary from Bengali text documents
- **Dataset**: introduce a highly abstractive dataset with document-summary pairs which is written by professional summary writers of National Curriculum and Text-book Board (NCTB)
- Performs hierarchical clustering and calculate Cosine Similarity using ULMFit pre-trained language model
- Performs Sentence fusion on Bengali texts
- Bengali Document Summarization Tool

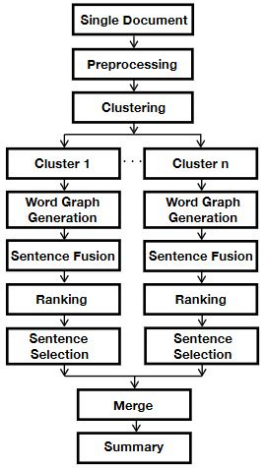
Why Unsupervised?

- Effective and Domain Independent
- No Need to Train Data
- Bengali: Low Resource Language

Why Document Clustering?

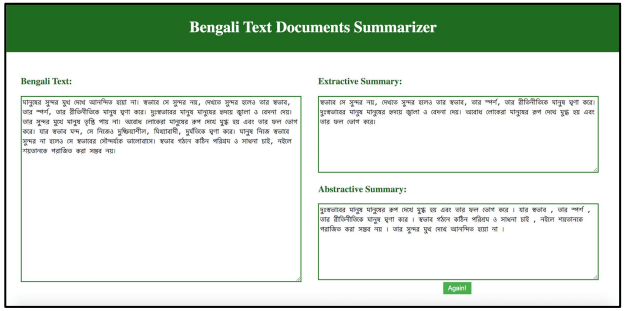
- Avoid incoherent summary and redundancy
- Ensure good coverage

Methodology (BenSumm Model)



Our **BenSumm** Unsupervised Model

Our Bengali Abstractive Summarizer Tool



Text Preprocessing

- Tokenization
- Removing Punctuation
- Removing Stopwords
- POS Tagging

Results (Human Evaluation)

Average Score in Scale (1-5):
- Content : 4.41
- Readability : 3.95
- Overall quality : 4.2

Here, 1= Poor, 5=Good

Dataset

- **Abstractive Dataset**: Created a set of 139 samples of human-written abstractive document-summary pairs written by professional summary writers of the National Curriculum and Textbook Board (NCTB)
- **Extractive Dataset**: Experiment with an Extractive Dataset Bangla Natural Language Processing Community (BNLPC)

	NCTB	BNLPC
Total #Sample	139	200
Source Length (Avg)	91.33	150.75
Human Reference Length (Avg)	36.23	67.06
Summary Copy Rate	27%	99%

Results (Automatic Evaluation)

NCTB (Abstractive)	R-1	R-2	R-L
Random Baseline	9.43	1.45	9.08
GreedyKL	10.01	1.84	9.46
LexRank	10.65	1.78	10.04
TextRank	10.69	1.62	9.98
SumBasic	10.57	1.85	10.09
BenSumm[Abs] (ours)	**12.17**	**1.92**	**11.35**

BNLPC(Extractive)	R-1	R-2	R-L
Random Baseline	35.57	28.56	35.04
GreedyKL	48.85	43.80	48.55
LexRank	45.73	39.37	45.17
TextRank	60.81	**56.46**	60.58
SumBasic	35.51	26.58	34.72
BenSumm[Abs] (ours)	**61.62**	55.97	**61.09**

Rough-1 , Rough-2 and Rough-L Score on our NCTB Dataset and BNLPC Dataset

Future Work

Increasing document-summary pair dataset, Implementing multi-sentence compression and paraphrasing