The 17th Conference of the European Chapter of the Association for Computational Linguistics

Shironaam: Bengali News Headline Generation using Auxiliary Information

Abu Ubaida Akash[1]*, Mir Tafseer Nayeem[2]*◊, Faisal Tareque Shohan[1], Tanvir Islam[3]

[1]Ahsanullah University of Science and Technology, [2]University of Alberta, [3]University of Hawaii at Manoa
{akash.ubaida@gmail.com, mnayeem@ualberta.ca, faisaltareque@hotmail.com, tislam@hawaii.edu}
*[equal contribution], ◊[supported by Huawei Doctoral Fellowship]

Introduction

Role of news headlines:

- Catching the reader's attention
- Providing Context
- Enhancing Search Engine Optimization (SEO)

A special case of abstractive summarization:

- Does not often maintain grammatical structure
- More extreme than extreme summarization
- Highly abstractive

Research Goal

Generating quality news headlines and establishing a category-robust benchmark in a low-resource language **Bengali** (7th most spoken language in the world with approximately 300 million native speakers) by incorporating auxiliary information

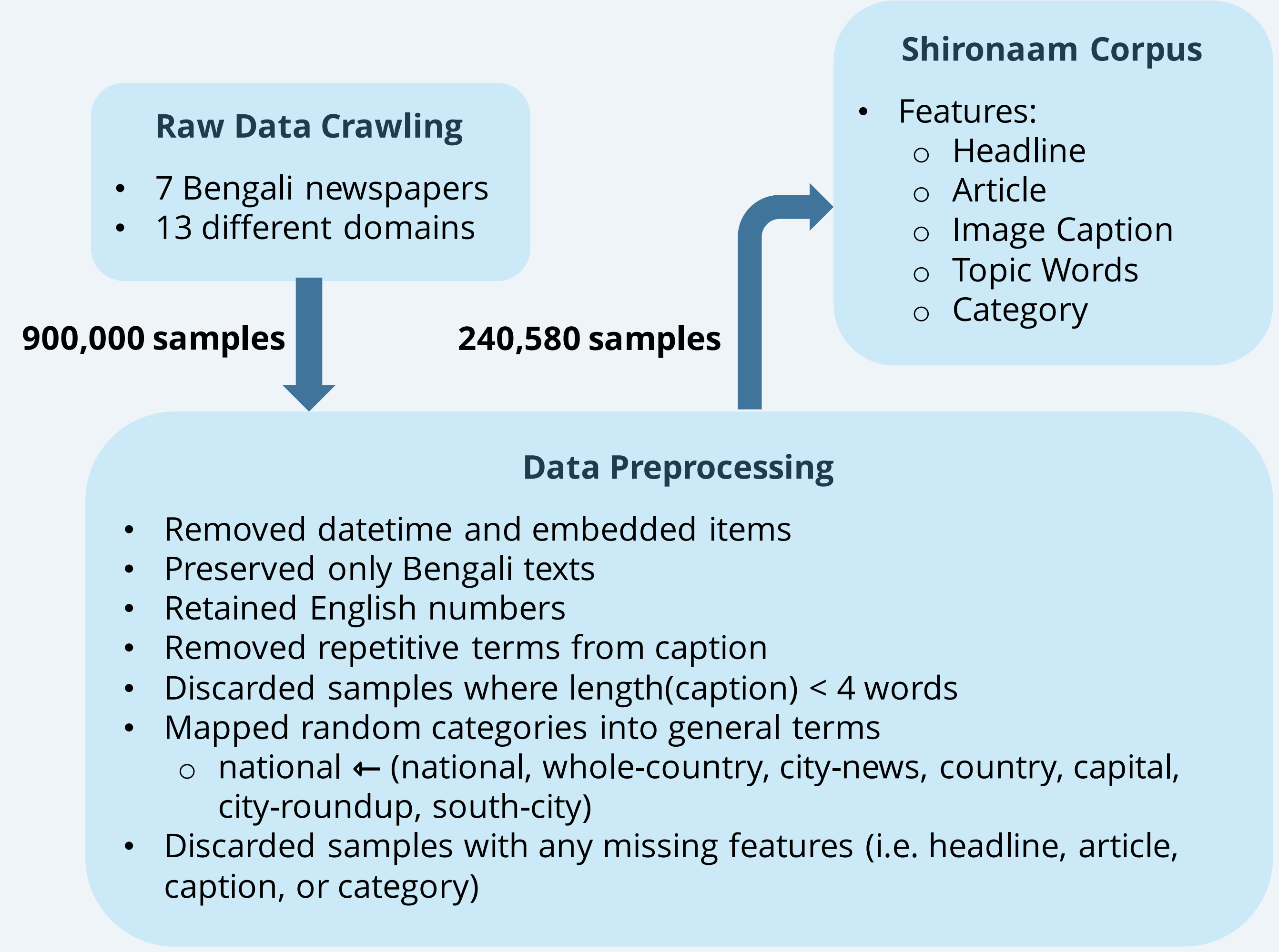
Why Our Work

- **Typically one-to-one mapping:**
  - Input is an article, output is an headline
- **Makes it difficult when the input is necessarily long:**
  - Contextualized language models suffer from a limited sequence
- **More challenging for low-resource languages:**
  - Unavailability of large-scale human-annotated dataset
  - Limited language models
  - Lack of SOTA models for the downstream task

Our Contributions

- **Provided Shironaam, a large-scale news headline generation dataset:**
  - Largest for a low-resource language i.e. Bengali
  - Contains auxiliary information along with article-headline pairs
- **Presented the concept of incorporating auxiliary information in headline generation:**
  - Developed an end-to-end SOTA model for headline generation
- **Developed BenSim, a module for measuring semantic similarity among Bengali sentences:**
  - Helps to encode long documents
- **Illustrated the utility and robustness by evaluating the performance with few-shot settings**

Dataset



Raw Data Crawling

- 7 Bengali newspapers
- 13 different domains

900,000 samples

Data Preprocessing

- Removed datetime and embedded items
- Preserved only Bengali texts
- Retained English numbers
- Removed repetitive terms from caption
- Discarded samples where length(caption) < 4 words
- Mapped random categories into general terms
  - national ← (national, whole-country, city-news, country, capital, city-roundup, south-city)
- Discarded samples with any missing features (i.e. headline, article, caption, or category)

240,580 samples

Shironaam Corpus

- Features:
  - Headline
  - Article
  - Image Caption
  - Topic Words
  - Category

Category	Total	Jaccard (%)	Category	Total	Jaccard (%)
Entertainment	17,565	13.56	Life-Health	6,933	17.83
National	128,226	24.60	Opinion	3,819	38.41
Nature	510	23.66	Politics	16,380	23.02
International	33,329	18.09	Edu-Career	4,372	53.58
Sports	19,235	17.82	Science-Tech	1,141	22.95
Economy	7,032	39.37	Religion	294	71.59
Life-Health	6,933	17.83	**Total/Average**	**240,580**	**28.94**

- **Train**, **Test**, and **Validation** set: Ratio of (92% - 220,574), (6% - 15,012), (2% - 4994) samples from all categories
- ***Jaccard scores*** represent the similarities of each domain in between the image captions and headlines

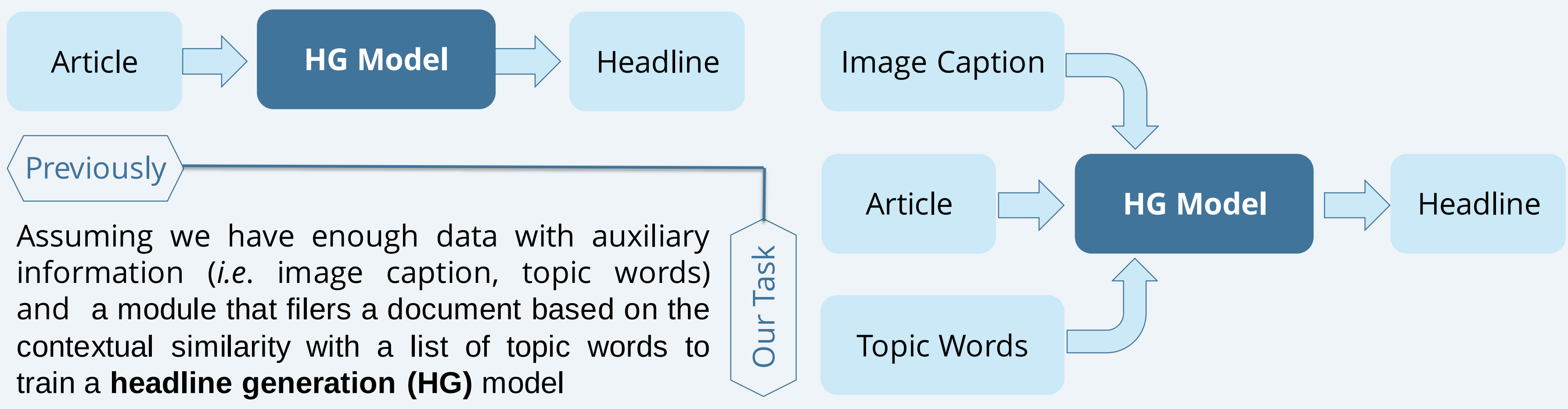
Dataset	% of novel n-gram			
	unigram	bigram	trigram	4-gram
Indic-BN[7]	26.59	66.12	82.71	86.49
Shironaam	46.38	78.92	90.39	94.77

Features	IndicNLG-BN	Shironaam
Article	Yes	Yes
Headline	Yes	Yes
Image Caption	No	Yes
Category	No	Yes
Topic Word	No	Yes
#Samples	142,731	240,580

Shironaam comprises a diverse range of headline styles and provides the largest collection of Bengali news articles that can also be used in document categorization, news clustering, keyword identification *etc.*

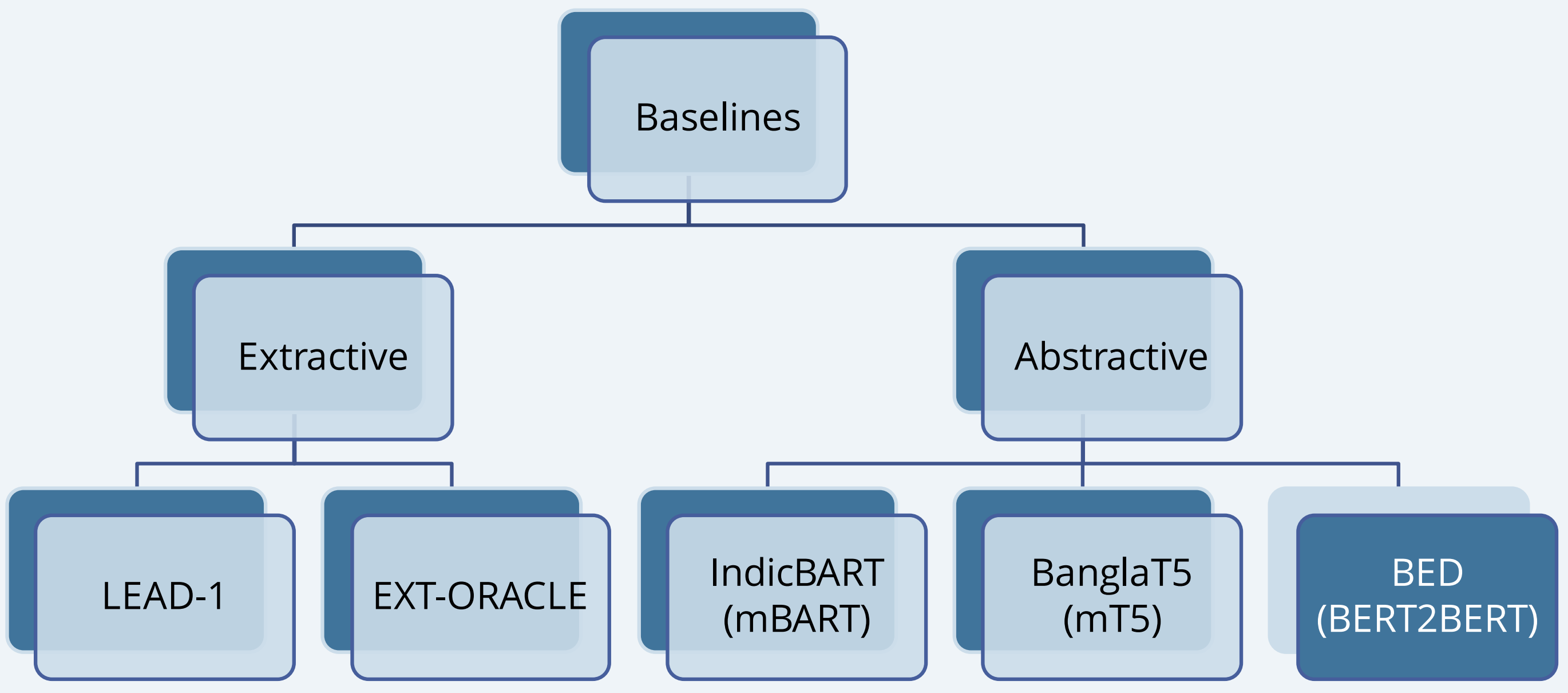
Dataset	Article	Headline	Image Caption	Topic Words
Average number of words				
Shironaam	252.01	6.53	6.80	3.21
Indic-BN[7]	199.83	10.03	-	-
Average number of sentences				
Shironaam	20.05	1.0	1.04	-
Indic-BN[7]	15.19	1.19	-	-
Vocabulary size				
Shironaam	605,750	76,732	87,644	-
Indic-BN[7]	614,374	65,553	-	-

Task

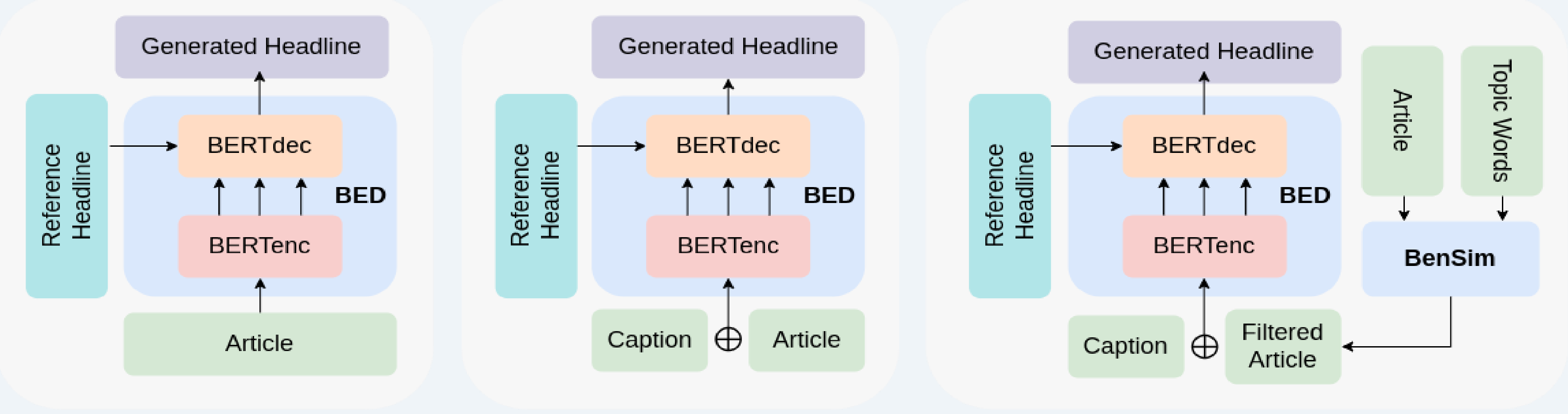


Assuming we have enough data with auxiliary information (*i.e.* image caption, topic words) and a module that filers a document based on the contextual similarity with a list of topic words to train a **headline generation (HG)** model

Approach



Proposed Models



(a) BED(base)

(b) BED(w/ Article + Caption)

(c) BED(w/ FilteredArticle + Caption)

BERT based Encoder Decoder (BED)

(a) Article Only

- Input: Article; Output: Headline
- Both encoder and decoder weights initialization with pre-trained BERT checkpoint (e.g. BanglaBERT)
- Only the cross attention weights randomly initialized
- Hugging Face encoder-decoder paradigm
- First SOTA baseline in Bengali language

(b) Article and Caption

- Input: Article, Image caption; Output: Headline
- Parallel fusion mechanism
- Separated by a special token

(c) Filtered Article and Caption

- Input: Article, Image caption, Topic words; Output: Headline
- Parallel fusion mechanism
- Separated by a special token
- Additionally BenSim#

(#) BenSim Module

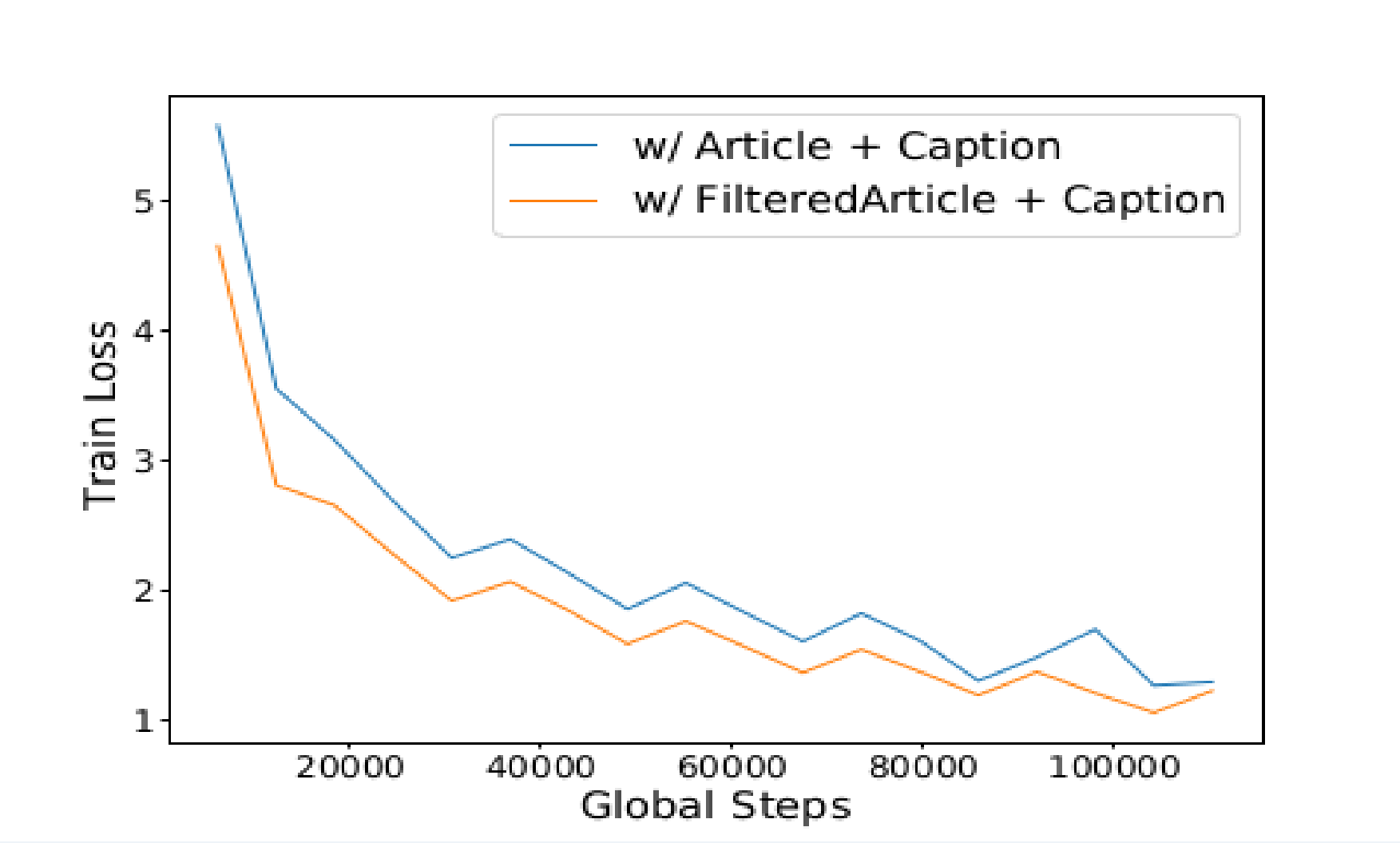
- Input: Article, Topic words; Output: Filtered article
- Measures semantic similarity between Bengali sentences utilizing bangla-bert-base embeddings
- Picks most relevant sentences from long articles (we consider top 40)
- Mean pool operation followed by Cosine similarity

Experiments

RQ#1: Can we use auxiliary information (e.g., image caption and topic words) to improve the performance of the headline generation?

RQ#2: Which domain(s) benefit from the auxiliary information in few-shot and non few-shot settings?

Models		Rouge			BLEU			BERT Score	METEOR Score
		R-1	R-2	R-L	BLEU Score	Brevity Penalty	Length Ratio		
Baselines	LEAD-1	30.50	13.86	28.00	5.65	97.71	2.48	74.63	29.90
	EXT-ORACLE	39.92	22.89	37.28	9.17	97.16	2.30	77.16	39.65
	IndicBART	28.76	12.65	27.11	15.03	99.91	1.14	74.95	20.39
	BanglaT5	44.13	23.03	42.12	13.05	91.33	1.15	80.13	34.65
Our Ablations	BED Base	44.22	24.18	42.28	22.06	94.47	0.94	80.53	34.16
	BED (Article+Caption)	51.62	33.62	49.94	31.39	96.02	0.96	82.93	42.57
	BED (FilteredArticle+Caption)	**52.19**	**34.27**	**50.31**	**31.80**	**98.57**	**0.99**	**83.10**	**43.52**



- Few lengthier articles in **Shironaam** corpus
- Slightly better performance
- Learns faster with the filtered articles
- Performance will increase with the number of longer articles

Domain Specific Analysis

- Compared our **BED (w/ FilteredArticle+Caption)** model with two baselines: BED (base), BanglaT5
- Domains in two folds: Few shot (<6500 samples), Non-Few shot (>6500 samples)
- Proposed model improves the baseline scores by satisfactory margin for all domains from both settings except *Entertainment*, and *Miscellaneous* category
- Highest improvement: *Politics* (from non-few-shot), *Religion* (from few shot)
- *Entertainment* domain are casual and clickbait-style and do not maintain the identical nature of a particular domain.
- *Miscellaneous* domain is comprised of different sorts of randomness from various domains

Future Work

Utilization of multimodal information, Human evaluation on generated samples, Development of language agnostic model