XL-HeadTags: Leveraging Multimodal Retrieval Augmentation for the Multilingual Generation of News Headlines and Tags

Faisal Tareque Shohan[1*], Mir Tafseer Nayeem[2*], Samsul Islam[1], Abu Ubaida Akash[3], Shafiq Joty[4,5]

[1]Ahsanullah University of Science and Technology
[2]University of Alberta
[3]Université de Sherbrooke
[4]Salesforce Research
[5]Nanyang Technological University

*[Equal Contribution]

News Headline and Tags

❑ Headline

- ❖ Providing brief context
- ❖ Catching readers attention
- ❖ Enhancing Search engine optimization

❑ Tags

- ❖ Semantic Markers
- ❖ Dynamically connects related articles
- ❖ Provide navigational aids

Headline and Tags Generation

❑ Headline Generation

- ❖ Special case of abstractive Summarization
- ❖ Do no often maintain grammatical structure
- ❖ Need to be brief and engaging
- ❖ Highly abstractive in nature

❑ Tags Generation

- ❖ Similar to key-phrase generation
- ❖ Focuses on broader overview
- ❖ Are often absent in the article
- ❖ Necessary for connecting to related article

Headline and Tags Generation

❑ Motivation

- ❖ Headline and Tags are **extreme compression** of the article
- ❖ Generating headline and tags in a **multilingual context**
- ❖ News article tags generation in **unexplored** in existing literature
- ❖ Simultaneous headline and tags generation are not often modeled together
- ❖ **Limited context window** of pretrained models hinder NLG task performance on long documents, leading to subpar results

XL-HeadTags Task

❑ What is XL-HeadTags Task?

- ❖ Simultaneously generate **Headline** and **Tags** in a **unified** learning framework
- ❖ Generate both **controlled** and **unrestricted** number of tags

XL-HeadTags Task

❑ Research Questions?

- ❖ Can the task of **simultaneous generation** of headline and task be modelled and learned?
- ❖ Can **improved content selection strategies** mitigate the constraints imposed by **limited context window** of pre-trained language models?
- ❖ How can **multimodal auxiliary information** (e.g., images, captions) be utilized as query to **effectively retrieve** the most salient information from lengthy articles?

Our Contributions

❑ XL-HeadTags Task

- ❖ **Simultaneous generation** of both headline and tags through **instruction tuning**
- ❖ Both **controlled** and **unrestricted** tags generation through natural language instruction

❑ MultiRAGen

- ❖ Present new **content selection approach** utilizing **multimodal** auxiliary information

Our Contributions

❑ Multilingual Tools

Developed multilingual tools by accumulating open-source resources

- ❖ **Multilingual Rouge Scorer –** Leveraging multilingual BPE tokenizer
- ❖ **Multilingual Sentence Tokenizer –** Covering 41 Languages
- ❖ **Multilingual Stemmer –** Supports 18 Languages

Our Contributions

❑ Tags Evaluation Metrics

- ❖ Introduce **Tags evaluation metrics** to evaluate both
  - ✓ Controlled Tags generation
  - ✓ Unrestricted Tags generation

❑ XL-HeadTags Dataset

- ❖ News articles with **multimodal** (e.g., images, captions) auxiliary information
- ❖ Covering **20** languages across **6** diverse language families

Dataset

- **M3LS** and **XL-Sum** are primary data source
- Both share **BBC** news are source
- **Minimal** Distributional and Structural changes are expected

❑ **M3LS –** Multilingual Multimodal Summarization Dataset

- Contains Headline, Article, Summary, Images, Captions, Tags, News links
- Auxiliary information's (e.g., images, captions) were utilized for retrieval augmentation framework

Dataset

- **XL-Sum –** Multilingual Abstractive Summarization Dataset
  - Contains Headline, Article, Summary, News links
  - **Arabic**, **Turkish** and **Persian** news articles were selected
  - Images, Captions and Tags were **absent**
  - Missing information's were **crawled** utilizing provided **URL's**

Dataset

❑ Statistics

- Total 415k data samples
- Average article
  - Words – 902
  - Number of sentences – 27.7
  - **Tokens – 1632**

- Average headline to article compression ratio 98.88%
- Average 3.47 tags per Article, where **44.64%** tags are **absent** in article

* Most pre-trained language models have a context window of 512

MultiRAGen – Multimodal **R**etrieval **A**ugmented **Gen**eration

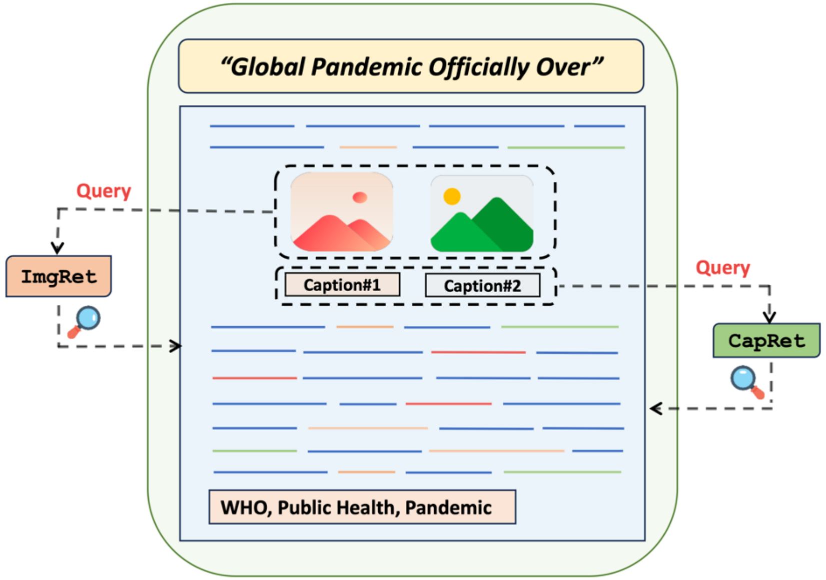
Our Approach **MultiRAGen** has two main component

- ✓ Multimodal Retrievers
- ✓ Instruction Tuning

MultiRAGen

❑ Multimodal Retrievers

- Tokenize article into sentences
- Use Images and Captions as queries to compute semantic similarity with sentences
    - Utilizing Multilingual CLIP-ViT-B32 that maps texts and images to a shared dense vector space
- Pick top-K sentences based of similarity scores
- Reorder top-K sentences to their original sequence in the article to preserve the narrative flow



MultiRAGen

- **Multimodal Retrievers –** Handling multiple Images and Captions
  - Each Image and Caption are treated as **distinct query** entity
  - Scores from each query are **aggregated**
  - **Greedy** approach is used to pick top K sentences

MultiRAGen

❑ Instruction Tuning

- ❖ Task specific prefixes to guide the model
- ❖ Two **instruction variations** are introduced

Determine optimal number of tags to generate

Instruction for Unrestricted Generation

`Input` → **Generate Headline and Tag Words:** *Selected Content*.

`Output` → **Headline is:** *Headline*. **Tag words are:** $T_1, T_2, \cdots, T_o$.

Instructed to generate specified number of tags

Instruction for Controlled Generation

`Input` → **Generate Headline and $\mathcal{N}$ Tag Words:** *Selected Content*.

MultiRAGen

❑ **Instruction Tuning –** Selected Content

❖ For the input in the instruction we utilize **three** settings

- **Original Article**
- **Top-K Retrieved Sentences**
- **Top-K Retrieved Sentences + Article**

Tags Evaluation

❑ Existing key-phrase evaluation metrics

- **Precision** ***(P)***, **Recall** ***(R)***, **F-measure** ***(F_1)*** are commonly used to measure predictive performance
- If $\overline{\gamma} = \{\overline{\gamma}_1, \overline{\gamma}_2, \ldots, \overline{\gamma}_m\}$ denotes **generated** key-phrases and $\boldsymbol{\gamma}$ denotes **ground truth** key-phrases

$$\boldsymbol{P} = \frac{|\overline{\boldsymbol{\gamma}} \cap \boldsymbol{\gamma}|}{|\overline{\boldsymbol{\gamma}}|}$$

$$\boldsymbol{R} = \frac{|\overline{\boldsymbol{\gamma}} \cap \boldsymbol{\gamma}|}{|\boldsymbol{\gamma}|}$$

$$\boldsymbol{F_1} = \frac{\boldsymbol{2 * P * R}}{\boldsymbol{P + R}}$$

Tags Evaluation

❑ Proposed Tags evaluation metrics

Inspired by the work of Yuan et al. (2020), we propose three metrics

Unrestricted Tags generation	Controlled Tags generation
❖ ***F_1@M,*** where ***M*** = | $\bar{\gamma}$ |. ***M*** varies with article ❖ Reflecting model's decision on the number of tags	❖ ***F_1@K,*** where ***K*** is user defined ❖ Here we defined ***K*** as **3** and **5** ❖ ***F_1@O,*** where ***O*** = | γ |. ❖ Number of tags in ground truth

Experiments

❑ Data

❖ Introduce Prefix Mixture Strategy

- ➢ Prefix mixture approach **during training** to improve the generalizability
- ➢ Enabling it to generate **both** controlled and unrestricted tags
- ➢ We maintain a **70:30** allocation ratio
- ➢ **70%** data for **controlled** tag word generation
- ➢ **30%** data for **unrestricted** tag words generation

❖ Data Split

- ➢ Split into train (95%), validation (1%) and test (%) sets for experiments

Experiments

❑ Models – Baselines

- ❖ We finetune following pre-trained models
  - o mT5-base
  - o mT0-base
  - o Flan-T5-large
- ❖ Selected Content is **Original Article**
- ❖ **LEAD-1** and **EXT-ORACLE** represents extractive baselines

Experiments

❑ Models – Baselines – LLM’s

- **Gemini-Pro** and **Mixtral** models for evaluating their efficacy in **XL-HeadTags** task
- **Zero-shot** prompting conditions
- Sampling **50** instances from each language

Experiments

❑ Models – MultiRAGen

- ❖ Two separate multimodal retrievers
  - **ImgRet** – Visual Retrievers (Images)
  - **CapRet** – Textual Retrievers (Captions)
- ❖ Number of sentences to be retrieved is determined by **K**
  - **5, 10** and **15** are explored as the value of **K**
- ❖ **Two** Selected Content approaches
  - Top-K Retrieved Sentences
  - Top-K Retrieved Sentences + Article

Results - Headline

		Selected Content	Models	Rouge-1	Rouge-2	Rouge-L	BLEU	Meteor	LR (↓)	BERT Score
Baselines		Article	mT5	37.86	17.20	33.53	12.95	25.55	0.84	75.79
			mT0	38.33	17.66	33.90	14.64	26.44	0.94	75.83
			Flan-T5	31.46	12.73	28.15	8.75	24.61	0.71	70.87
MultiRAGen	**Text (Caption)**	Top-K Retrieved + Article	mT5 *(K=10)*	39.04	18.20	34.51	14.03	26.86	0.87	76.23
			mT0 *(K=10)*	39.13	18.35	34.61	14.29	27.24	0.88	76.21
			Flan-T5 *(K=10)*	31.65	12.80	28.44	8.64	24.59	0.70	70.89
	Visual (Image)		mT5 *(K=10)*	38.94	18.17	34.44	14.08	26.87	0.87	76.18
			mT0 *(K=10)*	39.16	18.33	34.61	14.27	27.11	0.88	76.22
			Flan-T5 *(K=10)*	31.55	12.82	28.38	8.65	24.58	0.69	70.90

Results - Tags

		Selected Content	Models	Rouge-1	Rouge-2	Rouge-L	BLEU
Baselines		Article	mT5	45.01	39.82	44.67	46.79
			mT0	51.58	44.94	52.50	54.39
			Flan-T5	30.76	26.3	31.86	33.40
MultiRAGen	**Text (Caption)**	Top-K Retrieved	mT5 *(K=10)*	53.08	47.00	54.00	56.24
			mT0 *(K=10)*	53.88	47.95	55.29	57.49
			Flan-T5 *(K=10)*	31.18	26.65	32.16	33.77
	Visual (Image)		mT5 *(K=10)*	53.62	47.57	54.76	56.95
			mT0 *(K=10)*	53.79	47.69	55.00	57.12
			Flan-T5 *(K=10)*	30.74	26.25	31.40	33.21

Discussion

- Both Textual and Visual Retrieved Content Selections **help models outperform** their respective baselines
- Combining **retrieved sentences with article** is the superior strategy for headline
- While using **solely retrieved sentences** is more effective for tags generation
- The disparity indicates that
  - Tags, being concise, thrive on **focused inputs**
  - While headlines require **broader context**

Future Works

- Investigate the potential benefits of integrating both image and caption data for simultaneous retrieval process

Thank You!