The 62nd Annual Meeting of the Association for Computational Linguistics

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 & Technology, [2]University of Alberta, [3]Université de Sherbrooke, [4]Salesforce Research, [5]Nanyang Technological University

Headline & Tags

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

Why our work

- 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
- Improved **content selection** approach for overcoming **limited context window** of pretrained language models

Out Contributions

XL-HeadTags Task

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

MultiRAGen

- New **content selection approach** utilizing **multimodal** auxiliary information

Multilingual Tools

Multilingual Tools accumulating open-source resources

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

Tags Evaluation Metrics

Three Tags evaluation metrics

- **Controlled** Tags Generation
- **Unrestricted** Tags Generation

XL-HeadTags Dataset

- Contains Multimodal Auxiliary Info
- Covering 20 languages

Dataset

- M3LS and XL-Sum are primary data source, both share BBC news as source
- Minimal Distributional and Structural shifts are expected

M3LS

- Contains Headline, Article, Summary, Images, Captions, Tags, News links
- Auxiliary information utilized for retrieval

XL-Sum

- Arabic, Turkish, Persian articles selected
- Images, Captions and Tags missing
- Missing information's were crawled

Data Statistics

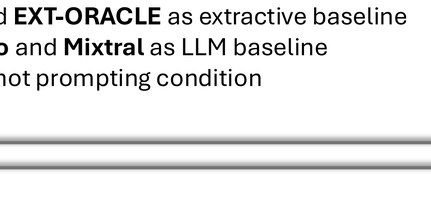
Samples	415117	**% of novel unigram**	33.60
Average # Words in Article	902	**% of novel bigram**	80.83
Average # Sentences in Article	27.7	**% of novel trigram**	94.37
Average # Tokens in Article	1632	**Average # Tags per Article**	3.47
Average # Words in Headline	10.13	**% of Tags present in Article**	44.64
Compression Ratio	98.88	**Average Image/Captions**	3.21

MultiRAGen – Multimodal Retrieval Augmented Generation

- MultiRAGen has two main component – **Multimodal Retrievers**, **Instruction tuning**

Multimodal Retrievers

- Tokenize article into sentences
- Compute semantic similarity between sentence and Image/Captions
  - Multilingual CLIP-ViT-B32 maps text and images to a shared vector space
- Pick top-K sentences based of similarity scores
- Reorder top-K sentences to their original sequence to preserve the narrative flow



Instruction Tuning

- Task specific prefixes
- Two instruction variation

Instruction for Unrestricted Generation

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

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

Instruction for Controlled Generation

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

Three selected content approach

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

Experiments

Data

Controlled 70 | 30 Unrestricted

- **Prefix Mixture strategy** during training
- Enabling both controlled and unrestricted tags generation

Model-Baselines

- Finetune following models with Original article
  - mT5-base
  - mT0-base
  - Flan-T5-large
- **LEAD-1** and **EXT-ORACLE** as extractive baseline
- **Gemini-Pro** and **Mixtral** as LLM baseline
  - Zero-Shot prompting condition

Model-MultiRAGen

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

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
MultiRAGen – Visual (Image)	Top-K Retrieved + Article	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

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
MultiRAGen – Visual (Image)	Top-K Retrieved	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

- 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 Work

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