Abstractive News Summarization for Vietnamese

Dec 25, 2024

OVERVIEW

Overview

What?

- Aims to condense news articles into concise, coherent summaries by rephrasing and synthesizing the main points.

Why?

- Helps avoid insignificant information and highlight knowledge have gone unnoticed

Xiang Jiang and Markus Dreyer. 2024. CCSum: A Large-Scale and High-Quality Dataset for Abstractive News Summarization. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7306–7336, Mexico City, Mexico. Association for Computational Linguistics.

Koh, Huan Yee, et al. "An empirical survey on long document summarization: Datasets, models, and metrics." ACM computing surveys 55.8 (2022): 1-35.

RELATED WORKS

RELATED WORKS

VLSP-2022

VLSP-2022:

- Pipeline:
  - Crawl data from websites
  - Cluster
  - Human annotator writes summary
  - Re-check data
  - Expert reviews

Tran, Mai-Vu, et al. "Overview of the VLSP 2022--Abmusu Shared Task: A Data Challenge for Vietnamese Abstractive Multi-document Summarization." *arXiv preprint arXiv:2311.15525* (2023).

RELATED WORKS

ViMs

ViMs:

- Pipeline:
  - Select important sentences
  - Remove redundant from those
  - Co-reference to preserve context
  - Sentence formulation-organization
  - Finalize summary
- Depends heavily on human annotation

Tran, NT., Nghiem, MQ., Nguyen, N.T.H. *et al.* ViMs: a high-quality Vietnamese dataset for abstractive multi-document summarization. *Lang Resources & Evaluation* **54**, 893–920 (2020). https://doi.org/10.1007/s10579-020-09495-4

DATASET CONSTRUCTION

STEP-BY-STEP

1. Get the News
2. Prepare before Filtering
3. Filter Design
4. Perform Bayesian Optimization
5. Finalize

Xiang Jiang and Markus Dreyer. 2024. CCSum: A Large-Scale and High-Quality Dataset for Abstractive News Summarization. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7306–7336, Mexico City, Mexico. Association for Computational Linguistics.

Raw data

BKAI

- 32M articles of Vietnamese news
- From 1970* to November 2023

Duc, Nguyen Quang, et al. "Towards Comprehensive Vietnamese Retrieval-Augmented Generation and Large Language Models." arXiv preprint arXiv:2403.01616 (2024).

Prepare before Filtering

- Remove first outlier entity, remove name entity in the end of some articles.
- We group the articles into 3-day windows, based on publication time.
- For each window, we use vi-sBERT to encode the main texts of the articles of this window, and then perform soft-clustering using Faiss

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547.

Filter Design

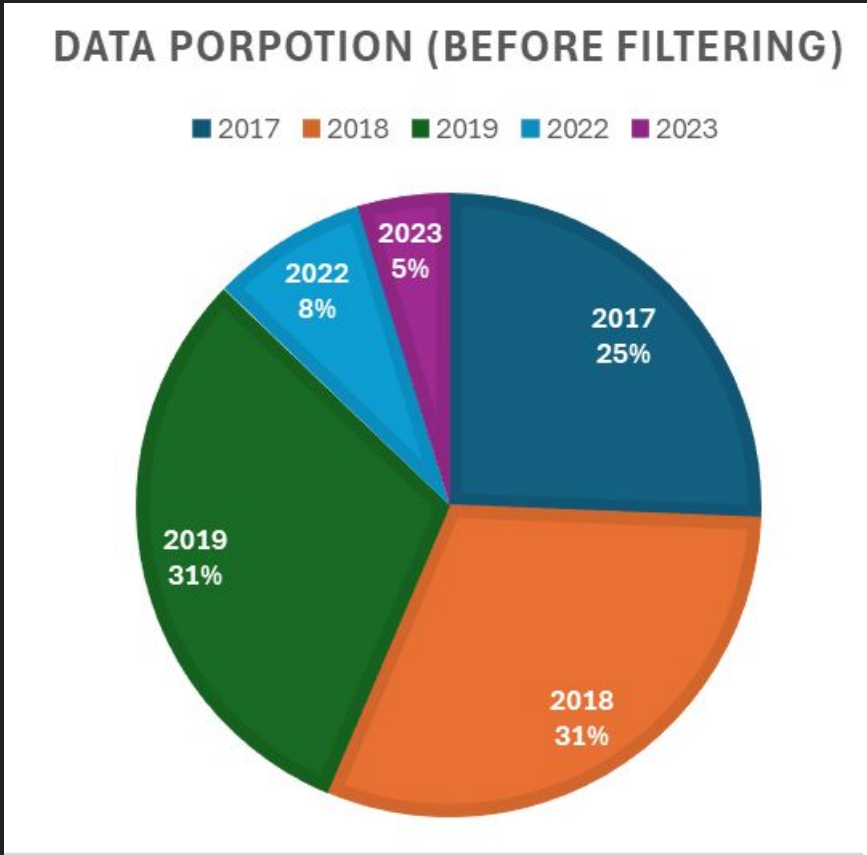
- Heuristics
  - Summary must have at least 1 entity, 25 or more words.
  - Summary must end in proper punctuations.
- Factual consistency:
  - Entity precision, BS-P, Quotation exact match.
- Coverage:
  - BS-R, Title-title similarity, Summary-title similarity.
- Abstractiveness:
  - MINT, Simhash.

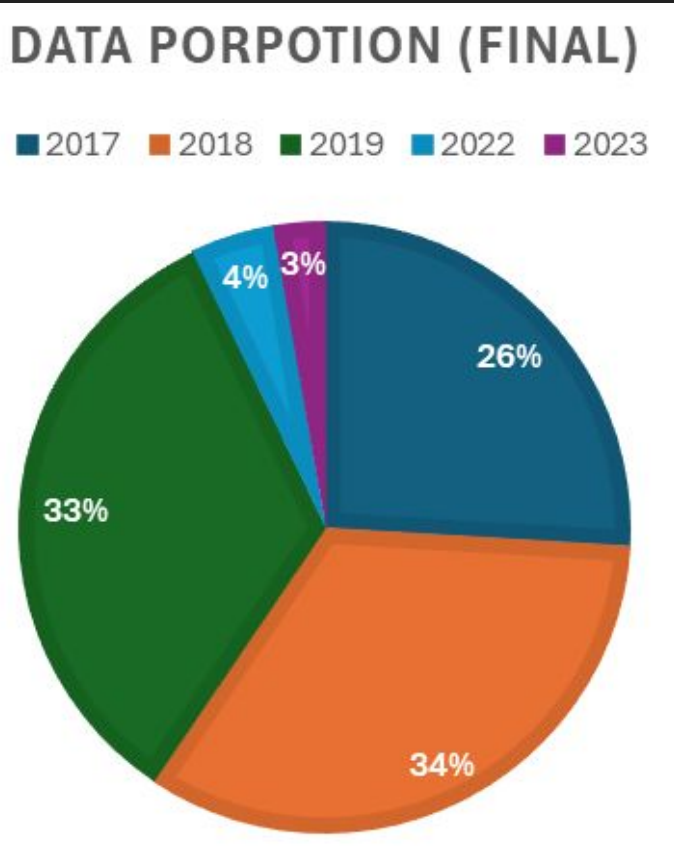
Bayesian Optimization

- Search space:
  - Perform on BS-P, BS-R, Title-title similarity, Summary-title similarity.
- Annotation:
  - Annotate 1K validation examples with the labels:
    - No factual error
    - Minor factual errors
    - Major factual errors.
- Optimization objective:

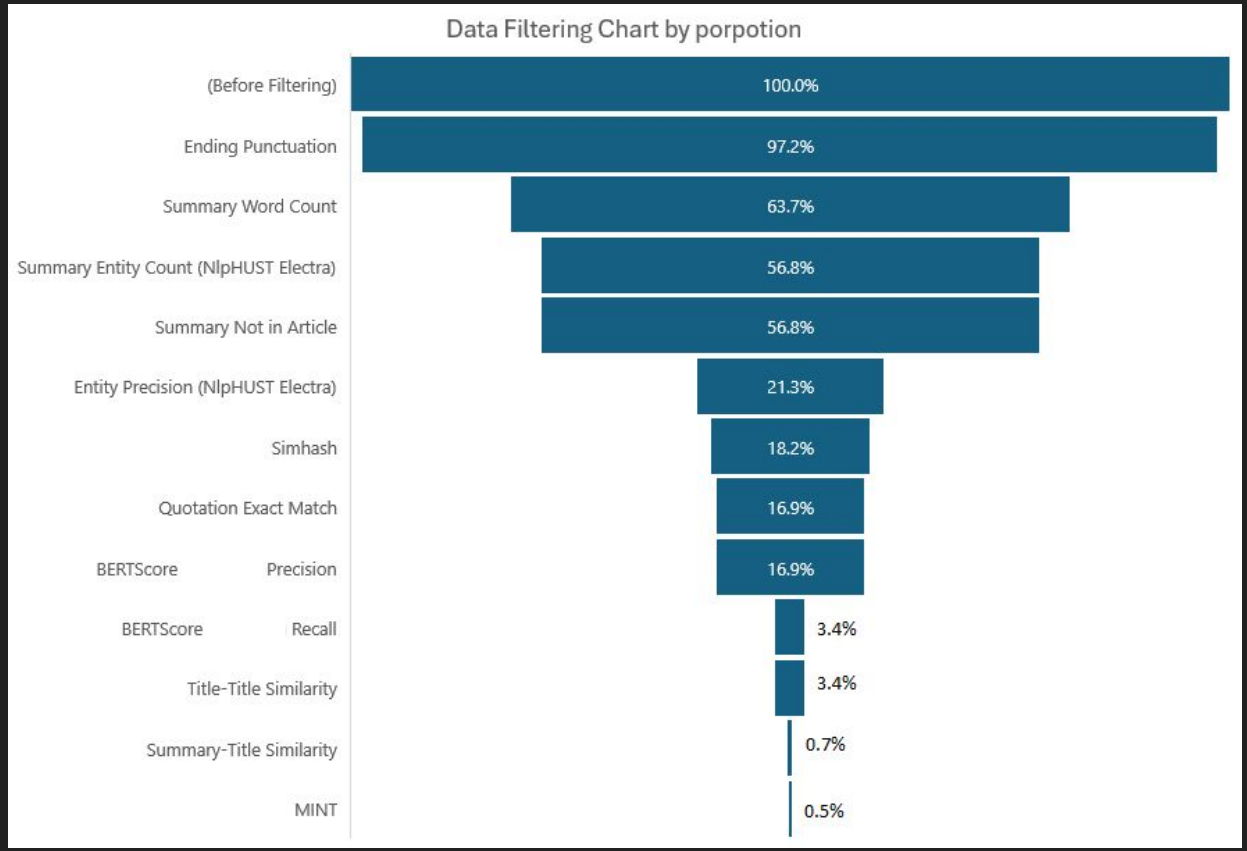
$$f = (0.03 - rate_{MajorFactualError}) + (rate_{Precision} - 0.8) + rate_{Recall}$$

Finalized dataset

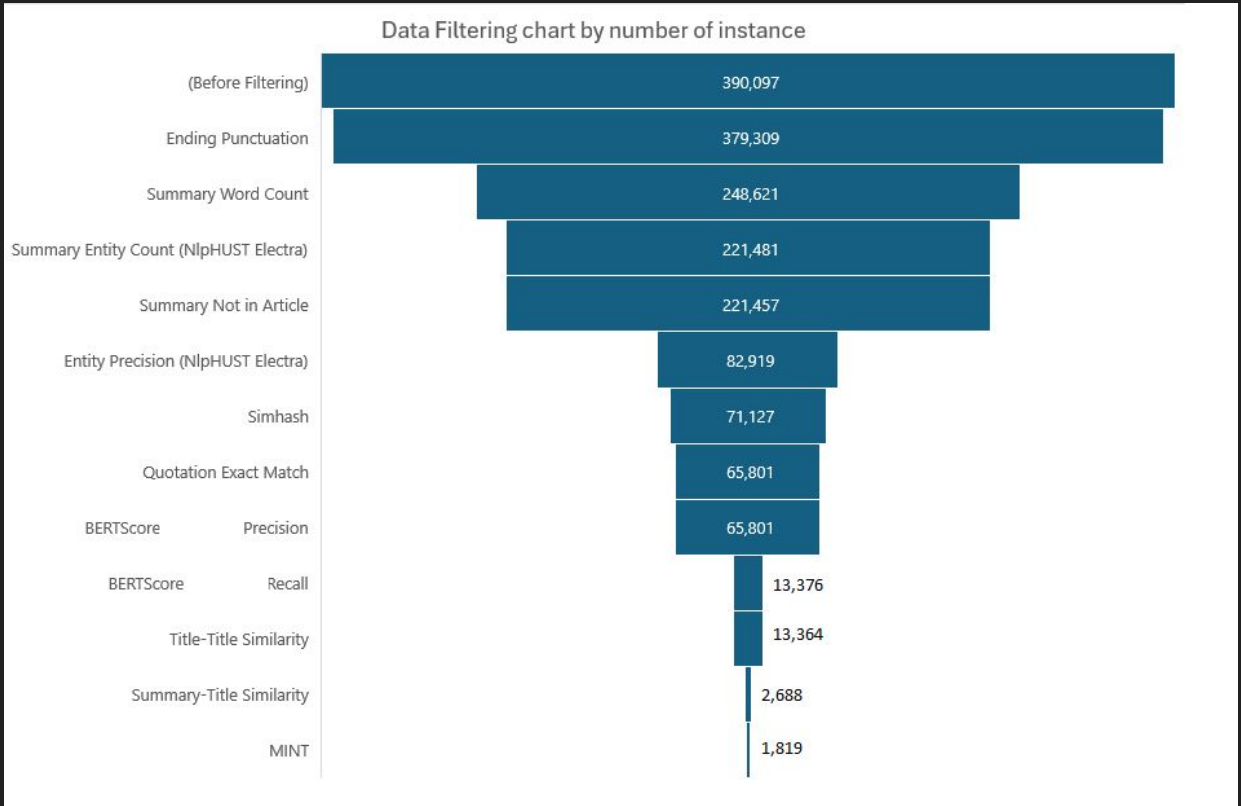




Finalized dataset



Finalized dataset



Finalize the Filtered Dataset

Final dataset: 1819 samples

Split into:

- Train: 1081 (2017-2018)
- Validation: 609 (2019)
- Test: 129 (2022-2023)

EXPERIMENTS

Automatic Evaluation

Model

- Generative text:
  - ROUGE-L
- Abstractiveness:
  - MINT.
- Coverage:
  - BS-Recall

Automatic Evaluation

Dataset

- Factual consistency:
  - BS-Precision
- Abstractiveness:
  - MINT.
- Coverage:
  - BS-Recall

Experiments

Jiang & Dreyer, 2024

- Summarization models:
  - Fine-tune FLAN-T5-Base, ViT5
  - Use FLAN-T5-Base, ViT5 pre-trained as baselines.

Results

On model

	ROUGE-L	BS-Recall	MINT
FLAN-T5-base	0.164	0.559	0.430
FLAN-T5-base (Fine-tuned)	0.310	0.657	0.430
ViT5	0.216	0.581	0.430
ViT5 (Fine-tuned)	0.469	0.815	0.430

Results

On dataset

	BS-Precision	BS-Recall	MINT
CCSum	0.806	0.481	0.480
Ours	0.798	0.673	0.489

Conclusion

- Presented a methods that automate the process of building a news summarization dataset
- Our dataset shows improvement on fine-tuned LLM compare to its pretrained.

Future development

- Filter more data
- Human Evaluation
- Finetune more models
- Find more filtering method specifically for Vietnamese

References

Xiang Jiang and Markus Dreyer. 2024. CCSum: A Large-Scale and High-Quality Dataset for Abstractive News Summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7306–7336, Mexico City, Mexico. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

THANK YOU

Team Members

Team 19

Natural Language Processing for Data Science (DS310.P11)

- Do Ba Huy
- Nguyen Mai Chi Tan
- Huynh Nhan Thap