SUMHOPE: An Automated Abstractive Vietnamese News Summarization Dataset Constructing Method

Nguyen Mai Chi Tan[§], Huynh Nhan Thap[§], Do Ba Huy[§]

VNU-HCM University of Information Technology, Ho Chi Minh City, Vietnam

{21521414, 21521457, 21522137}@gm.uit.edu.vn

Abstract

Abstractive summarization has garnered significant attention in the field of natural language processing (NLP) due to its ability to generate concise and coherent summaries by synthesizing the main ideas of a text rather than merely extracting sentences. However, obtaining datasets which both large and high quality is costly and contain a considerable amount of noise. For Vietnamese, the development of high-quality datasets is crucial to advancing research in this area. The CCSUM dataset paper introduced an automated pipeline to build such large-scale and high-quality dataset, specifically designed for abstractive news summarization. In this project, we apply those methods used to build the CCSUM dataset on Vietnamese news articles to demonstrate its effectiveness and practical potential. Furthermore, we apply some changes to make those methods more friendly for Vietnamese language.

1 Introduction

Abstractive news summarization aims to condense news articles into concise, coherent summaries by rephrasing and synthesizing the main points. Unlike extractive summarization, which uses verbatim sentences from the original text, abstractive summarization generates summaries through paraphrasing (Jiang and Dreyer, 2024). With advancements in transformer-based architectures (Vaswani et al., 2023), abstractive summarization has seen significant improvements, including larger models, refined objective functions (Cao and Wang, 2021), and strategies for enhancing factual accuracy (Zhu et al., 2021) and coherence (Gunel et al., 2020).

However, progress in this area has been limited by the quality and diversity of available datasets, such as CNN/Daily Mail (Hermann et al., 2015) and XSum (Narayan et al., 2018), which contain noise and factual inconsistencies (Tejaswin et al., 2021). To address these challenges, CCSUM (Jiang and Dreyer, 2024) introduces a large-scale dataset with 1.3 million article-summary pairs, emphasizing factual accuracy and coherence, but it primarily serves English-language news.

In Vietnam, several studies on text summarization have been conducted, such as the ViMs dataset (Tran et al., 2020) for abstractive multi-document summarization, which provides valuable resources but does not focus on news summarization; (Dang and Nguyen, 2023) use contrastive learning to prioritize relevant information and improve summary quality; (Le Ngoc) apllies a pre-trained transformer models like BERT (Devlin et al., 2019) to summarize online newspapers, fine-tuned on Vietnamese news datasets, to enhance summarization quality. However, these studies have not proposed specific methods tailored for summarizing news articles in Vietnamese, leaving a gap in the research.

With the development of large language models like GPT (Yenduri et al., 2023), natural language processing tasks have become more efficient. These models not only generate fluent text but also improve summarization tasks by learning from vast amounts of data. However, due to their efficiency, the news summarization task was considered dead since training those model no longer gives an improvement on document summarizing task anymore (Pu et al., 2023). In spite of that, CCSUM managed to create a dataset which shows improvement on news summarization task when training on LM like FLAN-T5 (Chung et al., 2022). This fine-tuned FLAN-T5 exceeded GPT 3.5, MixtralInstruct (Jiang et al., 2024) and Pegasus (Zhang et al., 2020) on news summarization task. CCSUM paper demonstrated its effectiveness on models like Flan-T5 in summarizing news articles. Inspired by this, our study adapts CCSUM methods to summarize Vietnamese news articles, using the BKAI News Corpus (Duc et al., 2024). Our main contributions can be summarized as follows:

[§]Equal contribution.

- We apply CCSUM method on Vietnamese news dataset.
- We remove outlier entities in Vietnamese news to reduce noise.
- We use Vietnamese-specific methods when it comes to models, e.g. we use Vietnamese Sentence-BERT instead of Sentence-BERT, vi-spaCy instead of spaCy,...
- We fine-tune and evaluate FLAN-T5-base and viT5-base on our newly constructed dataset.

2 Related Work

Automatic summarization, especially abstractive summarization, is a critical task in the field of Natural Language Processing (NLP), which has garnered significant attention in recent years. The primary focus of research in this area is developing deep learning models that can generate concise and accurate summaries while retaining the essential information from the original text.

One of the most well-known datasets for abstractive summarization is **CNN/Daily Mail (CNN/DM)** (Hermann et al., 2015), created from news articles on CNN and Daily Mail websites. This dataset is paired with bullet-point descriptions, which are used as summaries for the articles. However, a limitation of this dataset is that the descriptions were not initially designed to be abstractive summaries, and thus, some summaries may be incoherent or not fully representative of the article's content. This dataset has been widely used in summarization research, particularly in training and evaluating models for news summarization.

Similarly, **XSum** (Narayan et al., 2018) is another popular dataset used for abstractive summarization. It contains news articles from BBC, where the first sentence of each article is removed and used as the summary for the rest of the article. This dataset includes around 204k training samples. However, the introductory sentences were not specifically intended to be abstractive summaries and may contain information that is not present in the article, which can lead to summaries that are not fully accurate or comprehensive.

Another notable work is **CCSUM** (Jiang and Dreyer, 2024), where the authors introduce the CCSUM dataset, specifically designed for abstractive news summarization in English. This dataset includes millions of news articles and human-annotated high-quality summaries, providing a valuable resource for training abstractive summarization models for English news. While this dataset is not tailored for Vietnamese, it has had a significant impact on advancing summarization research in other languages, helping to produce more accurate and natural summaries.

For Vietnamese, a high-quality Vietnamese dataset for abstractive multi-Document summarization named **ViMs** (Tran et al., 2020) introduces the ViMs dataset, which is extractive summarization from multiple documents. Multi-document summarization is more complex than single-document summarization, as the model needs to aggregate information from several sources and generate a concise yet coherent summary. This dataset is essential for training models to handle multi-document inputs in Vietnamese.

VLSP-2022 (Tran et al., 2023) introduces a dataset of Vietnamese news summarizing which depends heavily on human annotation. By crawling raw data from websites, clustering, remove similar news. Annotator then write summary for these articles manually then perform cross checking, before get final checking by experts.

In (Dang and Nguyen, 2023) the authors propose a contrastive learning approach to enhance abstractive summarization for Vietnamese. This method helps the model to learn how to distinguish important information from the source text, improving the generation of summaries that are less likely to directly replicate the article's content. The contrastive learning technique proves effective in generating more accurate and natural summaries by focusing on relevant information.

Lastly, Le Ngoc (Le Ngoc) utilizes pre-trained models such as BERT (Devlin et al., 2019) to summarize Vietnamese online news articles. These pre-trained models enhance abstractive summarization by leveraging semantic and syntactic knowledge acquired from large-scale datasets, significantly reducing the need for extensive task-specific training. Experimental results demonstrate that pre-trained models perform well in producing concise, meaningful summaries.

3 Dataset Construction

3.1 Overview

The original news data before filtering is partially from the BKAI News Corpus (Duc et al., 2024), covering more than 32 million Vietnamese news articles from 1970 to 2023. However, due to our lim-

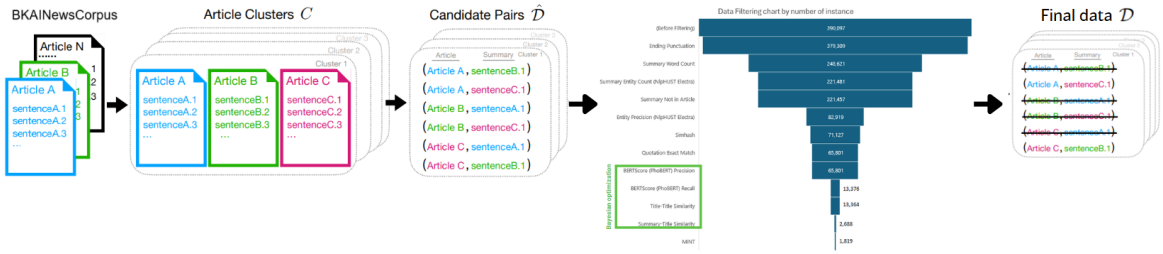

Figure 1: Process of constructing the final dataset.

ited computational resources, we selected 390,097 articles from the years 2017, 2018, 2019, 2022 and 2023 to perform filtering). Our method for extracting high-quality abstractive summaries is based on the CCSUM pipeline (Jiang and Dreyer, 2024).

Figure 1 illustrates the process of this method. After some preparatory steps on the original data, we begin by clustering news articles $A_i \in \mathcal{A}$ into news events $\mathcal{C}$. Within each event $C \in \mathcal{C}$, we select the first sentence A_i^1 of the article $A_i \in C$ as the summary for another article $A_j \in C$ to have candidate article-summary pairs. Then, we define a set of filters to ensure summary quality and use Bayesian optimization to fine-tune some filters' parameters. The final dataset contains 0.5% of the candidate pairs that have met all filter criteria.

3.2 Data Preprocessing

After observation of several news articles in the original dataset, we see some articles have the abbreviated names of their newsrooms (such as "LĐO" for "Lao Động Online", "TTO" for "Tuổi Trẻ Online") at the beginning of the first sentence. Thus, we collect a list of these abbreviated names and remove these names if they appear at the beginning of the articles.

3.3 Clustering for News Event Detection

Next, we group the articles into three-day windows $T_{[t_{\min}, t_{\max})}$ based on their publication times. This is to keep the temporal coherence of articles within each news event. For each window, we apply the Sentence-BERT (Reimers, 2019) model fine-tuned for Vietnamese[1] to encode the main text and perform soft clustering using FAISS (Johnson et al., 2021). In soft clustering, an article can be assigned to multiple clusters, which allows us to overgenerate candidate article-summary pairs followed-by strict filtering to obtain the final large-scale and high-quality dataset.

3.4 Candidate Summary Generation

We construct candidate summaries

$$\hat{D} = \left\{ \left(A_i, A_j^1\right) \mid i \neq j, A_i \in C, A_j \in C \right\}$$

for each cluster C, where the first sentence A_j^1 of article A_j is proposed as an abstractive summary for article A_i.

Simply having semantic similarity between news articles does not ensure that the first sentence will effectively function as a high-quality abstractive summary of the other article. To address this, we define a series of filters to ensure that the chosen summary is relevant, factual and encapsulates the key points of the article. Figure 1 illustrates the data filtering funnel.

3.5 Summary Quality Filter Definition

We use a set of metrics similar to that of CCSUM (Jiang and Dreyer, 2024), denoted as $\mathcal{F}$, to evaluate the quality of each candidate article-summary pair $\left(A_i, A_j^1\right) \in \hat{D}$. Each metric f_x is associated with a specific filter constraint, defined by some threshold λ_{f_x}. The final dataset D consists of article-summary pairs from $\hat{D}$ that meet all filter criteria, where $D = \left\{ \left(A_i, A_j^1\right) \in \hat{D} \mid \forall f_x \in \mathcal{F}, f_x\left(A_i, A_j^1\right) > \lambda_{f_x} \right\}$.

Table 1 defines our full list of filters, where filter constraints with λ denote paramters to be optimized using Bayesian optimization. Figure 2 presents a detailed filtering funnel, including the individual heuristics and BERTScores with BERT-base-multilingual model. Figure 3 depicts the proportions of the pairs by years before and after filtering.

Because of Vietnamese-specific nature, we have different methodology compared to CCSUM, whose details are described in this section. For named entity recognition, we use the Electra base

[1] https://huggingface.co/keepitreal/vietnamese-sbert

Metric name	Notation f_x	Value Range	Filter Constraint A_x
Ending punctuation	f_{punct}	$\{\text{True}, \text{False}\}$	$\{\text{True}\}$
Summary word count	$f_{\text{word\_count}}$	$[1, \infty)$	$[25, \infty)$
Summary entity count	$f_{\text{entity\_count}}$	$[0, \infty)$	$[1, \infty)$
Summary not in article	$f_{\text{not\_in\_article}}$	$\{\text{True}, \text{False}\}$	$\{\text{True}\}$
Entity precision	f_{ep}	$[0, 1]$	$\{1\}$
Simhash	f_{simhash}	$[0, 64]$	$(5, 64]$
Quotation exact match	$f_{\text{quotation}}$	$\{\text{True}, \text{False}\}$	$\{\text{True}\}$
BERTScore Precision	$f_{\text{BS—P}}$	$[0, 1]$	$[\lambda_{\text{BS–P}}, 1]$
BERTScore Recall	$f_{\text{BS—R}}$	$[0, 1]$	$[\lambda_{\text{BS–R}}, 1]$
Title-title similarity	$f_{\text{t—t}}$	$[0, 1]$	$[\lambda_{\text{t–t}}, 1]$
Summary-title similarity	$f_{\text{s—t}}$	$[0, 1]$	$[\lambda_{\text{s–t}}, 1]$
MINT	f_{MINT}	$[0, 1]$	$[0.2, 1]$

Table 1: Overview of the filters.

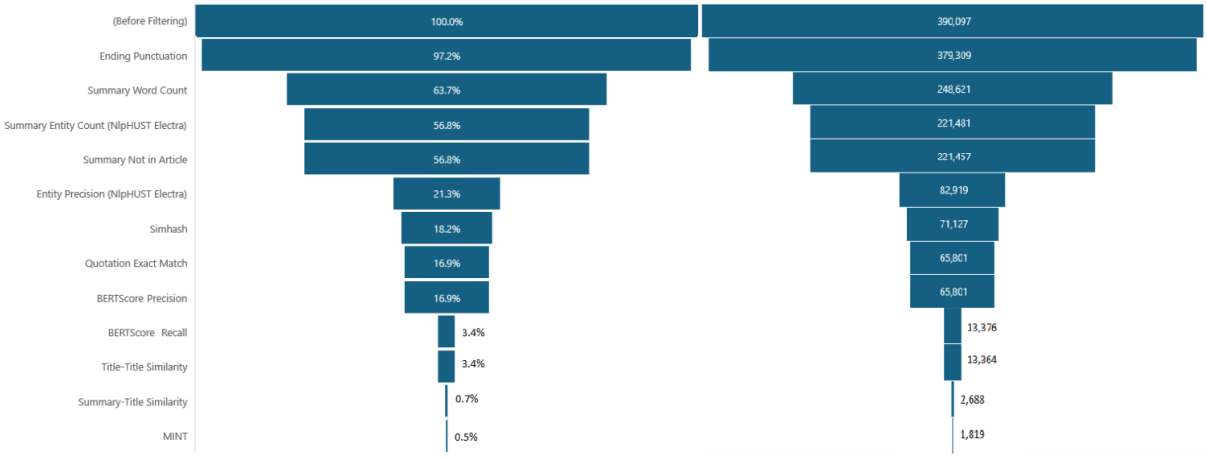

Figure 2: Detailed filter funnel in percentages (left) and in numbers of pairs (right).

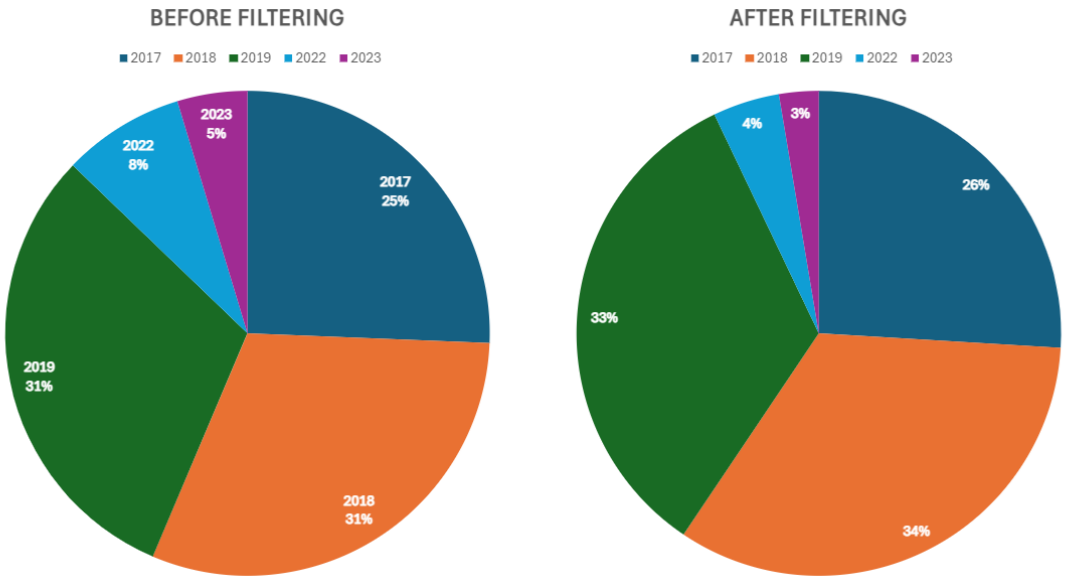

Figure 3: The proportions of article-summary pairs by years before filtering (left) and after filtering (right).

model (Clark, 2020) fine-tuned[2] on the VLSP 2018 dataset.

Heuristics. We define a set of heuristics specifically designed to mitigate challenges raised by this candidate summary generation approach. (i) *Summary entity count* $f_{\text{entity\_count}}$. A summary must have at least one entity; otherwise, the lead sentence could be a generic introductory sentence that does not capture the main points of the article. (ii) *Summary word count* $f_{\text{word\_count}}$. A summary must have at least 25 words. We segment the words from summaries using a word segmentation model for Vietnamese.[3]

Factual Consistency. A valid summary must accurately reflect the content of the original news article while avoiding introducing unsupported information. We use the following metrics for factual consistency: (i) *Entity precision* f_{ep} quantifies the proportion of entities in the summary that also appear in the article (Nan et al., 2021). (ii) *BERTScore-Precision* (BS-Precision) $f_{\text{BS-P}}$. BERTScore Precision works well to evaluate the factual consistency of an abstractive summary against an article (Zhang et al., 2019). We calculate BERTScores using multilingual BERT (Devlin et al., 2019), when CCSUM authors use BERT and BART for their English-specific dataset. (iii) *Quotation exact match* f_{quo}: We use regular expression to detect quotations from the summary and ensure their precense in the article.

Coverage. A well-crafted summary should capture the key points of an article while avoiding extraneous details. We use the following metrics to measure coverage: (i) *BERTScore-Recall* (BS-Recall) $f_{\text{BS-R}}$ aligns tokens in the article with tokens in the candidate summary and computes the average similarity score over tokens in the article (Zhang et al., 2019). (ii) *Title-title similarity* $f_{\text{t—t}}$ measures the semantic similarity between the summary's title, using cosine similarity of Vietnamese Sentence-BERT embeddings (Reimers, 2019). This metric determines if the summary's topic aligns with that of the article. (iii) *Summary-title similarity* $f_{\text{s—t}}$ measures the semantic similarity between the candidate summary A_j^1 and the title of the article A_i.

Abstractiveness. A well-crafted abstractive summary should effectively synthesize and condense the content of an article, rather than merely replicating sentences or phrases from it. We use MINT f_{MINT} (Dreyer et al., 2021) and *Simhash* f_{simhash} (Manku et al., 2007) to measure the abstractiveness of the candidate summary.

3.6 Bayesian Optimization

Given the wide range of filters and their intricate interactions, manually setting thresholds for each filter is impractical. We use Bayesian Optimization (Kushner, 1964; Frazier, 2018) to address this challenge by performing constrained optimization over filter parameters.

The Search Space. We perform filter parameter search on all the embedding-based metrics, i.e., BERTScore-Precision, BERTScore-Recall, title-title similarity and article-title similarity, which we denote as a vector λ.

Collecting Human Annotations. Bayesian optimization requires a measurable objective function to optimize in the parameter space. We use human judgements on a held-out set of candidate article-summary pairs to guide the search for filter parameters. The intuition is that the ideal filter parameterization should maximize the recall of high-quality summaries while reducing the presence of low-quality summaries. We select the first 1,000 article-summary pairs from 2021 for our candidate dataset $\hat{D}$ for human annotation. Each candidate article-summary pair is annotated into one of the three categories:

- No factual error
- Minor factual error
- Major factual error

We obtain an annotated dataset $\hat{D}_{labelled}$ where each example (A_i, A_{1j}, y) is associated with a factual consistency label y.

Optimization Objective. The optimization consists of a primary objective and two constraints. The primary objective is the recall of factually correct summaries $f_{recall}(\lambda, \hat{D}_{labelled})$, where

$$f_{recall} = \frac{|\{(A_i, a_j^1, y) \in D_{labelled} \,|\, y = correct\}|}{|\{(A_i, a_j^1, y) \in \hat{D}_{labelled} \,|\, y = correct\}|} \quad (1)$$

However, focusing solely on recall could result in a trivial solution, i.e. removing all filters. Therefore, we use two more constraints to ensure that the

[2] https://huggingface.co/NlpHUST/ner-vietnamese-electra-base

[3] https://huggingface.co/NlpHUST/vi-word-segmentation

Metric name	Constraint
Ending punctuation	{True}
Summary word count	$[25, \infty)$
Summary entity count	$[1, \infty)$
Summary not in article	{True}
Entity precision	{1}
Simhash	$(5, 64]$
Quotation exact match	{True}
BERTScore Precision	$[0.9386, 1]$
BERTScore Recall	$[0.0008, 1]$
Title-title similarity	$[0.9922, 1]$
Summary-title similarity	$[0.6175, 1]$
MINT	$[0.2, 1]$

Table 2: Final filter parameters.

final dataset is both large-scale and of high quality. The first constraint limits the major factual error rate $f_{error\,major}(\lambda, \hat{D}_{labelled})$, defined in Eq. 2, after filtering to be less than 3%.

$$f_{error_{major}} = \frac{|\{(A_i, a_j^1, y) \in D_{labelled} \,|\, y = error_{major}\}|}{|D_{labelled}|} \tag{2}$$

The second constraint focuses on the precision of factually correct summaries $f_{precision}(\lambda, \hat{D}_{labelled})$, defined in Eq. 3. It requires that the percentage of factually correct summaries exceeds eighty percent after filtering.

$$f_{precision} = \frac{|\{(A_i, a_j^1, y) \in D_{labelled} \,|\, y = correct\}|}{|D_{labelled}|} \tag{3}$$

Eq. 4 defines the overall optimization objective:

$$\begin{aligned} \max_{\lambda} \quad & f_{recall}(\lambda, D_{labelled}) \\ \text{subject to} \quad & f_{error_{major}}(\lambda, \hat{D}_{labelled}) \\ & f_{precision}(\lambda, \hat{D}_{labelled}) \end{aligned} \tag{4}$$

This filtering mechanism can also be applied to existing datasets for quality improvements. Table 2 summarizes the final filters' parameters after optimizing.

4 Dataset Overview and Evaluation

Table 3 presents the final dataset of 1,819 article-summary pairs. Table 7 shows some examples of summaries along with their corresponding articles in our dataset.

Train/validation/test splits. The dataset is divided into training, validation, and testing subsets, categorized by the publication data of each article (Table 3). The CCSUM pipeline (Jiang and Dreyer, 2024) goes with a similar approach to avoid the risk of news event leakage, where a summarization model might inadvertently recall and use details of a news event from the training set to generate summaries for the testing set, rather than relying on the content of the test article themselves.

	Total	Year range
Train	1,081	2017–2018
Val.	609	2019
Test	129	2022–2023

Table 3: Overview of our final dataset.

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

Table 4: Benchmarking abstractive summarization datasets using automatic metrics.

4.1 Automatic Evaluation

We perform automatic evaluation like that of CCSUM to benchmark the final dataset along with the CCSUM dataset. The automatic evaluation results on the test split on each dataset are presented in Table 4. The results indicate that our summaries have a good score in factual consistency, are comprehensible, and capture the main points of the article.

Factual consistency. We use BS-Precision to evaluate factual consistency. Table 4 shows that our dataset has a slightly lower score than CCSUM.

Abstractiveness. Abstractiveness quantifies the lexical abstraction in summaries; high MINT (Dreyer et al., 2021) indicates high abstractiveness. Table 4 shows that our dataset is the more abstractive when measured on the entire dataset.

Coverage. BS-Recall measures the extent to which summaries cover the articles' main contents. Our dataset has the highest coverage when measured on BS-Recall.

5 Experiments

5.1 Summarization Models

We use the models FLAN-T5 (Chung et al., 2022) and ViT5 (Phan et al., 2022) to fine-tune their weights with our dataset, because they were pretrained on the Vietnamese language. Due to our

	ROUGE-L	BS-Recall	MINT
FLAN-T5	0.164	0.559	0.430
ViT5	0.216	0.581	0.430

Table 5: Automatic evaluation before fine tuning.

	ROUGE-L	BS-Recall	MINT
FLAN-T5	0.310	0.657	0.430
ViT5	**0.469**	**0.815**	0.430

Table 6: Automatic evaluation after fine-tuning.

limited computational resources, we chose to fine-tune the base version of FLAN-T5, when the CCSUM authors used the large version. We then compared it with models that have been specifically trained on Vietnamese, such as the pre-trained ViT5 and FLAN-T5.

5.2 Performance Comparison of Summarization Models

Table 5 and 6 provides a comparative analysis of four summarization models based on three evaluation metrics: ROUGE-L, BS-Recall, and MINT. Notably, fine-tuning the FLAN-T5 model significantly enhances its performance, although the MINT score remains unchanged. Similarly, the ViT5 model demonstrates superior performance compared to the baseline FLAN-T5. The fine-tuned ViT5 model exhibits the highest performance overall. These findings underscore the substantial improvements achieved through fine-tuning in both FLAN-T5 and ViT5 models, particularly in ROUGE-L and BS-Recall metrics.

6 Conclusion

We present a dataset which adopt CCSUM method on Vietnamese data. Our dataset consist of 1819 samples filtered from 390,097 article-summary pairs from BKAI News Corpus dataset. By constructing this dataset, we propose Vietnamese-language friendly filtering pipeline that helps filter 95% of original data. The newly created proven itself high-quality when evaluate on many metrics that ensure factual consistency, abstractiveness and coverage. Moreover, lanuage models like FLAN-T5 and viT5 show significant improvement when fine-tuned on our dataset.

Although our dataset is not enough to be considered large-scale, our proposed methods can then be applied to any Vietnamese news corpus dataset to enlarge its scale.

Limitations

In this project, our dataset evaluation lacks metrics for comprehensivennes, which is crucial in news summary task. We also want to filter all data available in in BKAI News Corpus that we couldn't do in this work due to limited resources. Moreover, finetune and evaluate more models including LLMs to shows our data effectiveness on variety of models and gives you more information to compare is also our development direction.

References

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Nhat Minh Dang and Truong Son Nguyen. 2023. Contrastive learning for vietnamese abstractive summarization. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2021. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. *arXiv preprint arXiv:2108.02859*.

Nguyen Quang Duc, Le Hai Son, Nguyen Duc Nhan, Nguyen Dich Nhat Minh, Le Thanh Huong, and

Dinh Viet Sang. 2024. Towards comprehensive vietnamese retrieval-augmented generation and large language models. *arXiv preprint arXiv:2403.01616*.

Peter I Frazier. 2018. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2020. Mind the facts: Knowledge-boosted coherent abstractive text summarization. *Preprint*, arXiv:2006.15435.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Harold J Kushner. 1964. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise.

Thang Le Ngoc. Vietnamese online newspapers summarization using pre-trained model. *AKTUAL'NYYe ISSLEDOVANIYA*, page 9.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *Preprint*, arXiv:1808.08745.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. Association for Computational Linguistics.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449, Online. Association for Computational Linguistics.

Mai-Vu Tran, Hoang-Quynh Le, Duy-Cat Can, and Quoc-An Nguyen. 2023. Overview of the vlsp 2022–abmusu shared task: A data challenge for vietnamese abstractive multi-document summarization. *arXiv preprint arXiv:2311.15525*.

Nhi-Thao Tran, Minh-Quoc Nghiem, Nhung TH Nguyen, Ngan Luu-Thuy Nguyen, Nam Van Chi, and Dien Dinh. 2020. Vims: a high-quality vietnamese dataset for abstractive multi-document summarization. *Language Resources and Evaluation*, 54(4):893–920.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. 2023. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *Preprint*, arXiv:2305.10435.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

Summary	Article
Một đại gia tại khu dân cư Phước Nguyên Hưng, xã Phước Kiển, huyện Nhà Bè, TP. HCM thông báo bị trộm đục két sắt, mất gần 6 tỷ đồng.	Một đại gia ở TP.HCM đến công an trình báo bị kẻ trộm đột nhập, đục két sắt, cuỗm đi tài sản lên đến 5,7 tỷ đồng. Ngày 4/5, Công an huyện Nhà Bè phối hợp cùng các đơn vị nghiệp vụ Công an TP.HCM, điều tra vụ một đại gia bị mất trộm tại khu dân cư Phước Nguyên Hưng, xã Phước Kiển, huyện Nhà Bè, TP.HCM. Thông tin ban đầu, sáng 1/5, 2 người giúp việc phát hiện phòng ngủ tầng 2 của căn nhà có dấu hiệu đột nhập. Két sắt và tủ gỗ trong căn phòng bị cạy phá. Người giúp việc báo lại vụ việc cho chủ nhà. Chủ nhà kiểm tra thì phát hiện 30 lượng vàng, 120.000 USD, 3 nhẫn kim cương và đôi bông tai vàng bị lấy trộm. Tổng tài sản bị mất khoảng 5,7 tỷ đồng. Xã Phước Kiển nơi xảy ra vụ trộm. Ảnh: Google Maps.
Một vụ sập giàn giáo vừa mới xảy ra tại công trình thủy điện ở Lào đã khiến 2 lao động trú ở huyện Yên Thành (Nghệ An) tử vong.	Hai nạn nhân là anh Võ Văn Tuấn (SN 1985, trú xã Phú Thành, huyện Yên Thành, Nghệ An) và anh Phan Văn Thái (SN 1993, trú xã Tân Thành, huyện Yên Thành) tử vong ở Lào. Sáng 12-1, chính quyền địa phương nơi các anh sinh sống đã xác nhận sự việc trên. Theo đó, một số công nhân cùng đi sang Lào lao động với anh Tuấn, Thái cho biết, trong lúc đang làm việc tại công trình điện Mường Khống (Lào) thì bất ngờ giàn giáo bị sập khiến 2 nạn nhân nói trên tử vong vào sáng 11-1. Hiện, chính quyền xã Tân Thành và chính quyền xã Phú Thành cùng gia đình đã tiến hành làm thủ tục sang Lào để đưa thi thể 2 anh về quê.
Quân đội Hàn Quốc và Mỹ ngày 29/7 đã tiến hành cuộc tập trận tên lửa đạn đạo và tấn công chính xác nhằm đáp trả vụ phóng thử tên lửa vừa diễn ra đêm 28/7 của Triều Tiên.	Lực lượng Mỹ và Hàn Quốc vừa phóng một quả tên lửa đạn đạo vài giờ sau khi Triều Tiên thử một quả tên lửa mới đêm 28/7. Hàn Quốc và Mỹ tiến hành vụ phóng này nhằm đáp trả trực tiếp Bình Nhưỡng, hãng thông tấn Yonhap của Hàn Quốc cho biết. Thông cáo của Tham mưu trưởng liên quân Hàn Quốc nêu rõ, vụ thử tên lửa đạo đạo này nhằm tái khẳng định năng lực của đồng minh trong việc đáp trả chính xác các vụ tấn công của kẻ địch. Lực lượng Mỹ - Hàn đã phóng tên lửa Hyunmoo-2 và một quả tên lửa đất đối không có tầm bắn khoảng 300 km./.
Chỉ trong vòng 1 tiếng, hai người cùng trú lại xã Yang Tao (Đắk Lắk) đã bị sét đánh thương vong khi đang làm ruộng.	Ngày 30/5, tin từ UBND xã Yang Tao, huyện Lắk (Đắk Lắk) xác nhận, trên địa bàn vừa xảy ra 2 vụ sét đánh khiến 1 người tử vong, 1 người bị thương. Trước đó, vào trưa ngày 29/5 vừa qua, anh Y Nar H'Long (33 tuổi, ngụ tại buôn Dơng Bắc, xã Yang Tao) đi làm ruộng thì bị sét đánh bị thương nặng và được đưa đi cấp cứu. Ảnh minh họa. Chiều cùng ngày, bà H'gốc Kmăm (48 tuổi, ngụ cùng địa phương) cũng bị sét đánh tử vong trong lúc đi làm cỏ dưới ruộng. Nhận được tin báo, chính quyền địa phương đến nhà nạn nhân chia sẻ, hỗ trợ gia đình lo hậu sự.

Table 7: Example article-summary pairs in our dataset.