Journal of Computer Science and Cybernetics, V.40, N.4 (2024), 347–362
DOI no. 10.15625/1813-9663/18291

VLSP 2022 ABMUSU SHARED TASK: A DATA CHALLENGE FOR VIETNAMESE ABSTRACTIVE MULTI-DOCUMENT SUMMARIZATION

MAI VU TRAN, HOANG QUYNH LE*, DUY CAT CAN, QUOC AN NGUYEN

Faculty of Information Technology,
VNU University of Engineering and Technology
E3 Building, 144 Xuan Thuy Street, Cau Giay District, Ha Noi, Viet Nam

Abstract. This paper provides an overview of the Vietnamese abstractive multi-document summarization shared task (AbMuSu) for Vietnamese news, which is hosted at the 9^{th} annual workshop on Vietnamese Language and Speech Processing (VLSP 2022). The main goal of this shared task is to develop automated summarization systems that can generate abstractive summaries for a given set of documents on a specific topic. The input consists of several news documents on the same topic, and the output is a related abstractive summary. The focus of the AbMuSu shared task is solely on Vietnamese news summarization. To this end, a human-annotated dataset comprising 1,839 documents in 600 clusters, collected from Vietnamese news in eight categories, has been developed. Participating models are evaluated and ranked based on their `ROUGE2-F1` score, which is the most common evaluation metric for document summarization problems.

Keywords. Abstractive summarization, Vietnamese summarization dataset, multi-document summarization

1. INTRODUCTION

In the current age of information abundance, effectively extracting valuable insights from data is a challenging task that requires significant investment of time, resources, and human effort. Fortunately, multi-document summarization provides a promising solution to this problem. Leveraging natural language processing techniques, this approach involves analyzing a set of documents to identify and consolidate key information, resulting in a concise summary [1]. Despite its complexity, the research community has increasingly focused on advancing this field. Several past challenges and shared tasks have focused on summarization. One of the earliest summarization shared tasks is the series of document understanding conference (DUC) challenges [a], the Text Analysis Conference (TAC) summarization shared tasks [b]. In recent years, some summarization shared tasks have been launched to support

*Corresponding author.
E-mail addresses: lhquynh@vnu.edu.vn (H.Q Le); vutm@vnu.edu.vn (M.V Tran); catcd@vnu.edu.vn (D.C Can); annq@vnu.edu.vn (Q.A Nguyen)

[a] http://www-nlpir.nist.gov/projects/duc. DUC summarization challenges are organized 7 times from 2000 to 2007.

[b] http://tac.nist.gov/tracks/. TAC summarization shared tasks are organized 5 times on summarization news and biomedical text from 2008 to 2014.

research and development in this field for English, such as DocEng 2019 [2] and BioNLP-MEDIQA 2021 [3], etc.

Automatic summarization can be classified into two main approaches based on output characteristics: extractive and abstractive. Extractive summarization involves selecting the most crucial sentences or sections from the source documents, while abstractive summarization rewrites a new summary based on the original important information [4]. Since the 1950s, various extractive methods have been proposed, from frequency-based [5] to machine learning-based techniques [6]. However, extractive approaches still have significant disadvantages in arranging and combining information from several documents in multi-document summarization tasks. Because of this disadvantage, while these extractive methods are fast and simple, the summaries generated are often not comparable to those created manually. Abstractive methods have emerged as a promising solution, addressing this limitation [6]. Recently, transformer-based sequence-to-sequence learning methods have enabled significant improvements in abstractive summarization. Encoder-decoder models such as PEGASUS [7], BART [8], and T5 [9] have achieved remarkable results for abstractive multi-document summarization, attracting attention from the research community. Studies on this problem for Vietnamese text are still in the early stages with an initial achievement, especially in extractive approaches. In recent years, there has been a growing interest in developing automatic abstractive summarization systems. Despite these attempts, the lack of a comprehensive benchmarking dataset has limited the comparison of different techniques for Vietnamese. VLSP 2022 - AbMuSu shared task is set up to provide an opportunity for researchers to propose, assess, and advance their research, and further, promote the development of research on abstractive multi-document summarization for Vietnamese text.

Moreover, the performance of automatic summarization systems has significantly improved with the development of supervised approaches. In English, there are several summarization datasets available. The CNN/Daily Mail dataset is a popular news dataset that generates summaries based on news headlines. The BigPatent dataset [10] is a large-scale dataset that uses 1.3 million patents and related abstract paragraphs as summaries. However, these datasets are limited in their ability to provide multi-document summaries, which are more comprehensive and condensed than single-text summaries. To address this gap, the MEDIQA-AnS Dataset [11] has recently been published with expert-generated summaries for both documents and clusters. In Vietnamese, because of the limitation of available data sources, proposing a Vietnamese summarization dataset is a challenge. VNDS [12] is the first Vietnamese dataset to provide articles and automatically creates single-document summaries based on the introduction paragraph. The ViMs [13] and VietnameseMDS [c] datasets provide manual multi-document summaries. Nevertheless, as the demand for supervised methods increases, the existing datasets are insufficient for training a supervised model due to their limited number of documents. For instance, VietnameMDS and ViMs contain only 200 and 300 clusters, respectively. Notably, the VLSP AbMuSu shared task has contributed to the development of a new dataset for multi-document summarization, providing a valuable resource for researchers in this field.

The remainder of the paper is organized as follows: Section 2 describes the AbMuSu shared task. Section 3 describes the data construction, annotation methodologies and data

[c] https://github.com/lupanh/VietnameseMDS

collection. Section 4 shows the competition, baselines, approaches, and respective results, and the final section is the conclusions.

2. TASK DESCRIPTION

The VLSP 2022 AbMuSu shared task serves as an exciting opportunity for researchers to develop and showcase their skills in the field of abstractive multi-document summarization. The purpose of this task is to create summarization systems that can generate abstractive summaries automatically for a set of documents that are related to a specific topic. This is achieved by providing the model with multiple news documents on the same topic, and the model's output should be a relevant and informative abstractive summary. It is worth noting that the scope of the AbMuSu shared task is focused solely on Vietnamese news, which presents a unique challenge to participants as they must demonstrate their ability to summarize documents in a language that is not as widely studied as English. Vietnamese has some unique characteristics in terms of syllables, compound words, etc. In addition, the number of summarization models in Vietnamese is also much lower than in English.

To provide a comprehensive dataset for this task, a human-annotated collection of 1,839 documents was developed, with each document belonging to one of the eight categories of Vietnamese news. Furthermore, the AbMuSu task requires participants to summarize a group of documents that share the same topic, which we refer to as 'document clusters.' Each cluster contains two to five documents that provide information on a single topic. The goal of the shared task is to build models that can automatically create an abstractive summary for each cluster. Overall, the VLSP 2022 AbMuSu shared task offers an exciting opportunity for researchers to develop new techniques and strategies for abstractive multi-document summarization. The models submitted for evaluation will be ranked based on their `ROUGE2-F1` score, which is a widely accepted evaluation metric for document summarization tasks. The task provides a valuable resource for researchers to test their models and to contribute to the development of the field of abstractive multi-document summarization.

3. TASK DATA

The process of constructing the Abmusu dataset involves two main tasks, which are carried out sequentially: data collection and summary creation.

3.1. Data collection

Raw data was collected from Baomoi [d] - a Vietnamese E-news aggregator with about 200 Vietnamese official electronic news providers. There are four major steps in the data collection process, as illustrated in Figure 1. The data was initially obtained through web crawling and filtered. Afterward, the articles were semi-automatically grouped into clusters, and a similarity-based approach was employed to identify and remove duplicate documents. Subsequently, a selection process was conducted to choose appropriate clusters for the training, validation, and testing datasets.

[d] https://baomoi.com/

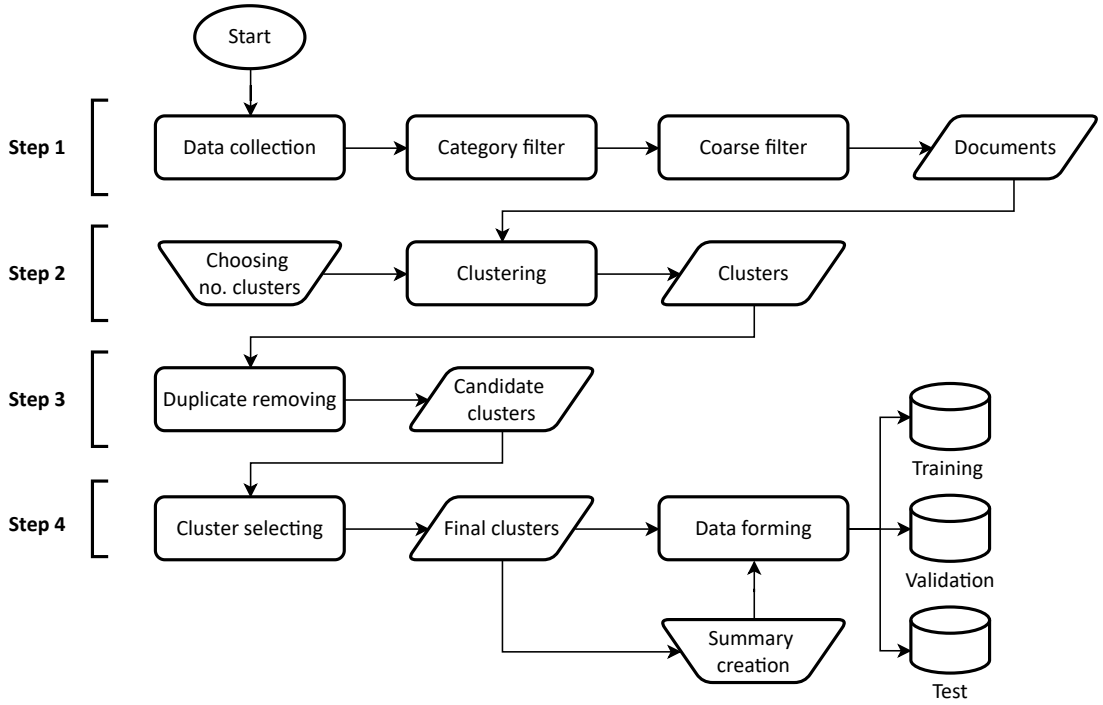

Figure 1: Data collection process

Step 1 - Data collection. The data have been automatically collected and filtered into 8 categories, including *economy*, *culture and society*, *science and technology*, *legal*, *entertainment*, *lifestyles*, *education*, and *world news*. The articles in the dataset comprise a title, anchor text, plain text, hashtags, and category tags. The anchor text, which is an introductory section written by the author, can serve as a summary in certain cases. Nonetheless, some documents may lack anchor text or may contain anchor text that is unrelated to the plain text. To remove such documents automatically, a coarse filter method was employed, and the following types of clusters were eliminated: `Isolated articles` - articles do not have hashtags or all hashtags are empty, `Short articles` - articles with a length is less than 400 Vietnamese characters or less than the length of an introduction paragraph, and `Video/Image articles` - the article contains only images, videos, and corresponding captions. As a result, after the coarse filter method, 7586 articles were collected and saved to the database.

Step 2 - Clustering. We observed a common pattern for articles in the same cluster - they often shared similar characteristics such as hashtags, categories, and publication dates. To incorporate this observation into the clustering process, we included these features in the document vector and used them to calculate the Euclidean distance between documents. Moreover, to ensure the relevance of the clusters, we limited the maximum number of documents in each cluster to 5. To identify the suitable number of clusters, we performed an analysis on the relationship among all documents in the new agglomerative clusters after each combining step. Based on the heuristic approach, we selected a threshold of 3000 clusters. Finally, we removed clusters that contained only one document, resulting in 1877 clusters from 6462 articles. Each cluster contains between 2 and 5 documents.

Step 3 - Duplicate removing. In this step, we use Jaccard similarity to compare the similarity between each pair of document vectors in each cluster, as follows

$$J_{sim}(S_{D1}, S_{D2}) = \frac{|S_{D1} \cap S_{D2}|}{|S_{D1} \cup S_{D2}|} \quad (1)$$

where S_{D1} and S_{D2} are set tokens of documents $D1$ and $D2$ respectively in the same cluster.

Based on the published timestamp, the latter article is eliminated if the Jaccard similarity score is more than 90%. Clusters containing one document are removed. The number of retained clusters is 1467.

Step 4 - Dataset forming. This step involves forming two parts of data. The first part contained 400 clusters, which were selected by the experts. The experts assessed the quality of the documents based on three criteria: consistency, diversity, and acceptable distribution. (i) *Consistency*: Articles within a single cluster must share one and only one common main topic, which must be the primary topic of each article. For example, an article about the impact of COVID-19 with some ideas about economic difficulties should not be classified as a topic about the economic situation. (ii) *Diversity*: prioritize clusters with many articles from many sources, clusters with articles on many different aspects of a topic. (iii) *Acceptable distribution*: Ensuring that each of the eight predefined categories had at least three clusters in each data subset (train, validation, and test). This guarantees a balanced representation of all categories across the different datasets. These 400 clusters were then separated into 200 clusters for the training set and 200 clusters for the test set. This systematic approach ensured that the 400 selected clusters were of high quality. To address the issue of unrelated or low-quality documents commonly encountered in real data, we randomly selected another 200 clusters, dividing them into 100 clusters for the validation set and 100 clusters for the test set.

As a result, 600 clusters with 1839 news articles were chosen. Each cluster has 2-5 documents that illustrate the same topic. Each single document contains five parts: title, anchor text, raw text, and category. All 600 clusters were then annotated with a multi-document abstractive summary. During the training and validation phases, the manually created reference abstractive summary is provided for each cluster. In contrast, to ensure the fairness of the results, the labels of the test set are not made public. We maintain an online evaluation system [e] , enabling users to submit predicted outputs for the testing result.

3.2. Summary creation

The summary creation process involved the manual creation of multi-document summaries after completing the data collection steps. We used INCEpTION [f] as the annotation tool.

There are four roles for the annotator process: *manager*, *annotator*, *supervisor*, and *expert*. *The manager* is directly responsible for operating and assigning members. *The annotator*, *supervisor*, and *experts* participate in the generated summaries process with four main steps, which are shown in Figure 2. The *annotator* first created a draft summary for a cluster (***summary draft annotation***). Then, the *supervisor* reviewed and classified

[e] https://aihub.ml/competitions/341
[f] https://inception-project.github.io

them into three quality levels: `type 1` - meeting all requirements, `type 2` - having fixable errors such as typos or punctuation mistakes, and `type 3` - having significant errors such as semantic misunderstandings (***summary review***). These errors were handled by *annotators* and *supervisors* (***error handling***). Finally, any remaining errors are reviewed and refined by *experts* (***expert curation***). All annotators, supervisors, and experts are native speakers and were trained carefully in the annotation process. They were provided with guidelines and participated in some training meetings. In detail, there are five graduate students (three in computer science and two in information systems) in the annotator role. The annotators are seniors who have experience labeling other datasets. There are three graduate masters in computer science in the supervisor role. There are two PhDs as experts who have experience in Natural Language Processing and have researched Vietnamese summarization tasks.

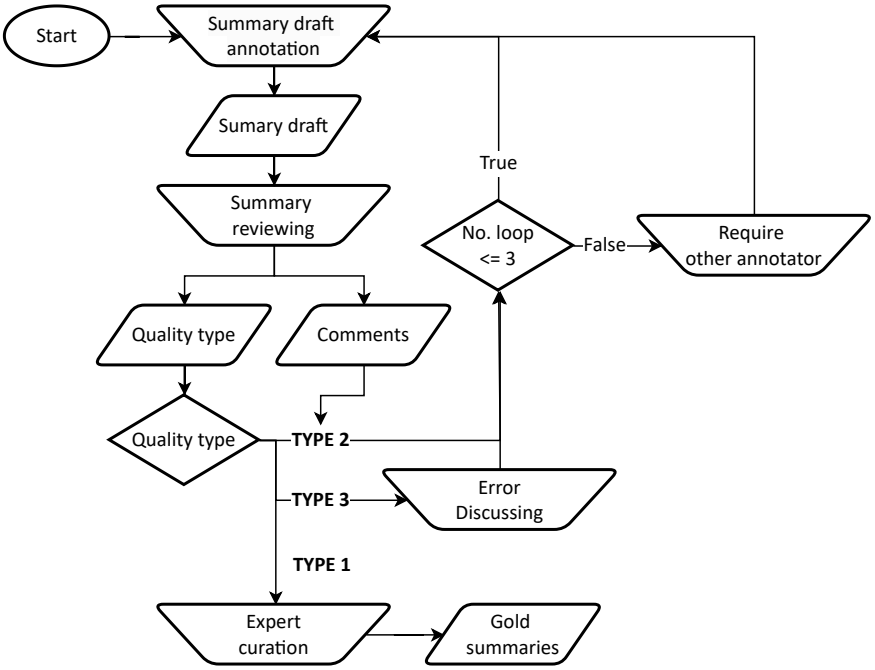

Figure 2: Summary creation process

Step 1 - Summary draft annotation. Annotators create draft summaries for the clusters based on an annotation guideline. Some key criteria of annotation guidelines are as listed follows: (i) The summary must be based on the original content of all articles, without any inclusion of the annotator's personal opinions, (ii) The summary must be coherent and cohesive and (iii) The annotator should pay attention to confusing words, such as adverbs of time (e.g., tomorrow, tonight, today) and pronouns (e.g., he, she, they) that may lead to conflicts between documents.

Step 2 - Summary reviewing. Based on the guidelines, the supervisor evaluated and scored the summary draft. There were three types for the manual summary quality, which are described in Table 1. If the quality of the summary is `Type 1`, it means that the summary meets all the requirements stated in the guidelines. `Type 2` indicates that the summary has some fixable errors, such as typos, punctuation, and wordiness, which are commented on in the review notes. `Type 3` implies that the summary has serious errors that directly impact its

quality. Examples of such serious errors include lacking vital information from documents, containing long and incoherent sentences, and providing false information.

Table 1: The description of three levels of manual summary quality

Types	Meaning
1	Satisfy all requirements
2	Having fixable error(s) such as typo, punctuation error
3	Having serious error(s) such as semantic error, understanding error

Step 3 - Error handling. Clusters are navigated based on the type of errors in Step 2, which are shown in Table 2. Accordingly, if the quality of the summary is `Type 1`, it is shifted to the next steps. For `Type 2`, the summary is carried back to step 1 and the annotator has to edit this based on the supervisor's previous review. `Type 3` means that the annotator and supervisor have to discuss the severe mistake(s). After that, this cluster is sent back to step 1 and the annotator generates the summary again. After 3 times of reviewing and annotating, the summary is discarded and the related cluster is given to a new annotator.

Table 2: Actions corresponding to each quality level

Types	Required members	Corresponding action
1	–	Passing to the next steps
2	Annotator	Returning to step 1 and editing the related summary based on the previous review
3	Annotator Supervisor	Having face-to-face discussions about mistake(s) Returning to step 1

Step 4 - Expert curation. After the supervisor approved the summary draft, it underwent a final review and refinement process by an expert. This stage focused on identifying and correcting any remaining errors. The revised summaries were then meticulously polished to ensure they were free from mistakes. Subsequently, they were paired with their respective clusters to form the Abmusu dataset.

3.3. Data description

The Abmusu dataset is divided into three parts: the training set consists of 621 documents (200 clusters), the validation set includes 304 documents (100 clusters), and the test set contains 914 documents (300 clusters). Figure 3 illustrates the distribution of categories within both the training/validation set and the test set.

Table 3 illustrates a representative sample from our dataset on the Education category. The input data comprises three distinct documents representing news articles. The output consists of a concise summary, typically spanning 3 to 6 sentences, aligning with the content of the original articles.

Figure 4 displays the statistics regarding the number of documents per cluster. Most clusters contain three documents, while the smallest number of clusters have five documents. This is due to the removal of duplicate documents during the filtering process.

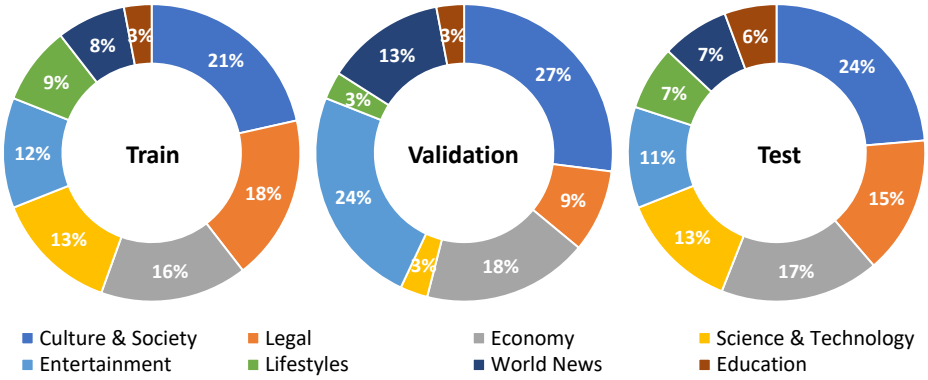

Figure 3: The data statistics by categories

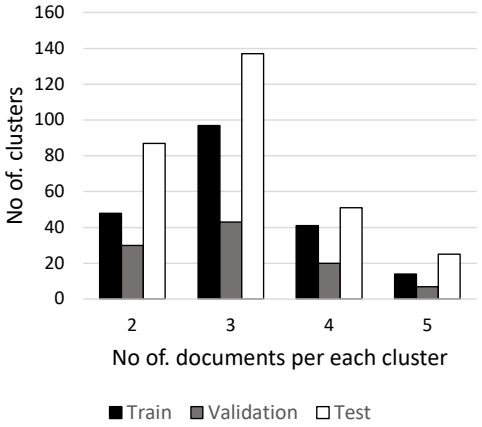

Figure 4: Statistics about the number of documents in each cluster

Table 4 and Table 5 describe the statistics of the Abmusu dataset in detail at the token and the sentence level. The compression ratio of the Abmusu dataset is $\sim 9\%$, the manually created summaries often contain 4 to 6 sentences.

Figure 5 displays the statistical information related to the number of unique tokens in raw texts, summaries, and anchor texts. It is noteworthy that the summaries and anchor texts have a significant overlap with raw texts in terms of the number of tokens. These tokens are the keywords that represent the events and topics discussed in the articles.

To assure progress and quality of annotating, statistics on three types of quality are updated in real-time. Figure 6 shows the final statistics across categories. *Science and technology* have the best ratio of passing documents because documents are usually clear and coherent. The lowest ratio of passing documents belongs to the *Culture, Society*, and *Legal* categories. The *Culture, Society*, and *Legal* documents are intended to summarize political events or mention plenty of related laws.

Table 6 presents some summarization datasets in English and Vietnamese. Some datasets have a considerable number of documents because the summaries are generated automatically from sources like news introductions and abstract paragraphs in scientific articles. While

Table 3: One example in the dataset

Data Types		Content
Single document 1	Title	Lời dặn dò tâm huyết của thầy cô trước khi học sinh bước vào kỳ thi THPT *Translation: Teachers' enthusiastic advice to students entering the National high school exam.*
	Anchor text	(...) kỳ thi quan trọng này, tâm lý các thí sinh không tránh khỏi lo lắng (...) *Translation: (...) this crucial exam, candidates inevitably feel anxiety. (...)*
	Raw text	(...) với môn thi đầu tiên là môn Ngữ văn (...) Các em học sinh hãy nắm chắc cấu trúc đề, ôn thật kĩ từng dạng câu hỏi theo chuyên đề giống với đề minh họa của Bộ giáo dục, và chú ý độ khó của từng dạng bài (...) Dạng bài đọc hiểu và đọc điền vốn là 2 dạng (...) môn Tiếng Anh. *Translation: (...) with the first exam being Literature (...) Students, make sure to grasp the exam structure, thoroughly review each question type according to the illustrated examples provided by the Ministry of Education, and pay attention to the difficulty level of each type of question (...) The comprehension and filling test are two question formats (...) English exam (...)*
Single document 2	Title	Giáo viên "bật mí" kinh nghiệm làm bài thi tốt nghiệp đạt điểm cao *Translation: Teachers share tips to achieve high scores in the National high school exam.*
	Anchor text	(...) 1 triệu thí sinh cả nước sẽ bắt đầu kỳ thi tốt nghiệp THPT (...) *Translation: One million candidates nationwide will begin the National high school exam.*
	Raw text	Các giáo viên bộ môn giàu kinh nghiệm đã có những chia sẻ, "mách nước" giúp học sinh bình tĩnh, có chiến lược ôn tập và kỹ năng làm bài đạt kết quả cao trong kỳ thi sắp tới. (...) cũng chia sẻ với thí sinh cách thức làm bài thi trắc nghiệm môn Toán (...) Một số điểm cần nhớ khi làm bài Khoa học xã hội, thí sinh cần phân tích và xử lý nhanh (...) song câu dễ phải làm nhanh và chắc (...) câu lạ và khó xử lí sau (...) *Translation: Experienced teachers have shared valuable advice to help students stay calm, develop study strategies, and the skills needed to achieve high results in the upcoming exam. (...) for multiple-choice questions in the Math exam (...) When doing Social Science exam, candidates need to analyze and process quickly, however, easy questions should be answered quickly and confidently (...) difficult ones require more careful handling (...)*
Single document 3	Title	"Chiến thuật" đạt điểm cao thi tốt nghiệp THPT năm 2022 *Translation: Strategies to achieve high scores in the 2022 National high school exam.*
	Anchor text	(...) thầy, cô giáo đã đưa ra những bước cần lưu ý để giúp thí sinh đạt được kết quả tốt (...) *Translation: (...) the teachers give notes to help candidates achieve the good results (...)*
	Raw text	Các thầy, cô giáo đã có những lời dặn dò tâm huyết để giúp các em đạt được điểm cao trong các môn thi (...) chiến thuật làm bài môn Toán, Theo (...) dễ trước, khó sau: Mặc dù đề thi đã được sắp xếp từ dễ đến khó, (...) Nháp cẩn thận, khoanh vùng rõ ràng: Nháp xong một câu thì cách ra một chút (...) Làm bài tiếng Anh hiệu quả về mặt thời gian (...) *Translation: The teachers have given enthusiastic advice to help students achieve high scores in the exams (...) Math strategies, according to (...) easy first, difficult later: Although Even though the exam questions are arranged from easy to difficult, (...) Draft carefully, and delineate: After drafting a question, prepare space (...) Optimizing the time in English test (...)*
Summary		Các giáo viên bộ môn giàu kinh nghiệm đã có những chia sẻ giúp học sinh đạt kết quả cao trong kỳ thi sắp tới. (...) Đối với môn Ngữ văn, cần nắm chắc cấu trúc đề, dạng câu hỏi theo chuyên đề (...). Chiến thuật làm môn Toán là: Dễ trước, khó sau, nháp cẩn thận, khoanh vùng rõ ràng (...). Đối với môn tiếng Anh, cần tối ưu hóa điểm số dựa trên mục tiêu (...) làm bài Khoa học xã hội, câu dễ cần làm nhanh và chắc, câu khó làm sau (...). *Experienced teachers have shared tips to help students achieve high results in the upcoming exams. (...) For Literature, the students need to firmly grasp the topic structure and question types according to the topic (...). The strategies for doing Math are Easy first, difficult later, draft carefully, and clearly delineate (...). For English, it is necessary to optimize scores based on the goal of (...) taking the Social Sciences test, easy questions need to be done quickly and carefully and difficult questions should be done later (...).*
Category		Giáo dục *Translation: Education*

The content in some text has been shortened by replacing with (...), Translations are not included in the dataset

summaries produced manually frequently contain fewer documents, the quality control of the summaries is strictly maintained. Compared to manual datasets, the Abmusu dataset has a large number of clusters and documents, which requires the effort of creating individual summaries for each larger cluster. In Vietnamese datasets, VNDS is created by using the introduction paragraphs as single-document summaries. ViMs and VietnameseMDS provide two manual multi-document summaries by two annotators per cluster, which can make

Table 4: Average statistics and compression ratio at the token level

Aspects	Training	Validation	Test
Average			
Documents per Cluster	3.11	3.04	3.05
Tokens per Cluster	1924.75	1815.41	1762.40
Tokens per Raw text	619.88	597.17	578.46
Tokens per Anchor text	41.65	35.58	40.33
Tokens per Summary	168.48	167.68	153.05
Compression ratio			
Multi-document Summary	0.09	0.09	0.09

Table 5: Average statistics and compression ratio at the sentence level

Aspects	Training	Validation	Test
Average			
Sentences per Cluster	66.93	60.69	61.07
Sentences per Raw text	21.56	19.96	20.04
Sentences per Anchor text	1.72	1.27	1.57
Sentences per Summary	4.82	4.94	4.93
Compression ratio			
Multi-document Summary	0.07	0.08	0.08

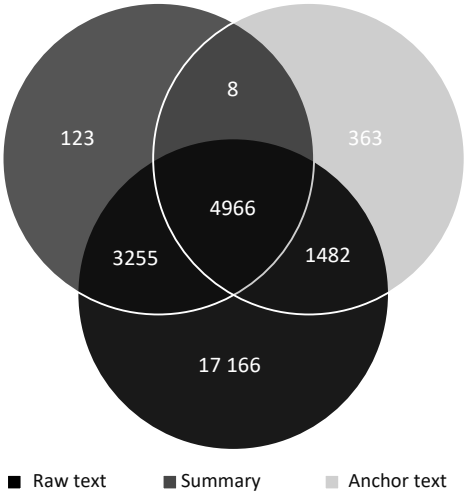

Figure 5: Statistics on the number of unique tokens

it difficult to train the model and compare the results of related models. With Abmusu construction pipelines, we create only one unified summary for each cluster.

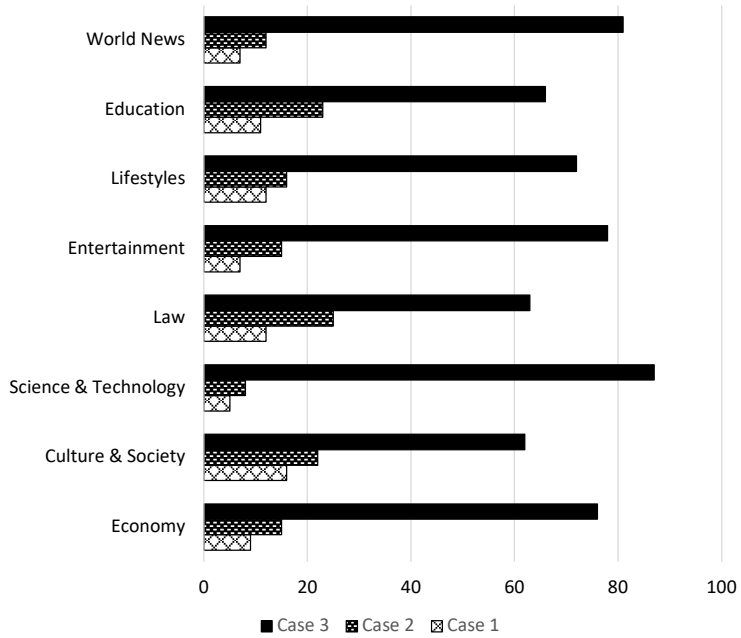

Figure 6: Final statistics of three quality types across 8 categories

Table 6: The statistics about some summarization datasets

Dataset	Language	Multi	Single	No. of cluster	No. of docs
DUC (2004)	English	x		50	500
Multi-News	English	x		56215	156270
BigPatent	English		x	-	1.3 million*
MEDIQA-AnS	English	x	x	156	348
VNDS	Vietnamese		x	-	150704
VietnameseMDS	Vietnamese	x		199	628
ViMs	Vietnamese	x		300	1945
Abmusu	Vietnamese	x		600	1839

Multi (Single) means the dataset provides Multi (Single) summaries

* *means approximate values are denoted, - means the dataset does not provide this information*

4. SHARED TASK RESULTS

4.1. Data format and submission

Each data example includes the title, anchor text, and body text of all single documents in a cluster. Each cluster also has a category tag and a manually created summary. The provided test set for the participated team is formatted similarly to the training and validation data, but without the manually created summary. The evaluation was performed on the AIhub [g] platform for 7 days. Test data was divided into two parts: Public Test and Private Test, each containing 50% of the test data. The Private Test was opened 4 days after the

[g] http://aihub.ml/

Public Test. Each team is allowed to submit a maximum of 35 submissions to the Public test (5 per day) and 5 submissions to the Private Test (not limited per day).

4.2. Evaluation metrics

The official evaluation measures are the `ROUGE-2` scores and `ROUGE-2 F1 (R2-F1)` is the main score for ranking. `ROUGE-2` scores are used in various summarization competitions and proposed models to evaluate and compare performance between models [6]. `ROUGE-2 Recall (R2-R)`, `Precision (R2-P)` and `R2-F1` between predicted summary and reference summary are calculated

$$\texttt{R2-P} = \frac{|\texttt{Matched bigrams}|}{|\texttt{Predicted summary bigrams}|}, \tag{2}$$

$$\texttt{R2-R} = \frac{|\texttt{Matched bigrams}|}{|\texttt{Reference summary bigrams}|}, \tag{3}$$

$$\texttt{R2-F1} = \frac{2 \times \texttt{R2-P} \times \texttt{R2-R}}{\texttt{R2-P} + \texttt{R2-R}}. \tag{4}$$

Besides ROUGE-2 at the bigram level, we also provide other metrics: ROUGE-1 at the unigram level and ROUGE-L at the Longest Common Sequence (LCS) level [14].

4.3. Baselines

Four summarization baselines are built to benchmark our dataset and show how it can be used to access automatically generated summaries, including:

- **Ad-hoc baseline**: Most Vietnamese news is written in an explanatory or inductive style. So, we concatenate the first and last sentences of each component document in each cluster to form the summary.
- **Anchor text baseline**: The anchor texts of all single documents in each cluster are concatenated to create the summary. This test will help test the hypothesis that a summary can be speculated from the introduction.
- **Extractive baseline**: The summary is created by the extractive summarization model which uses Lexrank [15] in the single-document summarization phase and MMR [16] in the multi-document summarization phase. Firstly, Lexrank is a graph-based method that generates single document summaries for each document. The single document summaries are then concatenated together. Finally, MMR is used to remove duplicate sentences.
- **Abstractive baseline**: The summary is created by ViT5 [17].

4.4. Participants

There are 46 registered teams from research groups in domestic and international Universities (VNU-HUS, VNU-UET, HUST, PTIT, etc.) and industries (Viettel, VinGroup, CMC, TopCV, VCCorp, etc). In which, 28 teams submitted the data agreement, and 16 teams participated officially by submitting at least 1 run on the evaluation platform. Participant

teams can use all possible tools and resources to build models. Participated teams made a total of 287 submissions. Post-challenge panels [h] are now opened on AIHUB for supporting research improvements.

4.5. Results

Table 7 shows the results of the private test were considered as the official results to rank the team in the AbMuSu shared task. An interesting finding from the evaluation is that the ad-hoc baseline achieved unexpectedly high results, ranking at 6th place. This can be explained by the fact that many news articles are written in an explanatory or inductive style, where the first and last sentences often contain important information. On the other hand, the extractive baseline model performed much better, ranking in 5th place, compared to the anchor text baseline model which ranked in 18th place. This is contrary to the assumption that anchor text can be considered as a simple summary of the news text. As for the abstractive baseline model, it only utilized the ViT5 model without any parameter tuning, resulting in a low ranking at 19th place.

The proposed models for the shared task on multi-document summarization followed two main approaches: (1) abstractive summarization and (2) hybrid approach, which first selects important sentences in an extractive phase and then generates the summary in an abstractive phase. Three leading teams–LBMT, the Coach, and CIST AI–adopted the hybrid approach.

The extractive phase saw a variety of methods employed. The *LBMT* team utilized similarity scoring techniques such as TF-IDF and Cosine, along with graph-based methods like TextRank and PageRank. In contrast, *The Coach* team implemented a long short-term memory (LSTM) model to predict an important score for each sentence. The *CIST AI* team combined the LexRank technique with a multi-layer perceptron (MLP) sentence classification model. Following the extractive phase, these teams fine-tuned pre-trained models in the Vietnamese language, such as BARTpho and ViT5. Overall, the hybrid model demonstrated high performance. However, it was noted that the hybrid model could yield low performance if the extractive phase inadequately filters important sentences, thereby diminishing the input quality for the abstractive phase.

The *FinalYear* team proposed a model that relied solely on extractive methods for creating summaries. They employed graph-based methods to create single-document summaries and used Maximal Marginal Relevance (MMR) to generate multi-document summaries.

In conclusion, while the hybrid approach showed significant potential for high performance, it is crucial to ensure the extractive phase effectively selects relevant sentences to maintain the quality of the final abstractive summary. The use of various innovative methods in both the extractive and abstractive phases highlights the diverse strategies employed by the participating teams.

5. CONCLUSION

In summary, the VLSP 2022 - AbMuSu shared task was launched to advance research in the field of abstractive multi-document summarization. By providing a standardized test-bed for comparing various summarization approaches, the task has the potential to significantly

[h] http://aihub.ml/competitions/341

Table 7: The official results of the Private Test

Rank	User	R2-F1	R2-P	R2-R	R1-F1	R1-P	R1-R	RL-F1	RL-P	RL-R
1	LBMT	**0.3035 (1)**	0.2298 (11)	**0.4969 (1)**	**0.5067 (1)**	0.4076 (16)	**0.7147 (1)**	**0.4809 (1)**	0.3868 (15)	**0.6780 (1)**
2	The coach	0.2937 (2)	0.2284 (12)	0.4463 (2)	0.4962 (2)	0.4072 (17)	0.6676 (4)	0.4701 (2)	0.3857 (16)	0.6326 (4)
3	CIST AI	0.2805 (3)	0.2629 (6)	0.3192 (6)	0.4876 (4)	0.4635 (6)	0.5352 (9)	0.4541 (4)	0.4314 (6)	0.4988 (7)
4	TheFinalYear	0.2785 (4)	0.2272 (13)	0.4040 (4)	0.4956 (3)	0.4221 (15)	0.6409 (5)	0.4612 (3)	0.3929 (14)	0.5964 (5)
5	NLP HUST	0.2689 (5)	0.2773 (4)	0.2829 (12)	0.4732 (6)	0.4903 (5)	0.4836 (12)	0.4373 (5)	0.4537 (5)	0.4465 (12)
6	*Extractive baseline*	*0.2625 (6)*	*0.2464 (7)*	*0.3174 (8)*	*0.4772 (5)*	*0.4582 (9)*	*0.5391 (8)*	*0.4339 (6)*	*0.4164 (9)*	*0.4905 (9)*
7	*Ad-hoc baseline*	*0.2611 (7)*	*0.2634 (5)*	*0.2947 (10)*	*0.4627 (8)*	*0.4601 (8)*	*0.5053 (11)*	*0.4273 (8)*	*0.4257 (7)*	*0.4659 (11)*
8	VNU Brothers	0.2544 (8)	0.3030 (2)	0.2406 (14)	0.4595 (9)	0.5315 (2)	0.4312 (17)	0.4194 (12)	0.4850 (2)	0.3937 (17)
9	FCoin	0.2544 (8)	0.2307 (9)	0.3027 (9)	0.4697 (7)	0.4302 (12)	0.5411 (7)	0.4296 (7)	0.3941 (13)	0.4938 (8)
10	vts	0.2448 (9)	0.2114 (15)	0.3188 (7)	0.4516 (12)	0.4048 (18)	0.5438 (6)	0.4208 (10)	0.3768 (18)	0.5074 (6)
11	Blue Sky	0.2412 (10)	0.2384 (8)	0.2610 (13)	0.4588 (10)	0.4604 (7)	0.4761 (13)	0.4194 (12)	0.4205 (8)	0.4358 (13)
12	HUSTLANG	0.2361 (11)	0.2880 (3)	0.2157 (17)	0.4360 (16)	0.5176 (4)	0.3981 (18)	0.4000 (15)	0.4750 (3)	0.3651 (18)
13	SGSUM	0.2322 (12)	0.2106 (16)	0.2896 (11)	0.4575 (11)	0.4279 (13)	0.5282 (10)	0.4235 (9)	0.3954 (12)	0.4897 (10)
14	vc-datamining	0.2304 (13)	0.1663 (20)	0.4371 (3)	0.4496 (14)	0.3450 (20)	0.7036 (2)	0.4201 (11)	0.3218 (20)	0.6590 (2)
15	TCV-AI	0.2288 (14)	0.1687 (19)	0.3976 (5)	0.4502 (13)	0.3485 (19)	0.6813 (3)	0.4190 (13)	0.3245 (19)	0.6340 (3)
16	Team Attention	0.2131 (15)	0.2159 (14)	0.2265 (16)	0.4274 (18)	0.4251 (14)	0.4514 (15)	0.3848 (18)	0.3835 (17)	0.4056 (15)
17	Cyber Intellect	0.2116 (16)	0.2085 (17)	0.2270 (15)	0.4464 (15)	0.4468 (10)	0.4627 (14)	0.4028 (14)	0.4030 (10)	0.4177 (14)
18	HHH	0.1919 (17)	0.1915 (18)	0.2076 (18)	0.4228 (19)	0.4350 (11)	0.4336 (16)	0.3888 (16)	0.4005 (11)	0.3984 (16)
19	*Anchor text baseline*	*0.1886 (18)*	*0.2306 (10)*	*0.1734 (19)*	*0.4321 (17)*	*0.5210 (3)*	*0.3900 (19)*	*0.3869 (17)*	*0.4659 (4)*	*0.3498 (19)*
20	*Abstractive baseline*	*0.1497 (19)*	***0.3061 (1)***	*0.1025 (20)*	*0.3226 (20)*	***0.5801 (1)***	*0.2299 (20)*	*0.2895 (19)*	***0.5205 (1)***	*0.2065 (20)*

The number highlighted in bold is the highest result in each column.
The number in the bracket () is the corresponding rank of a score. The baseline results are shown in italics.

contribute to future research. The carefully constructed AbMuSu dataset is expected to make notable contributions to related works. The task garnered attention from the research community, with participants utilizing various range of advanced technologies and resources to present exciting and promising results, which can serve as useful benchmarks for future research. We are pleased to conclude that the VLSP 2022 - AbMuSu shared task was executed successfully and has the potential to make significant contributions to the Vietnamese text mining and natural language processing communities.

ACKNOWLEDGMENT

This research has been done under the research project QG.22.61 “Research and Development of Vietnamese Multi-document Summarization Based on Advanced Language Models” of Vietnam National University, Hanoi. Quoc An Nguyen was funded by the

Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.ThS.002.

REFERENCES

[1] K. Ježek and J. Steinberger, "Automatic text summarization (the state of the art 2007 and new challenges)," in *Proceedings of Znalosti*, pp. 1–12, 2008.

[2] R. D. Lins, R. F. Mello, and S. Simske, "Doceng'19 competition on extractive text summarization," in *Proceedings of the ACM Symposium on Document Engineering 2019*, pp. 1–2, 2019.

[3] A. B. Abacha, Y. M'rabet, Y. Zhang, C. Shivade, C. Langlotz, and D. Demner-Fushman, "Overview of the mediqa 2021 shared task on summarization in the medical domain," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, pp. 74–85, 2021.

[4] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: A brief survey," *International Journal of Advanced Computer Science and Applications (ijacsa)*, vol. 8, no. 10, 2017.

[5] R. Khan, Y. Qian, and S. Naeem, "Extractive based text summarization using k-means and tf-idf," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 3, p. 33, 2019.

[6] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, pp. 1–66, 2017.

[7] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*, pp. 11328–11339, PMLR, 2020.

[8] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[10] A. R. Fabbri, I. Li, T. She, S. Li, and D. Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, 2019.

[11] M. Savery, A. B. Abacha, S. Gayen, and D. Demner-Fushman, "Question-driven summarization of answers to consumer health questions," *Scientific Data*, vol. 7, no. 1, pp. 1–9, 2020.

[12] V. H. Nguyen, T. C. Nguyen, M. T. Nguyen, and N. X. Hoai, "VNDS: A Vietnamese dataset for summarization," in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 375–380, IEEE, 2019.

[13] N. T. Tran, M. Q. Nghiem, N. T. Nguyen, N. L. T. Nguyen, N. Van Chi, and D. Dinh, "ViMs: a high-quality Vietnamese dataset for abstractive multi-document summarization," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 893–920, 2020.

[14] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, pp. 74–81, 2004.

[15] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

[16] J. Goldstein and J. G. Carbonell, "Summarization:(1) using mmr for diversity-based reranking and (2) evaluating summaries," in *TIPSTER Text Program Phase III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pp. 181–195, 1998.

[17] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained text-to-text transformer for Vietnamese language generation," *arXiv preprint arXiv:2205.06457*, 2022.

Received on April 21, 2023
Accepted on July 20, 2024