Interspeech 2024
1-5 September 2024, Kos, Greece

Real-time Speech Summarization for Medical Conversations

Khai Le-Duc[*1,2,6], *Khai-Nguyen Nguyen*[*3], *Long Vo-Dang*[4], *Truong-Son Hy*[5,6]

[1]University of Toronto, Canada [2]University Health Network, Canada
[3]College of William and Mary, United States [4]University of Cincinnati, United States
[5]Indiana State University, United States [6]FPT Software AI Center, Vietnam

duckhai.le@mail.utoronto.ca, knguyen07@wm.edu, TruongSon.Hy@indstate.edu

Abstract

In doctor-patient conversations, identifying medically relevant information is crucial, posing the need for conversation summarization. In this work, we propose the first deployable real-time speech summarization system for real-world applications in industry, which generates a local summary after every N speech utterances within a conversation and a global summary after the end of a conversation. Our system could enhance user experience from a business standpoint, while also reducing computational costs from a technical perspective. Secondly, we present VietMed-Sum which, to our knowledge, is the first speech summarization dataset for medical conversations. Thirdly, we are the first to utilize LLM and human annotators collaboratively to create gold standard and synthetic summaries for medical conversation summarization. Finally, we present baseline results of state-of-the-art models on VietMed-Sum. All code, data (English-translated and Vietnamese) and models are available online.

Index Terms: speech recognition, speech summarization, AI for healthcare, LLM

1. Introduction

In real-world conversations, the volume of information grows significantly in tandem with speaking rates, leading to information overload. Remembering every detail discussed, especially medical information, is beyond human capability. Yet, doctors and patients frequently make decisions by prioritizing crucial information and its significance. Consequently, the adoption of real-time speech summarization (RTSS) system is emerging as an effective approach to tackle this issue.

Compared to pre-recorded speech summarization, RTSS research has very little literature [1]. Besides, in industry settings, to the best of our knowledge, there is currently no RTSS system deployed for real-world applications[1].

In terms of medical domain, according to the latest survey by [2] and to the best of our knowledge, there is only one publicly available dataset for medical conversation summarization [3]. This dataset consists of written text in the Chinese language and was crawled from an online healthcare service provider. However, no speech summarization dataset for medical conversations is publicly available.

RTSS systems proposed by [1] constantly update and revise the current summary state in the course of a dialogue using additional components, such as flexible recognizer of utterance units, utterance lookahead-er, and information overrider. While these additions increase inference and training time, they also contribute to increased deployment and maintenance complexity. Furthermore, from a business standpoint, these RTSS systems can degrade user experience as users are unaware of the exact moment when a comprehensive summary concludes.

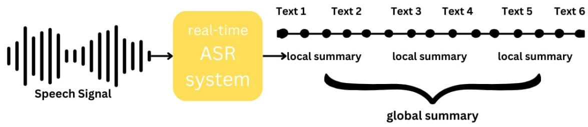

Figure 1: *Visualization of our proposed RTSS*

To tackle all the problems above, we propose a new approach to a RTSS system for medical conversations. Our contribution are as follows:

- We propose the first deployable RTSS system for real-world applications.
- We introduce *VietMed-Sum* - the first speech summarization dataset for real-world medical conversations, to the best of our knowledge.
- We conduct the first attempt to leverage ChatGPT and human annotators colaboratively to create gold standard and synthetic summaries for medical conversations.
- We present baseline results on our dataset using various state-of-the-art models.

All code, data (English-translated and Vietnamese) and models are published online[2,3].

2. Real-time Speech Summarization System

2.1. Previous Designs

RTSS system proposed by [1] has 3 major additional components: flexible recognizer of utterance units, utterance lookahead-er, and information overrider. Flexible recognizer of utterance units automatically split real-time Automatic Speech Recognition (ASR) transcript into segments with random lengths, while utterance lookahead-er seeks additional context from subsequent words generated by ASR, and information overrider continuously updates the summary in response to the latest contextual changes. We conducted surveys and gathered feedback from engineers and found that these extra components not only extend both inference and training time but also complicate RTSS systems, making it challenging for engineers to deploy and maintain them effectively. Furthermore, the

[*]Equal contribution
[1]In most papers, the term "real-time summarization" refers to the summarization of real-time news or events, instead of generating summaries in real-time.

[2]https://github.com/leduckhai/MultiMed
[3]https://github.com/HySonLab/VietMed-Sum

 10.21437/Interspeech.2024-2250

summary is constantly updated after each utterance generated by the ASR system. This results in increased computational costs when compared to a scenario where a solid summary is generated after a set of utterances. The continuously updated summary creates confusion as users are unable to keep track, given the uncertainty of when a summary is completed.

2.2. Our Design

In contrast, our approach is much simpler. Our design generates a local summary after every N utterances of speech within a conversation and a global summary after the end of a conversation. Every local summary is generated using the corresponding local context of N utterances, without the need to continuously update the new context generated by real-time ASR utterances. Meanwhile, the global summary serves as an "overrider" using the context of the entire conversation.

2.3. Balance for System Delay

RTSS system by [1] generates a summary with a delay of one utterance. A large number of delayed utterances results in longer waiting time for users to receive the generated summary. Conversely, a low number of delayed utterances means that the context necessary for accurate summaries is missing, making summarization unnecessary. After analyzing the context within the *VietMed* corpus [4] and conducting our internal user survey, we found that setting N = {4, 5} (or a maximum of around 30 seconds) strikes a suitable balance. This ensures that each summary includes an adequate amount of context without keeping users waiting excessively.

3. Data

3.1. Labeling strategy

We used GPT-3.5 Turbo[4] (or ChatGPT) to generate summaries for every transcript in our dataset, which we refer to as GPT-annotated summaries. We then split the dataset into two subsets: **Gold standard (*GOLD*)** set and the **Synthetic (*SYN*)** set. On the *GOLD* set, we performed **human editing** where the human annotator edit the GPT summary according to the annotation guideline, while on the *SYN* set, we did not. More information on GPT annotation is on section 4.

3.2. Data Collection

Real-world dataset (REAL): We choose the *VietMed* dataset [4], a real-world medical ASR dataset in Vietnamese, for annotating summaries. This choice is driven by the fact that *VietMed* currently stands as the world's largest and most generalizable publicly-available medical ASR dataset.

Simulated dataset (SIM): To make the dataset more generalizable and to extend the scale of the existing *VietMed* dataset, we used extra medical text data[5]. We simulated real-world conversations by imitating the speaking style found in the *VietMed* corpus. This includes incorporating hesitations, disfluencies, and stuttering words at a rate similar to that of *VietMed* utterances. Pseudo Python code for simulation is in the Appendix.

Our *GOLD* set contains gold standard data from *REAL* and *SIM*, while *SYN* contains GPT summaries from the extra medical text mentioned above.

[4]https://platform.openai.com/docs/models/gpt-3-5-turbo
[5]https://github.com/duyvuleo/VNTC

3.3. Annotation Process and Data Quality Control

Details are in Data Annotation section in the Appendix.

3.4. Data Statistics

	Gold-standard Data			Synthethic	All
	Real-world		Simulated		
	Local	Global	Local	Local	
Train	837	382	560	18981	20760
Dev	874	624	70	0	1568
Test	1192	767	70	0	2029
Total	2903	1773	700	18981	24357

Table 1: *Data statistics for VietMed-Sum. Full table is in the Appendix.*

Table 1 shows the statistics of our dataset. To construct our *VietMed-Sum* dataset, we keep the original split of *REAL* by [4] as 5-5-6 hours for the corresponding train-dev-test set. We split our *SIM* set with a ratio of 8:1:1 for the corresponding train-dev-test, while the entire *SYN* is used for training.

Acquisition and annotation of medical dataset is challenging and costly, resulting in medical summarization datasets typically being smaller compared to those in the general domain. Compared to other public medical written text summarization dataset, such as MeQSum corpus of summarized consumer health questions [5], our dataset has 23 times more summaries. Besides, compared to the Chinese medical text summarization dataset by [3], ours is half the size.

3.5. English-translated *VietMed-Sum*

We also introduce *VietMed-Sum-en*, the English version of *VietMed-Sum* which was translated using Google Translate[6]. Results are in the Appendix.

4. GPT for Annotation

4.1. Motivation

To the best of our knowledge, existing prominent Vietnamese summarization datasets, such as VietNews [6] and FAQSum [7], utilize their titles and abstracts as summaries. Our dataset, however, lacks these pre-made summaries which would traditionally require human annotation. However, finding high-quality annotators for either low-resource languages like Vietnamese or medical domain is hard [8].

In recent year, there has been an increasing focus on utilizing Large Language Models (LLMs) for annotation [9, 10, 11]. Experimental results from [12] showed that fully GPT-3 labeling can ourperform fully human labeling in low-budget settings. GPT has shown to have adequate medical knowledge [13]. Furthermore, it achieves reasonable performance as an annotator in sequence generation tasks in Vietnamese [10].

4.2. GPT Setup

We used GPT-3.5 Turbo to generate GPT summaries. Full setup details are in the Appendix.

4.3. Cost-Efficiency Evaluation

Following the design from [10, 12], we evaluated the performance difference of human-annotated summaries versus GPT-

[6]https://translate.google.com/

Cost	Method	R-1	R-2	R-L
\$2.5	250 Human Summaries	60.73	45.35	55.67
	6k GPT Summaries	56.69	40.13	50.61
\$5	500 Human Summaries	62.85	47.19	57.50
	6k GPT → 250 Human-reuse	**62.88**	**47.81**	**57.67**
	6k GPT → 250 Human-new	**63.45**	**47.65**	**57.49**

Table 2: *ROUGE on FAQSum on two budgets: \$2.5 and \$5. ViT5 is trained on the data from each method. GPT → Human-reuse refers to two-step finetuning on the 250 human summaries from the \$2.5-budget setting while GPT → Human-new refers to two-step finetuning on new 250 human summaries.*

annotated summaries under a fixed budget. In particular, we evaluated ViT5 performance on FAQSum, a medical summarization dataset, trained with human-annotated summaries versus with GPT summaries on fixed budgets of \$2.5 and \$5, corresponding to 250 and 500 human-annotated summaries accordingly[7]. At \$2.5, GPT can generate around 6000 summaries[8].

The experimental results from the \$2.5-budget setting in Table 2 demonstrate the importance of human annotation in the Vietnamese medical summarization task. Since summaries are heavily influenced by the annotators' medical knowledge and writing style, we hypothesize that GPT summaries still provide useful medical knowledge but require additional training on human summaries for writing style transfer.

To verify our hypothesis, we devise the the \$5-budget setting where we fine-tuned ViT5 on GPT summaries, then on human summaries, and compared this two-step process with fine-tuning only on human summaries. We also doubled the number of human summaries for fair comparison. Experimental results show that the two-step process achieves slightly better results while costing significantly less time. We hypothesize that fine-tuning on GPT summaries helps provide medical knowledge and fine-tuning on human summaries helps the model aligns its output's text style.

Method	R-1	R-2	R-L
6k Human Summaries	65.80	50.82	60.83
6k GPT Summaries	56.69	40.13	50.61
6k GPT → 250 Human-reuse	62.88	47.81	57.67
6k GPT → 250 GPT	56.93	39.81	50.14

Table 3: *ROUGE scores on FAQSum on 6k human summaries, 6k GPT summaries and the previously mentioned two-step fine-tuning process of ViT5.*

Table 3 further support our argument. When we perform two-step finetuning on ViT5, the performance gap between human summaries and GPT summaries is significantly closed. As such, our labeling strategy involves creating the *SYN* set and refining the *GOLD* set to balance between time-consumption, cost and performance.

4.4. Human Evaluation

We further investigate the characteristics of the GPT summaries empirically. In this experiment, we sample 50 arbitrary transcripts of varying lengths from *VietMed-Sum* and let two annotators independently summarize them, one with GPT summaries as references (human editing) and one without.

Human Editing Time-Efficiency: We quantify the overall improvement in time taken to annotate the transcripts between human editing and manually writing the summary. We found that annotators who perform human editing is approximately 70% faster than those who perform manual summary writing. As such, we perform human editing on the *GOLD* set.

Hallucination: To make sure the GPT summary is factually aligned with the transcript, we further cross-check the GPT summaries with the transcripts. We found that (1) the GPT summaries sometimes contain details implied but not explicitly mentioned in the transcript and (2) GPT is easily confused by transcripts with a lot of spoken language characteristics (i.e. hesitations, disfluencies and stuttering words). This typically leads to hallucinate and we found that around 25% of our samples have hallucination. Furthermore, the GPT summaries are often more lengthy with an average compression rate of 30%, higher than required in the guideline. As such, we ask the annotators to strictly adhere to the annotation guideline when editing.

5. Experimental Setup

5.1. Evaluation Metrics

We use ROUGE [14], a metric commonly used for summarization, to evaluate our models. More details are in the Appendix.

5.2. Baseline Summarization Models

We employed $\text{BARTpho}_{\text{syllable}}$ and $\text{BARTpho}_{\text{word}}$ [15], ViT5, ViT5-vietnews [16] (ViT5 fine-tuned on Vietnews summarization dataset [17]), and ViPubmedT5 [18] models in our experiments. More information about the models is in the Appendix.

5.3. Downstream Tasks

Summarization on Human Transcript: We train the models on the abstractive summarization task on VietMed-Sum. To evaluate their performance, we calculate their ROUGE scores on the local and global summaries in the test set.

Summarization on ASR Transcript: We also evaluated the models' performance from transcripts obtained from audio speech recognition (ASR). We employed the best ASR model on *VietMed* from [4] with a Word-Error-Rate (WER) of 28.8% to generate the ASR transcripts for summarization. This creates noisier text which is more challenging for the baseline models.

6. Experimental Results

We report the ROUGE of our baseline models on the Global summaries and Local summaries subset from *GOLD*.

6.1. Gold Standard Data Summarization

Table 4 shows the ROUGE scores on our *GOLD* test set for the baseline models fine-tuned on the combination of *GOLD* local and global summaries. ViT5, ViPubmedT5, and ViT5-vietnews consistently outperforms the BARTpho variants.

Results from Table 5 shows the models have a noticeable drop in performance. On the local summaries, the ROUGE scores from the BARTpho variants are much lower than that of the ViT5 variants. Conversely, when fine-tuned on the global summaries, both variants of BARTpho performs much better than the ViT5 variants except ViT5-vietnews, probably because

[7] Based on the minimum fee of \$0.01 per assignment on MTurk

[8] With an average of 700 input tokens and 20 output tokens per sample with rate \$0.50 per 1M input tokens and \$1.50 per 1M output tokens

Data GOLD	Global Summaries R-1	 R-2	 R-L	Local Summaries R-1	 R-2	 R-L
$BARTpho_{syllable}$	60.92	40.71	49.38	59.07	38.27	47.69
$BARTpho_{word}$	58.83	39.71	48.07	57.76	37.43	46.87
ViT5	61.65	40.56	49.62	59.95	38.66	48.66
ViT5-vietnews	61.90	41.07	49.61	59.94	38.69	48.25
ViPubmedT5	61.73	40.17	48.81	59.99	38.30	47.67

Table 4: *Experimental results on VietMed-Sum's GOLD test set of each model fine-tuned on local + global summaries of GOLD.*

Data Global	Global Summaries R-1	 R-2	 R-L	Local Summaries R-1	 R-2	 R-L
$BARTpho_{syllable}$	61.69	41.19	49.87	59.13	37.80	47.59
$BARTpho_{word}$	59.45	39.05	47.79	57.8	36.28	46.35
ViT5	58.89	37.2	46.76	56.79	34.97	45.31
ViT5-vietnews	60.67	39.20	48.17	58.64	36.83	46.61
ViPubmedT5	58.57	36.65	45.75	56.83	34.70	44.72
Local	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
$BARTpho_{syllable}$	58.33	39.62	47.49	57.37	37.43	46.41
$BARTpho_{word}$	56.46	38.62	46.20	55.65	36.36	45.13
ViT5	59.82	40.47	48.74	59.11	38.41	47.87
ViT5-vietnews	61.04	41.43	49.57	60.10	39.39	48.52
ViPubmedT5	59.00	38.27	47.36	58.99	37.38	47.11

Table 5: *Results on VietMed-Sum's Global and Local subset of GOLD test set. Each baseline model is fine-tuned on the global summaries (above) and the local summaries (below) of GOLD*

ViT5-vietnews was previously fine-tuned on other Vietnamese abstractive summarization dataset.

6.2. Synthetic Data Summarization

Data SYN	Global Summaries R-1	 R-2	 R-L	Local Summaries R-1	 R-2	 R-L
$BARTpho_{syllable}$	60.37	40.27	48.48	58.68	38.17	47.14
$BARTpho_{word}$	59.64	40.21	48.01	57.50	37.31	46.01
ViT5	61.71	42.36	50.17	59.74	39.83	48.65
ViT5-vietnews	60.78	40.79	49.07	58.18	38.21	47.01
ViPubmedT5	60.58	40.95	48.72	58.44	38.40	47.09
SYN + GOLD	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
$BARTpho_{syllable}$	61.13	41.55	49.83	59.10	39.10	48.06
$BARTpho_{word}$	61.11	42.08	49.66	59.16	39.36	48.12
ViT5	63.23	43.92	51.64	60.9	40.93	49.83
ViT5-vietnews	62.68	43.59	51.58	60.31	40.52	49.32
ViPubmedT5	62.15	42.92	50.45	59.96	40.22	48.82
SYN → GOLD	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
$BARTpho_{syllable}$	62.37	41.87	50.49	60.5	39.78	49.24
$BARTpho_{word}$	60.93	41.80	49.89	59.38	39.43	48.67
ViT5	**64.52**	**45.12**	**52.95**	**62.56**	**42.41**	**51.61**
ViT5-vietnews	**63.34**	**43.29**	**51.58**	**61.73**	**41.20**	**50.23**
ViPubmedT5	61.70	40.13	48.80	59.99	38.31	47.68

Table 6: *Experimental results on VietMed-Sum's GOLD test set of each baseline model fine-tuned on SYN, (SYN + GOLD) and SYN → GOLD. + refers to concatenating the datasets while → refers to two-step fine-tuning. Bolded text refers to two best performing models.*

Table 6 shows the ROUGE scores of each model fine-tuned on *SYN*-only and with *GOLD*. We found that fine-tuning only on *SYN* data did not improve the models' performance. However, when we incorporated *GOLD* into fine-tuning the models either by concatenating the *GOLD* and *SYN* data or by performing two-step fine-tuning *SYN* → *GOLD*, the performance of the models drastically improved compared to training only on *GOLD* on all metrics. This is consistent with our observations from Subsection 4.3.

6.3. ASR Transcript Summarization

Model	Global Summaries R-1	 R-2	 R-L	Local Summaries R-1	 R-2	 R-L
ViT5	58.95	36.82	46.63	56.78	34.43	45.52
ViT5-vietnews	58.22	35.34	46.02	56.48	33.65	45.10

Table 7: *Experimental results on ASR transcripts of the two best performing models from Section 6.2. Full table with all results is in the Appendix.*

We report the results of our baseline models on the ASR transcripts on Table 7. The performance of the models is worse than that on *VietMed-Sum*, which we attribute to the noisy nature of the text generated by the ASR models. Nevertheless, the ROUGE scores remain fairly reasonable, which is proof to our model's robustness.

6.4. Human Evaluation

While ROUGE is commonly used to evaluate the performance of the models, does not measure the fluency and factual alignment of the summaries. As such, we adopt the human evaluation methodology from [19, 20] and report the results in Table 8. Details of experiments are in the Appendix.

7. Conclusion

In this work, we propose a novel RTSS system that generates a local summary after every N utterances within a conversation and a global summary for the entire conversation. Unlike previous works that continuously update the summary after each utterance generated by ASR systems which might be hard for users to follow, our system could improve user experience and lower computational costs. Secondly, we present *VietMed-Sum*, the first speech summarization dataset for medical conversations. Thirdly, our proposed labeling strategy strikes a balance between performance of summarization models, annotation cost, and annotation time (approximately 70% time reduction). We report the use of our proposed synthetic data generated by LLM, which improves models' performance across all metrics. Notable is an average improvement of 2.74 in the R-1 score for ViT5.

Summary	Fluency	Consistency	Relevance	Coherance
ChatGPT	3.8	3.3	5.0	4.2
GOLD	5.0	5.0	5.0	5.0
ViT5	4.0	4.2	4.3	4.2

Table 8: *Results for human evaluation on 50 samples. Scores range from 1 (worst) to 5 (best). GOLD is the baseline which has all scores of 5. ViT is the best model for ROUGE scores.*

8. Acknowledgement

We thank Bao Tran at Chubb Canada and Linh Nguyen from Bucknell University for helping the initial annotation. We appreciate David Thulke at RWTH Aachen University and AppTek GmbH for his precious feedback. We thank Ralf Schlüter at RWTH Aachen University for supporting computing resource to conduct experiments.

9. References

[1] M. Kameyama, G. Kawai, and I. Arima, "A real-time system for summarizing human-human spontaneous spoken dialogues," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996.

[2] R. Jain, A. Jangra, S. Saha, and A. Jatowt, "A survey on medical document summarization," *arXiv preprint arXiv:2212.01669*, 2022.

[3] Y. Song, Y. Tian, N. Wang, and F. Xia, "Summarizing medical conversations via identifying important utterances," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 717–729.

[4] K. Le-Duc, "Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 17 365–17 370.

[5] A. B. Abacha and D. Demner-Fushman, "On the summarization of consumer health questions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2228–2234.

[6] 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)*, 2019, pp. 375–380.

[7] N. Minh, V. H. Tran, V. Hoang, H. D. Ta, T. H. Bui, and S. Q. H. Truong, "ViHealthBERT: Pre-trained language models for Vietnamese in health text mining," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 328–337. [Online]. Available: https://aclanthology.org/2022.lrec-1.35

[8] E. Pavlick, M. Post, A. Irvine, D. Kachaev, and C. Callison-Burch, "The language demographics of Amazon Mechanical Turk," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 79–92, 2014. [Online]. Available: https://aclanthology.org/Q14-1007

[9] M. Li, T. Shi, C. Ziems, M.-Y. Kan, N. F. Chen, Z. Liu, and D. Yang, "Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation," 2023.

[10] J. Choi, E. Lee, K. Jin, and Y. Kim, "Gpts are multilingual annotators for sequence generation tasks," 2024.

[11] S. Latif, M. Usama, M. I. Malik, and B. W. Schuller, "Can large language models aid in annotating speech emotional data? uncovering new frontiers," 2023.

[12] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," *CoRR*, vol. abs/2108.13487, 2021. [Online]. Available: https://arxiv.org/abs/2108.13487

[13] A. Meyer, J. Riese, and T. Streichert, "Comparison of the performance of gpt-3.5 and gpt-4 with that of medical students on the written german medical licensing examination: Observational study," *JMIR Medical Education*, vol. 10, p. e50965, 2024.

[14] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[15] N. L. Tran, D. M. Le, and D. Q. Nguyen, "Bartpho: Pre-trained sequence-to-sequence models for vietnamese," 2022.

[16] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "Vit5: Pretrained text-to-text transformer for vietnamese language generation," 2022.

[17] M.-T. Nguyen, H.-D. Nguyen, T.-H.-N. Nguyen, and V.-H. Nguyen, "Towards state-of-the-art baselines for vietnamese multi-document summarization," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 85–90.

[18] L. Phan, T. Dang, H. Tran, T. H. Trinh, V. Phan, L. D. Chau, and M.-T. Luong, "Enriching biomedical knowledge for low-resource language through large-scale translation," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 3131–3142.

[19] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, "Dialogsum: A real-life scenario dialogue summarization dataset," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 5062–5074.

[20] W. Kryscinski, N. S. Keskar, B. McCann, C. Xiong, and R. Socher, "Neural text summarization: A critical evaluation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[22] 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," 2019.

[23] 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," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[25] A. Roberts, H. W. Chung, A. Levskaya, G. Mishra, J. Bradbury, D. Andor, S. Narang, B. Lester, C. Gaffney, A. Mohiuddin, C. Hawthorne, A. Lewkowycz, A. Salcianu, M. van Zee, J. Austin, S. Goodman, L. B. Soares, H. Hu, S. Tsvyashchenko, A. Chowdhery, J. Bastings, J. Bulian, X. Garcia, J. Ni, A. Chen, K. Kenealy, J. H. Clark, S. Lee, D. Garrette, J. Lee-Thorp, C. Raffel, N. Shazeer, M. Ritter, M. Bosma, A. Passos, J. Maitin-Shepard, N. Fiedel, M. Omernick, B. Saeta, R. Sepassi, A. Spiridonov, J. Newlan, and A. Gesmundo, "Scaling up models and data with `t5x` and `seqio`," 2022.

[26] J. Heek, A. Levskaya, A. Oliver, M. Ritter, B. Rondepierre, A. Steiner, and M. van Zee, "Flax: A neural network library and ecosystem for JAX," 2023. [Online]. Available: http://github.com/google/flax