An empirical study of deep learning approaches toward a multi-task learning framework for Vietnamese text summarization and keyword extraction

Dinh Thu Khanh [1,2,3,4], Hoang Duc Trung[4], Vu Duc Thi[5], Le Minh Tuan[6,*]

and Nguyen Long Giang[1]

[1] Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

[2] Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

[3] Faculty of Information Technology, Electric Power University, Hanoi, Vietnam

[4] Artificial Intelligence Research Center, VNU Information Technology Institute, Vietnam National University, Hanoi, Vietnam

[5] VNU Information Technology Institute, Vietnam National University, Hanoi, Vietnam

[6] Faculty of Technology and Engineering, Trade Union University, Hanoi, Vietnam
khanhdt@epu.edu.vn, ductrunghoang2002@gmail.com,
vdthi@vnu.edu.vn, tuanlm@dhcd.edu.vn, nlgiang@ioit.ac.vn

Abstract. Text summarization and keyword extraction are essential for processing Vietnamese natural language processing (VNLP), a low-resource language with limited datasets and complex features like tonal structure and morphology. Multi-task learning (MTL) offers a promising approach by sharing knowledge across tasks, potentially improving performance in VNLP. This study evaluates single-task models such as TextRank, LexRank, and T5-small for summarization, and Bi-LSTM, T5-small and KeyBERT [18] for keyword extraction. A custom collection of 32,521 Vietnamese news articles, annotated with abstractive summaries and human-generated keywords, was employed in this study. Experiments on single-task settings show that transformer-based models like T5-small (ROUGE-L = 0.3390) and PhoBERT (F1-score = 0.3586) outperform traditional models such as TextRank and Bi-LSTM, with higher inference cost. To address the trade-off between accuracy and computational efficiency, we propose MTL strategies that employ shared encoder architectures and combined loss functions to jointly optimize both summarization and keyword extraction. A multi-task baseline is recommended at least ROUGE-L and F1-score of 0.3, with

* Corresponding author

a reduction in inference time of at least 10-30% compared to separate single-task models. These findings support the potential of MTL as a unified and efficient approach to Vietnamese text processing and lay the groundwork for future development of real-world VNLP applications.

Keywords: Vietnamese NLP, low-resource languages, single-task, multi-task.

1 Introduction

With the rise of online news and social media, the amount of Vietnamese text data has grown rapidly, creating a need for better natural language processing (NLP) tools to understand and summarize content [1]. Tasks like text summarization, which creates concise versions of articles, and keyword extraction, which identifies important words or phrases, are essential for applications such as news apps, search engines, and content recommendations [2]. However, Vietnamese is a tonal language, which complicates NLP due to its complex word forms and tone-dependent meanings [3, 4]. Most existing datasets for Vietnamese focus on single-task setting, such as word segmentation or part-of-speech tagging, text summarization [1, 4]. This gap hinders the development of models capable of handling multiple tasks together, limiting efficiency in processing the growing volume of Vietnamese text data. Addressing these challenges requires robust models and datasets that can support both single-task and MTL approaches for VNLP.

Text summarization and keyword extraction are foundational for enabling efficient information retrieval and content analysis in VNLP [2]. Single-task learning has been the predominant approach, with models like TextRank [5] and LexRank [6] for extractive summarization, which use graph-based algorithms to rank sentences based on similarity or lexical centrality. For abstractive summarization, transformer-based models like T5-small [7] leverage pre-trained architectures fine-tuned on Vietnamese data, often processed by tools like VnCoreNLP [4]. For keyword extraction, Bi-LSTM models [8] capture sequential dependencies, while KeyBERT [9] leverages BERT embeddings for semantic relevance. However, single-task models are computationally intensive and fail to exploit shared linguistic patterns, limiting generalization in low-resource settings where annotated data is sparse [3, 9]. This study evaluates these models to establish a performance baseline, highlighting their strengths and inefficiencies in addressing VNLP's data scarcity and linguistic complexity.

MTL addresses these limitations by training a single model on multiple tasks simultaneously, sharing representations to improve performance and efficiency [10, 11]. In VNLP, summarization and keyword extraction share semantic dependencies, as both involve identifying key content and contextual relevance [2, 9]. MTL can leverage these similarities through shared encoder architectures, such as PhoBERT, a transformer-based model pre-trained on Vietnamese text [3]. By optimizing a joint loss function, MTL balances performance across tasks, reducing overfitting and computational overhead in data-scarce environments [10]. Recent advancements in transfer learning, exemplified by models like BERT [9] and T5 [7], demonstrate MTL's potential to enhance generalization in low-resource languages [2, 3]. For instance, combining summarization and keyword extraction can improve a model's ability to capture salient information, as keywords often reflect core ideas needed for summaries [9].

In this study, we aim to establish strong single-task baselines for Vietnamese text summarization and keyword extraction, and to propose a framework that can serve as the foundation for future multi-task learning research. The objectives of this study are to: (1) evaluate single-task model performance was assessed ROUGE metrics for summarization and Precision, Recall, and F1-score for keyword extraction, (2) assess computational efficiency via processing time, and (3) propose MTL strategies to enhance VNLP performance. The research is structured as follows: Section 2 details the methodology, including model descriptions and evaluation metrics; Section 3 presents the experimental setup and results; Section 4 discusses findings and MTL directions; and Section 5 concludes with a summary of insights and future work.

2 Methodology

2.1 Text summarization

Text summarization creates concise and coherent summaries that capture a text's main ideas while reducing redundancy [2, 13]. Methods include extractive approaches, which select key sentences (e.g., TextRank [5], LexRank [6]), and abstractive approaches, which generate new sentences using models like T5 [7]. Summarization is vital for applications such as news aggregation [2]. In VNLP, tonal dependencies and limited datasets pose challenges, requiring preprocessing with tools like VnCoreNLP [4, 15] to handle linguistic complexities [1, 3]. In this study, we evaluated three models: TextRank, LexRank, and T5-small. TextRank [5] is a graph-based algorithm inspired by PageRank that ranks sentences using TF-IDF similarity, offering language independence, low cost, and no training data requirements. LexRank [6] extends TextRank with centrality measures to prioritize salient sentences and better handle redundancy. T5-small [7], an abstractive transformer model, generates fluent summaries and captures complex semantics, though it demands substantial computational resources.

2.2 Keyword extraction

Keyword extraction identifies key words or phrases that represent a text's main topics, supporting indexing and search [2, 9, 12]. Approaches include statistical methods, neural models such as Bi-LSTM-CRF for sequence labeling [8] and embedding-based methods like KeyBERT [18]. Vietnamese's tonal morphology increases the challenge, requiring accurate preprocessing [3, 4]. In this study, we evaluated three methods: TextRank, Bi-LSTM, and KeyBERT/PhoBERT. TextRank [5] is a graph-based algorithm that ranks words or sentences based on co-occurrence, widely used for unsupervised keyword extraction due to its domain independence. Bi-LSTM models [8] capture sequential dependencies and assign importance to tokens, making them useful in low-resource languages like Vietnamese. KeyBERT [9, 18] leverages BERT [16] embeddings to identify semantically relevant keywords via cosine similarity, and when combined with PhoBERT [3], it better handles Vietnamese linguistic features such as tonal diacritics and morphology.

2.3 Evaluation Metrics

We evaluate summarization with ROUGE-1, ROUGE-2, ROUGE-3, and ROUGE-L [10], and assess keyword extraction with Precision, Recall, and F1-score [9].

3 Experiments

3.1 Dataset description

The dataset was initially collected with 35,406 Vietnamese news articles sourced from reputable online platforms, spanning diverse topics. To ensure high-quality data for benchmarking text summarization and keyword extraction. The preprocessing steps were as follows: First, records with missing content, summaries, or keywords were removed, reducing the dataset to 34,940 articles. Next, erroneous data were filtered out, including articles with summaries shorter than 50 characters, content shorter than 300 characters, or summaries longer than their corresponding content. Noise, such as HTML tags, special characters, and other artifacts, was stripped from the content and summaries to ensure clean text. Finally, keywords longer than 9 words were removed to maintain relevance and conciseness. After these steps, the final dataset comprises 32,521 articles, each annotated with a human-generated abstractive summary and manually extracted keywords.

The final dataset includes 32,521 Vietnamese news articles, covering topics such as human (37%), science (20%), economy (14%), education (9%), politics (9%), healthcare (5%), environment (4%), and other domains (1%). To characterize the dataset, we computed several basic statistics. The average article length is 661 words (standard deviation: 401 words), with summaries averaging 31 words (standard deviation: 8 words). The keyword set per article has an average of 4.2 keywords and the average number of words per keyword is 2.7.

Preprocessing includes tokenization performed the Underthesea library and normalization to handle Vietnamese diacritics. Underthesea's robust tokenization capabilities ensure accurate segmentation of Vietnamese text, making it suitable for the dataset's linguistic diversity. The dataset's size and variety make it appropriate for benchmarking both summarization and keyword extraction tasks.

3.2 Experimental results and analysis

To evaluate the dataset, we implemented single-task models for text summarization (TextRank, LexRank, T5-small) and keyword extraction (TextRank, Bi-LSTM, KeyBERT). The models were trained on Google Colab Pro using an NVIDIA A100 GPU with 40GB VRAM, a CPU with 25GB RAM, and a 150GB SSD, ensuring sufficient computational resources for fine-tuning and evaluation. Summarization performance was measured using ROUGE-1, ROUGE-2, ROUGE-3 and ROUGE-L metrics, which assess unigram, bigram, trigram and longest common subsequence overlap between generated and reference summaries. Keyword extraction performance was evaluated using Precision, Recall, and F1-score, measuring the accuracy of extracted keywords against human-annotated keywords. All experiments were conducted on the full dataset

of 32,521 articles, split into 60% training (19,513 articles), 20% validation (6,504 articles), and 20% testing sets (6,504 articles).

As illustrated in Fig. 1, the experimental framework begins with preprocessing a dataset of 32,521 Vietnamese news articles, which are split into training, validation, and testing sets. Single-task experiments are then performed for summarization (TextRank, LexRank, T5-small) and keyword extraction (TextRank, Bi-LSTM, PhoBERT), each evaluated with widely used metrics such as ROUGE scores, Precision, Recall, and F1-score. The outcomes are further integrated into a proposed MTL combined evaluation, establishing baseline performance and efficiency targets to guide future multi-task implementations.

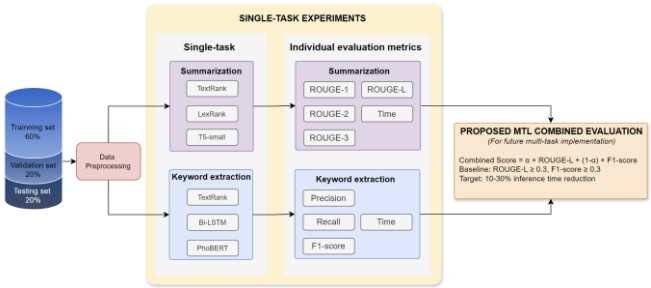

Fig. 1. Single-task experimental framework and proposed multi-task evaluation methodology

Summarization results. TextRank and LexRank exhibit comparable performance across ROUGE metrics, with optimal configurations at top_k = 2. TextRank performs better with a fully connected graph (threshold = 0.0), whereas LexRank benefits from filtering out low-similarity sentence pairs (threshold = 0.1). T5-small, a transformer-based model, significantly outperforms both extractive methods by leveraging deep semantic representations, achieving the highest ROUGE scores. However, this improvement comes with a substantial computational cost, requiring approximately 11,820 seconds for inference, around eight times longer than TextRank and LexRank (~1,440 seconds). These results underscore a trade-off between summarization quality and computational efficiency: transformer models like T5 are more suitable for high-resource settings, while graph-based methods offer advantages in time-sensitive or resource-constrained scenarios.

Table 1. Summarization performance

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L	Time (s)
TextRank	0.4772	0.1805	0.0922	0.2953	1,480
LexRank	0.4804	0.1750	0.0881	0.2946	**1,440**
T5-small	**0.5443**	**0.2089**	**0.1124**	**0.3390**	11,820

Bolded values indicate the best performance among the compared methods.

Keyword extraction results. PhoBERT emerges as the top performer with a Precision of 0.4028, Recall of 0.3145, and F1-score of 0.3586, slightly outperforming Bi-LSTM (F1-score: 0.3263) and significantly surpassing TextRank (F1-score: 0.1050). This

suggests KeyBERT's effectiveness in leveraging PhoBERT embeddings [3, 9, 18] for contextual keyword extraction, though its longer runtime (13,500s) reflects the computational overhead of transformer-based models. Bi-LSTM's F1-score (0.3263) indicates strong contextual modeling, likely due to its bidirectional sequence approach, but it falls short of PhoBERT's precision. TextRank lower and variable performance (Precision 0.0873, Recall 0.1315) and minimal runtime (4,054s) position it as a fast but less accurate baseline, consistent with its statistical nature.

Table 2. Keyword extraction performance

	Precision	Recall	F1-score	Time (s)
TextRank	0.0873	0.1315	0.1050	4,054
Bi-LSTM	0.3657	0.2945	0.3263	**3,420**
PhoBERT	**0.4027**	**0.3145**	**0.3586**	13,500

4 Discussion

In this study, an empirical evaluation of statistical, neural, and transformer-based models was conducted as the foundation for developing multi-task learning approaches to Vietnamese text summarization and keyword extraction. Statistical methods (TextRank, LexRank) are fast and interpretable but limited in capturing semantics, while deep learning models (Bi-LSTM, PhoBERT) achieve better contextual performance with higher computational cost. Transformer-based models (e.g., T5-small) deliver the best summarization quality but face challenges of fine-tuning cost and inference time.

To identify the best combination for MTL, we evaluated pairs of summarization and keyword extraction methods [11] use the combined score $\{Combined\ Score\} = \alpha\{ROUGE\_L\} + (1 - \alpha)\{F1\_score\}$. Table 3 shows the results.

Table 3. Combined performance scores for MTL combinations

Summarization	Keyword Extraction	ROUGE-L	F1-score	Combined Score	Time (s)
TextRank	TextRank	0.2953	0.1050	0.2002	5,534
	Bi-LSTM	0.2953	0.3263	0.3108	4,900
	PhoBERT	0.2953	0.3586	0.3270	14,980
LexRank	TextRank	0.2946	0.1050	0.1998	5,494
	Bi-LSTM	0.2946	0.3263	0.3105	**4,860**
	PhoBERT	0.2946	0.3586	0.3266	14,940
T5-Small	TextRank	0.3390	0.1050	0.2220	15,874
	Bi-LSTM	0.3390	0.3263	0.3327	15,240
	PhoBERT	0.3390	0.3586	**0.3488**	25,320

We evaluated summarization and keyword extraction pairs evaluated through a combined score with $\alpha = 0.5$. Results show that T5-Small + PhoBERT achieves the highest combined score (0.3488) but with long processing time (25,320s), while LexRank + Bi-LSTM offers the best trade-off, reaching a reasonable score (0.3105) with the lowest

inference time (4,860s). This suggests that lightweight summarization with a moderately complex extractor can balance accuracy and efficiency for practical use.

Based on these findings, we propose an MTL design with a shared encoder and task-specific decoders to capture common semantics while preserving task-specific objectives. A combined loss function (ROUGE for summarization, F1 for keyword extraction) is recommended to optimize both tasks fairly. As a practical baseline, MTL models should aim for ROUGE-L and F1 ≥ 0.30, while reducing inference time by 10–30% compared to running separate models.

In this study, strong single-task baselines were established, and possible strategies for MTL were discussed. As the next step, future work will test and evaluate MTL models on our dataset of 32,521 Vietnamese news articles. Future work will be conducted with architecture incorporating shared encoders, task-specific decoders, dynamic loss weighting, and feature reduction, to see whether MTL can reach or even improve on the performance of single-task models while being more efficient for real-world uses such as news aggregation, content indexing, and search.

One important challenge of MTL is negative transfer, where knowledge that helps one task may reduce the performance of another [11]. This issue is relevant in our case because summarization works at the level of overall meaning in a text, while keyword extraction focuses on short phrases. Such differences can make shared representations less effective. To reduce this risk, future work can explore task-specific decoders, selective parameter sharing, and dynamic loss weighting. More advanced methods such as gradient normalization or adversarial training may also help limit harmful interference and improve generalization.

5 Conclusion

This paper presented an empirical study on Vietnamese text summarization and keyword extraction, focused on both single-task baselines and potential strategies for multi-task learning (MTL). Using a curated dataset of 32,521 news articles, we evaluated statistical, neural, and transformer-based models, showing that T5-small and PhoBERT achieve the best accuracy, while LexRank combined with Bi-LSTM offers a more efficient trade-off for practical use.

This study lays the groundwork for applying MTL to Vietnamese summarization and keyword extraction by highlighting its potential to reduce redundancy and improve efficiency within a unified framework. At this stage, our work does not include the implementation of MTL; instead, we provide analysis and design considerations as a foundation for future research. In the next step, we plan to experiment with architectures that combine shared encoders, task-specific decoders, and dynamic loss balancing to evaluate their effectiveness and address challenges such as negative transfer in real-world Vietnamese NLP applications.

References

1. Nguyễn, V.-H., Nguyễn, T.-C., Nguyễn, M.-T., Nguyễn, H.: VNDS: A Vietnamese Dataset for Summarization. In: 2019 NAFOSTED Conference on Information and Computer Science (NICS), pp. 375–380 (2019).
2. Radev, D.R., Hovy, E., McKeown, K.: Introduction to the special issue on summarization. Computational Linguistics 28(4), 399–408 (2002).
3. Nguyễn, D.Q., Nguyễn, A.T., Vũ, T.: PhoBERT: Pre-trained language models for Vietnamese. In: Findings of EMNLP, pp. 1037–1042 (2020).
4. Vu, T., Nguyễn, D.Q., Nguyễn, D.Q., Dras, M., Johnson, M.: VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In: NAACL Demonstrations, pp. 56–60 (2018).
5. Mihalcea, R., Tarau, P.: TextRank: Bringing order into text. In: EMNLP 2004, pp. 404–411 (2004).
6. Erkan, G., Radev, D.R.: LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22, 457–479 (2004).
7. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR 21(140), 1–67 (2020).
8. Alzaidy, R., Caragea, C., Giles, C.L.: Bi-LSTM-CRF Sequence Labeling for Keyphrase Extraction from Scholarly Documents. In: The Web Conference 2019, pp. 2551–2557 (2019).
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL 2019, pp. 4171–4186 (2019).
10. Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries. In: ACL Workshop on Text Summarization Branches Out, pp. 74–81 (2004).
11. Chen, S., Zhang, Y., Yang, Q.: Multi-task learning in natural language processing: An overview. ACM Computing Surveys 56(12), 1–32 (2024).
12. Nomoto, T.: Keyword extraction: a modern perspective. SN Computer Science 4(1), 92 (2022).
13. Wibawa, A.P., Kurniawan, F.: A survey of text summarization: Techniques, evaluation and challenges. Natural Language Processing Journal 7, 100070 (2024).
14. Giarelis, N., Karacapilidis, N.: Deep learning and embeddings-based approaches for keyphrase extraction: a literature review. Knowledge and Information Systems 66(11), 6493–6526 (2024).
15. Nguyen, T.H., Do, T.N.: Pre-Training Clustering Models to Summarize Vietnamese Texts. Vietnam Journal of Computer Science 12(1), 83–100 (2025).
16. Luo, Y.: Keywords extraction algorithm based on attention mechanism of BERT model. In: International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), vol. 13395, pp. 1168–1173. SPIE, November (2024).
17. Ngo, T.M., Ngo, B.H., Valerievich, S.V.: Fine-tuned PhoBERT for sentiment analysis of Vietnamese phone reviews. CTU Journal of Innovation and Sustainable Development 16(Special issue: ISDS), 52–57 (2024).
18. Issa, B., Jasser, M.B., Chua, H.N., Hamzah, M.: A comparative study on embedding models for keyword extraction using keybert method. In: 2023 IEEE 13th International Conference on System Engineering and Technology (ICSET), pp. 40–45. IEEE, October (2023).