ViSoLex: An Open-Source Repository for Vietnamese Social Media Lexical Normalization

Anh Thi-Hoang Nguyen[1,2], Dung Ha Nguyen[1,2], Kiet Van Nguyen[1,2,*]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
***Correspondence:** kietnv@uit.edu.vn
Contributing authors: 20520134@gm.uit.edu.vn, dungngh@uit.edu.vn

Abstract

ViSoLex is an open-source system designed to address the unique challenges of lexical normalization for Vietnamese social media text. The platform provides two core services: Non-Standard Word (NSW) Lookup and Lexical Normalization, enabling users to retrieve standard forms of informal language and standardize text containing NSWs. ViSoLex's architecture integrates pre-trained language models and weakly supervised learning techniques to ensure accurate and efficient normalization, overcoming the scarcity of labeled data in Vietnamese. This paper details the system's design, functionality, and its applications for researchers and non-technical users. Additionally, ViSoLex offers a flexible, customizable framework that can be adapted to various datasets and research requirements. By publishing the source code, ViSoLex aims to contribute to the development of more robust Vietnamese natural language processing tools and encourage further research in lexical normalization. Future directions include expanding the system's capabilities for additional languages and improving the handling of more complex non-standard linguistic patterns.

1 Introduction

The increasing presence of Non-Standard Words (NSWs) in social media has introduced significant challenges for natural language processing (NLP) systems. In Vietnamese, these challenges are particularly pronounced due to the informal, abbreviated, and non-canonical nature of social media language. Lexical normalization, which transforms NSWs into their standard forms, is essential for improving the performance of downstream tasks such as sentiment analysis, hate speech detection, and machine translation. While research on lexical normalization has made significant advancements globally, Vietnamese has lagged behind due to a lack of resources and standardized datasets.

To address these challenges, we introduce ViSoLex[1], an open-source repository for Vietnamese lexical normalization. ViSoLex provides a comprehensive solution by integrating multitask learning capabilities to simultaneously detect and normalize NSWs. This is achieved by leveraging pre-trained language models and weak supervision techniques, reducing the dependency on extensive manual labeling. Furthermore, ViSoLex incorporates a growing dictionary of NSWs for dictionary lookup, enabling efficient identification and normalization of non-standard words.

The repository is designed to address the unique linguistic challenges of Vietnamese social media text and fosters customization, allowing researchers to adapt the system for various datasets and languages. By offering a scalable and open-source solution, ViSoLex supports broader research and practical applications, advancing the field of Vietnamese NLP. This paper presents the system architecture, multitask training framework, and the extensive efforts made to improve the quality of Vietnamese NLP tasks through lexical normalization.

2 Related Works

The study of lexical normalization has seen significant advancements worldwide, especially in addressing the challenges of non-standard text. Early approaches like the Abbreviation Expander by Ciosici and Assent (2018) tackled abbreviation expansion in technical documents, providing a web-based solution for easy understanding of domain-specific terms. In 2019, MoNoise by van der Goot (2019) was introduced for vocabulary normalization, using spelling correction and word embeddings with a Feature-Based Random Forest Classifier. Initially for English, it later expanded to support multiple languages, becoming a widely used

[1] https://github.com/HaDung2002/visolex

multilingual tool. Furthermore, Muller et al. (2019) marked a shift toward using pre-trained language models for handling noisy text in user-generated content, framing normalization as a token prediction task. Nguyen et al. (2021) introduced the idea of capturing not just lexical meaning but also social context, a key aspect in understanding informal and non-canonical language. These developments paved the way for more robust normalization systems across various languages, demonstrating the potential of combining linguistic insights with modern NLP techniques.

In the Vietnamese context, Tran et al. (2021) employed deep learning models like Bidirectional-GRU to solve the problem of missing diacritics. Do et al. (2021) further advanced the field by using VSEC model, a Transformer-based approach, to correct Vietnamese spelling errors, significantly improving upon prior methods. Nguyen et al. (2024c) introduced the first corpus, called ViLexNorm, for Vietnamese lexical normalization, a critical resource for normalizing social media text and improving downstream tasks. Additionally, Nguyen et al. (2024a) introduced a Seq2Seq approach for normalizing NSWs, with a publicly available dataset for further research. Building upon this foundation, Nguyen et al. (2024b) proposed a novel framework that integrates semi-supervised learning with weak supervision techniques, leveraging pre-trained language models to enhance dataset quality, reduce manual labeling efforts, and normalize NSWs.

In this paper, we further advance our previous work proposed in Nguyen et al. (2024b), now named ViSoLex, by incorporating multitask learning capabilities to simultaneously detect and normalize NSWs. Additionally, we have integrated a dictionary lookup feature for non-standard word detection. The ViSoLex repository is designed as an open-source solution for Vietnamese lexical normalization, specifically addressing the unique linguistic challenges presented by social media text, and is made publicly available for broader use and development.

3 ViSoLex: Vietnamese Social Media Lexical Normalization

3.1 System Architecture

ViSoLex is designed to provide two key services: NSW Lookup and Lexical Normalization. Users can input a NSW for interpretation or enter a sentence containing NSWs for normalization. The architecture of ViSoLex, as illustrated in Figure 1, follows a modular design that integrates various components to streamline Vietnamese social media text normalization. At the core, user inputs flow through distinct paths depending on the requested service. Communication between the components ensures dynamic interaction and updates. This architecture enables independent updates to different system components while maintaining overall functionality.

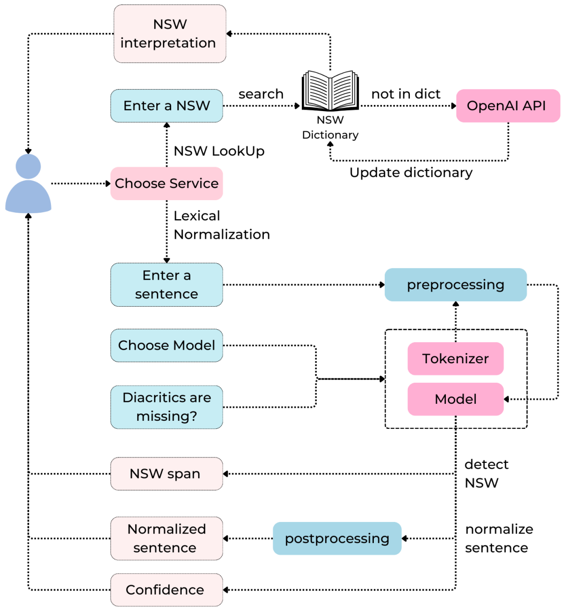


Figure 1: The Architecture of ViSoLex. The diagram illustrates the modular components enabling NSW Lookup and Lexical Normalization services, including their interactions and flow of user inputs.

3.1.1 NSW LookUp Service

The NSW LookUp service enables users to retrieve potential standard forms and interpretations of NSWs from an established dictionary. Upon choosing this service, users are asked to input an NSW, which is then checked against the existing dictionary. If found, the system returns the standard forms and definitions, along with relevant examples. If the word is not in the dictionary, the system consults the OpenAI GPT-4o API to suggest a possible normalization, which is then added to the dictionary for future use. This approach allows NSWs to be resolved either by utilizing existing data or dynamically learning from external models.

The NSW dictionary was built by leveraging the OpenAI GPT-4o API to generate definitions and examples for each entry in the Vietnamese Non-Standard Words Dictionary[2].

3.1.2 Lexical Normalization Service

The lexical normalization service transforms NSWs in a sentence into their standard forms. When users select this service, they input a sentence that may contain NSWs. The system tokenizes and preprocesses the input before applying a multitask-trained model (discussed in detail in Section 3.2) to identify non-standard tokens and predict their corresponding standard forms, each accompanied by confidence scores. The predicted output undergoes post-processing, where redundant spaces before punctuation are removed, and proper sentence capitalization and punctuation are applied. The final result provides a fully normalized sentence, along with a breakdown of each NSW, its standard equivalent, and the confidence score, ensuring precise normalization for Vietnamese social media text.

3.2 Lexical Normalizer Training

The updated lexical normalizer builds on the framework presented in our previous work Nguyen et al. (2024b), introducing multitask learning to enhance its capabilities. As illustrated in Figure 2, this weakly supervised framework leverages both labeled and unlabeled data to identify and standardize NSWs in Vietnamese social media text. Inspired by the ASTRA framework Karamanolakis et al. (2021), it incorporates two key components: the Lexical Normalizer, now enhanced with multitask learning as a student model, and a Rule Attention Network, acting as a teacher by embedding weak supervision rules. This integration of data-driven and rule-based approaches enables the model to generalize more effectively, handling the diverse and evolving NSW patterns in social media discourse.

3.2.1 Lexical Normalizer

The Lexical Normalizer is trained using a multitask framework to predict the standard forms of NSWs in Vietnamese social media text. It leverages pre-trained models, such as BARTpho Tran et al. (2022) and ViSoBERT Nguyen et al. (2023), fine-tuned for text normalization. The input consists of sentences with NSWs, and the model outputs both NSW detection and their normalized forms, combining token classification with sequence generation for effective normalization.

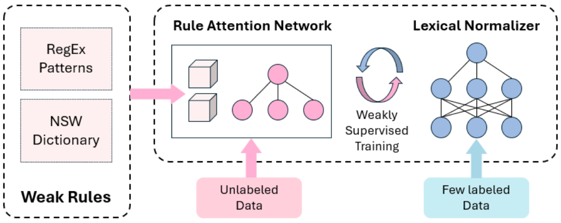


Figure 2: Weak Supervision Training. This figure illustrates the training process of the lexical normalizer, which integrates multitask learning and a Rule Attention Network guided by weak supervision rules to effectively standardize NSWs in Vietnamese social media text.

ViSoLex introduces multitask learning to simultaneously handle NSW detection and lexical normalization. A shared encoder extracts input features, followed by task-specific heads that generate predictions. The model minimizes the binary cross-entropy loss $\mathcal{L}_{NSW}$ for NSW detection and cross-entropy loss $\mathcal{L}_{Norm}$ for normalization, with the total loss:

$$\mathcal{L}_{Total} = \alpha\mathcal{L}_{Norm} + \beta\mathcal{L}_{NSW} \tag{1}$$

where α and β balance the contributions of each task. This multitask approach enhances efficiency and performance in normalizing noisy social media text.

3.2.2 Rule Attention Network with Weak Rules

To further enhance the model's performance, a Rule Attention Network (RAN) is integrated, guided by weak supervision rules. These weak rules are derived from NSW dictionary and regular expression rules, capturing common patterns of NSWs in Vietnamese social media text. As shown in Figure 2, the RAN learns to assign different levels of attention to these rules during the training process. This network dynamically weighs the influence of each rule based on the context and reliability of the prediction, allowing the model to flexibly adapt to both well-defined and ambiguous cases of NSWs. The combination of weak supervision with the rule attention mechanism allows the model to effectively learn from limited labeled data, improving both NSW detection and normalization accuracy.

[2] https://github.com/AnhHoang0529/vn-nsw-dictionary

4 Evaluation

In this section, we evaluate the performance of multitask learning in comparison to the lexical normalization approach presented by Nguyen et al. (2024b). The evaluation employs three key metrics: the F1-score, which specifically measures the accuracy of normalizing NSWs; the Integrity Score, which assesses the preservation of words that do not require normalization; and Accuracy, which evaluates the overall correctness of the predicted sentences. These metrics are defined and explained in detail by Nguyen et al. (2024b).

Table 1 demonstrate the impact of multitask learning on model performance across different metrics, with p representing the diacritics removal ratio. The results show that multitask learning consistently enhances model performance. For BARTpho, improvements in F1-score are modest, with increases of 0.34% and 0.25% for different diacritics removal ratios ($p = 0$ and $p = 1$). ViSoBERT, however, benefits from multitask learning, particularly when all diacritics are removed ($p = 1$), with a notable 3.74% increase in F1-score.

In terms of Accuracy, both models see slight improvements, but again, ViSoBERT shows stronger gains, reinforcing its ability to normalize sentence with diacritics removal in traning and development dataset. Despite a small decrease in the Integrity Score for BARTpho, ViSoBERT improves, particularly in high diacritics removal scenarios.

Overall, multitask learning proves especially effective for ViSoBERT, leading to performance improvements in handling noisy text data with diacritic variations.

5 System Demonstration

In this section, we outline the system demonstration of ViSoLex, tailored to meet the needs of various user groups within the NLP community. We offer two distinct entry points to accommodate both technical and non-technical users.

5.1 For Researchers and Developers

We have published the source code on GitHub to allow researchers and developers to leverage the system's capabilities through the following features:

- **Model Training and Evaluation**: Users can utilize the top-level script `main.py` to retrain models, reproduce results, and evaluate performance. This offers comprehensive insight into the system's underlying methodologies and processes.
- **Demo Functionality**: For a quick overview, users can run an interactive terminal session using `demo.py`, which demonstrates the core functionalities of the system with minimal setup.

The framework is also designed for flexibility, allowing users to customize the model and its components for specific datasets and model selection, enhancing its adaptability to various research applications. Key customization options include:

- `data/`: Users can replace the default data with their own, ensuring they provide three labeled data files (`train.csv`, `dev.csv`, `test.csv`) and an unlabeled data file (`unlabeled.csv`).
- `dict/`: Users can integrate a custom NSW dictionary to further align the framework with their specific language or domain requirements.
- `aligned_tokenizer.py`: The token-level alignment tokenizer can be modified to suit the characteristics of different datasets and languages.
- `normalizer/model_construction/`: New models for lexical normalization, tailored to different languages or datasets, can be added here.
- `project_variables.py`: Global constants such as data directories or language-specific tokens can be modified to fit custom requirements.
- `arguments.py`: Users can configure additional settings for their projects and reset the default argument values.

This modular and customizable design allows researchers to tailor the system to meet their unique needs in lexical normalization tasks.

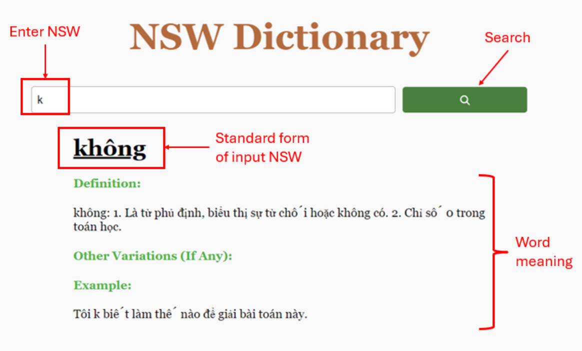
5.2 For Non-Experts

To accommodate non-technical users, we developed a user-friendly front-end interface using a Flask web application. The interface provides two main services, accessible through distinct endpoints:

Metric	**Task**	**BARTpho** $p = 0.0$	**BARTpho** $p = 1.0$	**ViSoBERT** $p = 0.0$	**ViSoBERT** $p = 1.0$
F1-score (%)	Single task	84.94	85.64	75.79	72.19
	Multitask	85.28	85.89	77.22	75.93
	Improvement	↑0.34	↑0.25	↑1.43	↑3.74
Integrity Score (%)	Single task	98.88	98.62	98.26	96.92
	Multitask	98.83	98.50	98.64	98.27
	Improvement	↓0.05	↓0.12	↑0.38	↑1.35
Accuracy (%)	Single task	96.06	96.16	95.42	94.42
	Multitask	96.14	96.26	95.08	94.76
	Improvement	↑0.08	↑0.10	↑0.34	↑0.34

Table 1: Comparison of Single Task and Multitask Learning Performance on BARTpho and ViSoBERT models with different diacritics removal ratios ($p = 0\%$ and $p = 100\%$

- **Interactive Dictionary Service**: This service, available via the `/dict_lookup` endpoint, allows users to search for non-standard words and retrieve their standard equivalents and definitions from our extensive dictionary. The interface for this service is illustrated in Figure 3.
- **Sentence Normalization**: Through the `/normalize_text` endpoint, users can input sentences containing non-standard words and receive real-time normalized outputs. The UI for this service is shown in Figure 4.



NSW Dictionary

k

không

Definition:

không: 1. Là từ phủ định, biểu thị sự từ chố i hoặc không có. 2. Chỉ số o trong toán học.

Other Variations (If Any):

Example:

Tôi k biế t làm thế nào để giải bài toán này.

Figure 3: User Interface of NSW LookUp Service. This interface allows users to search for non-standard words and retrieve their standard equivalents, definitions, and examples from the dictionary.

The Flask application enables seamless interaction between the front-end and back-end components, ensuring efficient and responsive user experiences. A self-hosting tutorial for deploying the UI is available in the project's GitHub repository, along with a demonstration video on how to use the interface, accessible via this Youtube URL[3].

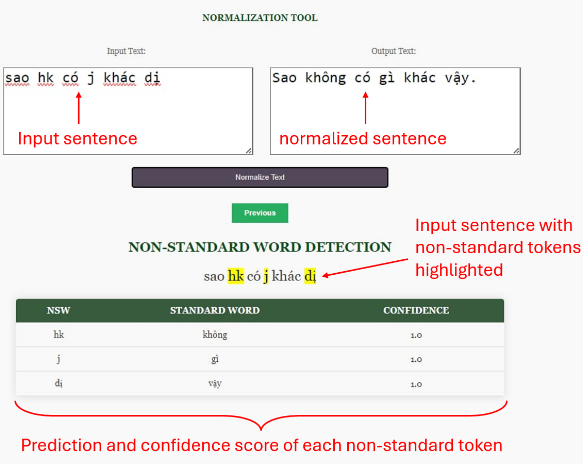
6 Discussion and Future Directions

ViSoLex represents a significant advancement in the lexical normalization of Vietnamese social media text, offering researchers and developers a powerful and flexible tool to address the challenges posed by non-standard language. The system exhibits robust performance in both NSW detection and normalization, leveraging its integration of pre-trained models and weakly supervised learning. Notably, ViSoLex achieves consistent improvements in F1-score and accuracy across multitask settings when compared to the original framework Nguyen et al. (2024b), which was exclusively designed for lexical normalization without NSW detection. Through its open-source availability, ViSoLex encourages further research and application in Vietnamese NLP.

Looking ahead, future work on ViSoLex will focus on expanding its capabilities to support more complex non-standard patterns and handling additional languages. Efforts will also be made to enhance the model's adaptability, allowing it to better manage evolving trends in social media language. Furthermore, expanding the NSW dictionary and refining the system's ability to predict social-contextual meanings are promising directions. Lastly, improving user experience through more intuitive front-end interfaces and incorporating additional downstream NLP tasks will enhance ViSoLex's practical applications in real-world sce-

[3] `https://youtu.be/XBIAogDpF3o?si=PUXiMCuu9qDTfM3B`

narios.



NORMALIZATION TOOL

Input Text: sao hk có j khác dị

Output Text: Sao không có gì khác vậy.

NON-STANDARD WORD DETECTION

sao hk có j khác dị

NSW	STANDARD WORD	CONFIDENCE
hk	không	1.0
j	gì	1.0
dị	vậy	1.0

Input sentence

normalized sentence

Input sentence with non-standard tokens highlighted

Prediction and confidence score of each non-standard token

Figure 4: User Interface of Lexical Normalization Service. This interface enables users to input sentences with non-standard words and receive fully normalized outputs in real-time.

Acknowledgement

This research is funded by University of Information Technology-Vietnam National University Ho Chi Minh City under grant number D1-2024-75.

References

Manuel R. Ciosici and Ira Assent. 2018. Abbreviation expander - a web-based system for easy reading of technical documents. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 1–4, Santa Fe, New Mexico. Association for Computational Linguistics.

Dinh-Truong Do, Ha Thanh Nguyen, Thang Ngoc Bui, and Hieu Dinh Vo. 2021. Vsec: Transformer-based model for vietnamese spelling correction. In *PRICAI 2021: Trends in Artificial Intelligence*, pages 259–272, Cham. Springer International Publishing.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 845–863, Online. Association for Computational Linguistics.

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2019. Enhancing BERT for lexical normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 297–306, Hong Kong, China. Association for Computational Linguistics.

Anh Thi-Hoang Nguyen, Dung Ha Nguyen, Nguyet Thi Nguyen, Khanh Thanh-Duy Ho, and Kiet Van Nguyen. 2024a. Automatic textual normalization for hate speech detection. In *Intelligent Systems Design and Applications*, pages 1–12, Cham. Springer Nature Switzerland.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Dung Ha Nguyen, Anh Thi Hoang Nguyen, and Kiet Van Nguyen. 2024b. A weakly supervised data labeling framework for machine lexical normalization in vietnamese social media. *Preprint*, arXiv:2409.20467.

Nam Nguyen, Thang Phan, Duc-Vu Nguyen, and Kiet Nguyen. 2023. ViSoBERT: A pre-trained language model for Vietnamese social media text processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5191–5207, Singapore. Association for Computational Linguistics.

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Nguyen. 2024c. ViLexNorm: A lexical normalization corpus for Vietnamese social media text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1421–1437, St. Julian's, Malta. Association for Computational Linguistics.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022. Bartpho: Pre-trained sequence-to-sequence models for vietnamese. *Preprint*, arXiv:2109.09701.

Quang-Linh Tran, Gia-Huy Lam, Van-Binh Duong, and Trong-Hop Do. 2021. A study on diacritic restoration problem in vietnamese text using deep learning based models. In *2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pages 306–310.

Rob van der Goot. 2019. MoNoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206, Florence, Italy. Association for Computational Linguistics.