A Chinese-to-Vietnamese Machine Translation Program

Vũ Đình Dũng
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student624@ictu.edu.vn

***Abstract*—This undergraduate thesis investigates a Chinese-to-Vietnamese machine translation program. Translation between Chinese and Vietnamese requires a parallel corpus and a model that can represent word sequence and context. The student surveys neural translation, recurrent networks, attention, and Transformers, then prepares a dataset and trains a Transformer-based model. The thesis documents example translations and reports training settings and a BLEU result. The report gives a BLEU value of 0.73, a training loss of 0.1592, and paired example translations. These values are reproduced as reported by the student, without an independent test or cross-system comparison. The source itself notes limits in corpus quality and simple word segmentation. Dataset split, benchmark comparability, and generalization to other domains are insufficiently documented for broader performance claims. The contribution is a documented student study or prototype whose stated scope can guide further testing.**

***Index Terms*—machine translation, Chinese, Vietnamese, Transformer, BLEU**

I. Introduction

Translation between Chinese and Vietnamese requires a parallel corpus and a model that can represent word sequence and context. This thesis investigates a Chinese-to-Vietnamese machine translation program. It begins from a practical engineering need and explains the technologies and workflow chosen for the case. The objective is to document a student implementation or comparative investigation, together with the evidence available in the manuscript. Broader operational or economic effects require separate measurements and are not assumed from the existence of a prototype [1]. Earlier ICTU student work on related designs is included here for context [2]. Related published studies provide context for the subject and methods [3], [4], [5], [6], [7], [8], [9], [10]. These contextual works were selected for this English summary; they are not asserted to appear in the original thesis bibliography.

II. Methods

The student surveys neural translation, recurrent networks, attention, and Transformers, then prepares a dataset and trains a Transformer-based model. The thesis documents example translations and reports training settings and a BLEU result. The source is summarized through its stated design, implementation material, examples, and conclusion. Features shown in the manuscript are distinguished from proposals for later development. This review does not add a new experiment, fill missing measurements with estimates, or treat a planned capability as an observed result. The source thesis remains the provenance for the reported work [1].

Supervisor: Ngô Hữu Huy, Thai Nguyen University of Information and Communication Technology (nhhuy@ictu.edu.vn).

III. Results

The report gives a BLEU value of 0.73, a training loss of 0.1592, and paired example translations. These values are reproduced as reported by the student, without an independent test or cross-system comparison. The report's screens, configurations, examples, or measurements constitute the evidence for this summary. A capability described only as a design intention is not counted as a validated operational outcome. No additional performance, adoption, financial, or safety figure has been introduced here. The result is interpreted within the scope of the student's documented setup [1].

IV. Discussion

The source itself notes limits in corpus quality and simple word segmentation. Dataset split, benchmark comparability, and generalization to other domains are insufficiently documented for broader performance claims. Further work could curate a stronger parallel corpus, report a reproducible evaluation split, and compare output against a documented translation baseline. These limitations affect how the evidence should be read. The current manuscript can support understanding of the design and reported demonstration, while independent tests would be needed for stronger claims. Any later deployment should document its environment and separate observed outcomes from intended benefits [1].

V. Conclusion

The thesis documents a Chinese-to-Vietnamese machine translation program. The report gives a BLEU value of 0.73, a training loss of 0.1592, and paired example translations. These values are reproduced as reported by the student, without an independent test or cross-system comparison. The source itself notes limits in corpus quality and simple word segmentation. Dataset split, benchmark comparability, and generalization to other domains are insufficiently documented for broader performance claims. Further work could curate a stronger parallel corpus, report a reproducible evaluation split, and compare output against a documented translation baseline.

Accordingly, the present contribution is a source-grounded account of the student work and its tested or illustrated scope. Claims about broader effectiveness should be assessed in later work using explicit evaluation conditions [1].

References

[1] Vũ Đình Dũng, ”A Chinese-to-Vietnamese Machine Translation Program,” internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2024.

[2] ICTU Student 197, ”Application of Genetic Algorithm in Text Classification of Vietnamese Language,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2010.

[3] Hong-Hai Phan-Vu et al., ”Neural Machine Translation between Vietnamese and English: an Empirical Study,” Journal of Computer Science and Cybernetics, 2019. doi:10.15625/1813-9663/35/2/13233.

[4] Zhiqiang Yu et al., ”Linguistic feature template integration for Chinese-Vietnamese neural machine translation,” Frontiers of Computer Science, 2022. doi:10.1007/s11704-022-1297-0.

[5] Hongzheng Li et al., ”Revisiting Back-Translation for Low-Resource Machine Translation Between Chinese and Vietnamese,” IEEE Access, 2020. doi:10.1109/access.2020.3006129.

[6] Zhiqiang Yu et al., ”Improving Chinese-Vietnamese Neural Machine Translation with Linguistic Differences,” ACM Transactions on Asian and Low-Resource Language Information Processing, 2022. doi:10.1145/3477536.

[7] Phuoc Tran et al., ”A Method of Chinese-Vietnamese Bilingual Corpus Construction for Machine Translation,” IEEE Access, 2022. doi:10.1109/access.2022.3186978.

[8] Huu-Anh TRAN et al., ”Preordering for Chinese-Vietnamese Statistical Machine Translation,” IEICE Transactions on Information and Systems, 2019. doi:10.1587/transinf.2018edp7211.

[9] Anh Tran Huu et al., ”Integrating pronunciation into Chinese-Vietnamese statistical machine translation,” Tsinghua Science and Technology, 2018. doi:10.26599/tst.2018.9010006.

[10] Nghia Luan Pham et al., ”A Data Augmentation Method for English-Vietnamese Neural Machine Translation,” IEEE Access, 2023. doi:10.1109/access.2023.3252898.