VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name: Ngô Thị Vinh 2. Sex: Female

3. Date of birth: 02/08/1984 4. Place of birth: Thai Nguyen

5. Admission decision number: 654/QĐ-ĐT Dated 05/09/2016

6. Changes in academic process: 166/QĐ-ĐT Dated 22/02/2019

7. Official thesis title: Improving the quality of neural machine translation systems in the low-resource issue.

8. Major: Computer Science 9. Code: 948.01.01.01

10. Supervisors:

1. Assoc. Prof. Nguyen Phuong Thai
2. Assoc. Prof. Nguyen Le Minh

11. Summary of the **new findings** of the thesis:

The thesis have proposed some different strategies to enhance the quality of machine translation in low-resource situations on translation tasks which are relative to Vietnamese:

i) Firstly, we have proposed the translation system based on segmentation of texts for Japanese-Vietnamese language pairs including subword, word and character and the translations have obtained improvements up to +3.92 BLEU points. Besides, we have suggested an unsupervised method for Vietnamese texts and have achieved an equivalent improvement compared to the pyvi tool. Experiments have investigated separate bilingual sets which are crawled from various sources on the Internet.

ii) Secondly, we have proposed a simple and fast method for generating synthetic data in the under-resource circumstance of parallel data. The method does not require external resources such as pre-trained language or translation models, bilingual dictionaries, syntax parser, ... We have obtained significant improvements (up to +4.0 BLEU points) on the translation tasks such as Chinese-Vietnamese and Japanese-Vietnamese.

iii) The third, we have proposed using multilingual machine translation systems for the language pairs which could share common information through vocabulary, syntax and lexical when using pre-trained language models. We have suggested

different segmentations for Japanese texts in the translation task from Chinese, Japanese to Vietnamese. Our proposal has shown substantial improvements up to +7.8 BLEU scores. Moreover, we propose to use artificial labels to augment sharing information among translation units in multilingual translation space. Furthermore, we have proposed two strategies to improve translation of rare words in multilingual space for the translation task from English, French to Vietnamese and have gained improvements up to +1.93 BLEU points..

iv) The fourth, the thesis have proponed some solutions to advance the translation rare words in low-resource issues including: (1) improving the decode process base on labeling rare words, our experiments have achieved improvement up to +1.8 BLEU scores; (2) combining word embeddings to the output predict probability, segmenting English texts with the unsupervised method, leverage relative synonym on lexical WordNet database. The combination of these methods has shown an improvement on translation performance up to +0.9 points BLEU.

v) Lastly, we publish some datasets for research purposes such as English-Vietnamese, French-Vietnamese, Chinese-Vietnamese and Japanese-Vietnamese.

12. Practical applicability, if any:

- The outcomes of the thesis could serve as valuable reference materials for studying and researching in natural language processing areas, especially in the machine translation field.
- The proposed methods in the thesis could be applied to machine translation systems in fact.

13. Further research directions, if any:

- We will investigate the proposed methods in the thesis for other low-resources language pairs such as Lao-Vietnamese, Khmer-Vietnamese, Malaysian-Vietnamese, etc.
- We will leverage existing large language models (such as ChatGPT, PhoBert), multilingual translation to enhance translation quality for language pairs in the low-resources issue.
- We shall also consider the incorporation of linguistic knowledge to improve machine translation quality for low-resource language pairs.
- We will enhance the proposed methods in the thesis and propose more new methods to improve the translation quality of rare words, names, and terms.

14. Thesis-related publications:

[1]. **Thi-Vinh Ngo**, Van-Tan Bui, Phuong-Thai Nguyen, and Le-Minh Nguyen. Improving Multilingual Neural Machine Translation with Artificial Labels. Proceedings of the 12th International Symposium on Information and Communication Technology (SOICT 2023), Pages 79-84, DOI: https://doi.org/10.1145/3628797.3628964, Association for Computing Machinery (ACM).

[2]. **Thi-Vinh Ngo**, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha, and Le-Minh Nguyen (2022), "An Efficient Method for Generating Synthetic Data For Low-Resource Machine Translation: An empirical study of Chinese, Japanese to Vietnamese Neural Machine Translation", Applied Artificial Intelligence, Volume 36, Issue 1, 2022, Open Access, DOI: 10.1080/08839514.2022.2101755, Taylor & Francis, SCIE.
[3]. Minh-Cong Nguyen-Hoang, **Thi-Vinh Ngo**, Van-Vinh Nguyen (2022), "A Simple and Fast Strategy for Handling Rare Words in Neural Machine Translation", In Proceedings of the 2st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop (SRW AACL-IJCNLP 2022). November 21-23, 2022, Taiwan, China, Online. Association for Computational Linguistics.
[4]. **Ngô Thị Vinh**, Nguyễn Phương Thái (2021), "Nâng cao hiệu quả dịch từ hiếm cho cặp ngôn ngữ Trung-Việt và Nhật-Việt", Hội thảo quốc gia lần thứ XXIV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông (VNICT 2021), Trang 325-330, Thái Nguyên, 13-14/12/2021. Nhà xuất bản khoa học kỹ thuật.
[5]. **Thi-Vinh Ngo**, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh, and Le-Minh Nguyen (2020), "Improving Multilingual Neural Machine Translation For Low-Resource Languages: French, English - Vietnamese", In Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages (LoResMT 2020), pages 55–61, Suzhou, China, Online. Association for Computational Linguistics.
[6]. **Thi-Vinh Ngo**, Minh-Thuan Nguyen, Minh Cong Nguyen Hoang, Hoang-Quan Nguyen, Phuong-Thai Nguyen, Van-Vinh Nguyen (2020), "The UET-ICTU Submissions to the VLSP 2020 News Translation Task". In Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing (VLSP 2020), pages 71–76, December 18, 2020, Hanoi, Vietnam. Association for Computational Linguistics.
[7]. **Thi-Vinh Ngo**, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen (2019), "How Transformer Revitalizes Character-based Neural Machine Translation: An Investigation on Japanese-Vietnamese Translation Systems". In Proceedings of the 16th International Conference on Spoken Language Translation (IWSLT 2019), November 2-3, 2019, Hong Kong, China. Association for Computational Linguistics.
[8]. **Thi-Vinh Ngo**, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen (2019), "Overcoming the Rare Word Problem for Low-Resource Language Pairs in Neural Machine Translation". In Proceedings of the 6th Workshop on Asian Translation (WAT

2019), pages 207–214, November 4, 2019, Hong Kong, China. Association for Computational Linguistics.
[9]. **Thi-Vinh Ngo**, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen (2018), "Combining Advanced Methods in Japanese-Vietnamese Neural Machine Translation", 10th International Conference on Knowledge and Systems Engineering (KSE 2018), pages 318–322, November 1-3, 2018, Ho Chi Minh, Vietnam. Springer.
[10]. **Ngô Thị Vinh** (2018), "Dịch máy Nhật-Việt sử dụng mô hình mạng nơron học sâu", Tạp chí khoa học công nghệ Đại học Thái Nguyên, Tập 178, số 2, Trang 9-14.

Date: January 22th, 2024

Signature: ……

Full name: Nguyen Phuong Thai

Date: January 22th, 2024

Signature: ……

Full name: Ngô Thị Vinh

.