VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name : Pham Nghia Luân 2. Sex: Male

3. Date of birth: March 17, 1983 4. Place of birth: Hai Phong

5. Admission decision number: 642/QĐ-CTSV Dated : September 15th, 2014

6. Changes in academic process:

7. Official thesis title: ***Research on Domain Adaptation techniques in English-Vietnamese Statistical Machine Translation***

8. Major: Information system 9. Code: 9480104.01

10. Supervisors:

1. PhD. Nguyen Van Vinh
2. PhD. Pham Viet Thang

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

The dissertation has achieved some main results as follows:

- Firstly, proposing a method for refining the phrase-table for SMT machine translation system. As the phrase-table contains a list of translation probabilities of phrases from the source language to the target language in both translation directions, these probabilities are automatically learned from bilingual data. The proposed method performs domain classification for phrases in the phrase-table, thereby adjusting and updating the translation probabilities of these phrases in a more prioritized direction in the target domain.
- Secondly, proposing a method for automatically generating bilingual data for machine translation. As NMT machine translation is always in a state of lacking bilingual training data, especially in the domain of bilingual data. Therefore, in the dissertation, a method using Google translate as a component model in the steps of the back-translation technique was proposed to generate fake bilingual data.
- Finally, the proposal for improving the quality of automatically generated pseudo-language data in the second proposal is presented. As the input of the second proposal is a text, but this text is often noisy due to possible errors in spelling and grammar,

which affects the quality of the output, the proposed method contributes to reducing noise by automatically correcting spelling and grammar errors for the input text. This proposal helps to improve the quality of automatically generated pseudo-language data.

12. Practical applicability, if any:

Currently, machine translation is increasingly widely applied. The statistical machine translation method is currently the best approach. Moreover, each domain must have a suitable approach and translation strategy, so researching domain adaptation in statistical machine translation is meaningful both scientifically and practically.

13. Further research directions, if any:

In the near future, graduate students will focus on research to address some of the remaining limitations of the thesis. Specifically, they will focus on conducting the following studies:

- Methods for collecting, and creating monolingual and bilingual data, domain data, and improving data quality.
- Improving the quality of machine translation by leveraging knowledge from collected data sources and applying large language models.
- Enhancing the quality of automatic English-Vietnamese translation across multiple domains and expanding the ability to translate between Vietnamese and other languages by building, developing, and improving multilingual and multi-domain translation systems.

14. Thesis-related publications:

- Nguyễn Quang Huy, Nguyễn Văn Vinh, Phạm Nghĩa Luân, Nguyễn Quỳnh Anh (2014). "Nghiên cứu phương pháp dóng hàng câu cho cặp ngôn ngữ Anh – Việt". *Hội thảo quốc gia lần thứ XVII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, trang 188-195.
- Phạm Nghĩa Luân, Nguyễn Văn Vinh, Nguyễn Quang Huy (2015). "Một phương pháp thích ứng miền cho dịch máy thống kê". *Hội thảo quốc gia lần thứ XVIII: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, trang 174-180.
- Viet Tran Hong, Huyen Vu Thuong, Trung Le Tien, Luan Nghia Pham and Vinh Nguyen Van (2015). "The English - Vietnamese Machine Translation System for IWSLT 2015". *In Proceedings of the 12th International Workshop on Spoken Language Translation*, pp. 80-83. (SCOPUS).

- Phạm Nghĩa Luân, Nguyễn Văn Vinh (2019). "Thích ứng miền trong dịch máy nơ ron cho cặp ngôn ngữ Anh - Việt". *Hội nghị khoa học quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR 2019)*, pp.436-442.
- Nghia Luan Pham and Van Vinh Nguyen (2019). "Adapting Neural Machine Translation for English-Vietnamese using Google Translate system for Back-translation". *In the 33rd Pacific Asia Conference on Language, Information and Computation*, pp. 567-575 (SCOPUS).
- Nghia Luan Pham, Tien Ha Nguyen and Van Vinh Nguyen (2019). "Grammatical error correction for Vietnamese using Machine Translation". *In 16th International Conference of the Pacific Association for Computational Linguistics*, pp.505-512. ISBN 978-981-15-6167-2. DOI: https://doi.org/10.1007/978-981-15-6168-9_41 (SCOPUS).
- Nghia Luan Pham and Van Vinh Nguyen (2020). "Adaptation in Statistical Machine Translation for low-resource domains in English-Vietnamese language". *In VNU Journal of Science: Computer Science and Communication Engineering*, [S.l.], v.36, n.1. ISSN 2588-1086.
- Nghia Luan Pham, Van Vinh Nguyen and Thang Viet Pham (2023). "A Data Augmentation Method For English-Vietnamese Neural Machine Translation," *In IEEE Access*, vol. 11, pp. 28034-28044, 2023, doi: 10.1109/ACCESS.2023.3252898 (Q1, SCOPUS).

Date: June 9, 2023	Date: June 9, 2023
Signature: ……	Signature: ……
Full name: Nguyen Van Vinh	Full name: Pham Nghia Luan