THÔNG TIN VỀ LUẬN ÁN TIẾN SĨ

1. Họ và tên nghiên cứu sinh: Nguyễn Tiến Hà 2.Giới tính: Nam

3. Ngày sinh: 04/08/1977 4. Nơi sinh: Vĩnh Phúc

5. Quyết định công nhận nghiên cứu sinh: Quyết định số 4374/QĐ-KHTN-CTSV ngày 03/12/2012 của Hiệu trưởng Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.

6. Các thay đổi trong quá trình đào tạo: Quyết định gia hạn đào tạo và bảo vệ luận án tiến sĩ số 741/QĐ-ĐHKHTN ngày 31/03/2016 và số 1034/QĐ-ĐHKHTN ngày 25/04/2017; Quyết định trả NCS về địa phương số 5034/QĐ-ĐHKHTN ngày 29/12/2017 của trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.

7. Tên đề tài luận án: **Nghiên cứu xây dựng tài nguyên song ngữ Việt-Anh ứng dụng cho dịch máy theo miền.**

8. Chuyên ngành: Cơ sở Toán học cho Tin học 9. Mã số: 9460117.02

10. Cán bộ hướng dẫn khoa học: Hướng dẫn chính: TS. Nguyễn Thị Minh Huyền

Hướng dẫn phụ: PGS. TS. Nguyễn Hữu Ngự

11. Tóm tắt các kết quả mới của luận án:

Với việc thực hiện đề tài nghiên cứu này chúng tôi đã thu được một số kết quả sau:

- Đề xuất kỹ thuật cải tiến công cụ dóng hàng XAlign cho cặp ngôn ngữ Việt-Anh. Sử dụng công cụ dóng hàng này luận án đã xây dựng được kho ngữ liệu song ngữ có dóng hàng câu gồm trên 20.000 cặp câu miền du lịch và trên 270.000 cặp câu miền chung. Luận án đã chứng tỏ bằng thực nghiệm rằng việc khai thác các kho ngữ liệu này nâng cao đáng kể chất lượng dịch máy theo miền du lịch.

- Đề xuất phương pháp trích rút từ và cụm từ song ngữ từ kho ngữ liệu song ngữ và kho ngữ liệu đơn ngữ. Sử dụng các phương pháp này luận án đã xây dựng được kho ngữ liệu trên 40.000 cặp từ và cụm từ song ngữ, bao gồm: trên 1.000 cặp cho miền du lịch; trên 600 cặp cho miền y tế; còn lại thuộc miền chung.

- Đề xuất kỹ thuật tiền xử lý câu dài trong dịch máy nơ-ron cải thiện chất lượng dịch.

- Triển khai một phương pháp sinh chú giải tiếng Việt tự động cho hình ảnh dựa vào dịch máy Anh-Việt, đề xuất một kỹ thuật khai thác từ điển để xử lý các từ mới (unknown words) nhằm nâng cao chất lượng hệ thống dịch.

12. Khả năng ứng dụng thực tiễn:

Các công cụ và ngữ liệu song ngữ xây dựng trong luận án, bao gồm kho văn bản song ngữ và các kho từ/cụm từ song ngữ, có thể ứng dụng trong việc nâng cao chất lượng của các hệ thống dịch Anh-Việt. Kết quả của luận án liên quan tới sinh chú thích tiếng Việt cho ảnh dựa vào dịch máy cũng có thể được ứng dụng nhằm nâng cao hiệu suất xây dựng các bộ dữ liệu ảnh có chú thích, phục vụ nghiên cứu ứng dụng trong lĩnh vực trí tuệ nhân tạo.

13. Các hướng nghiên cứu tiếp theo:

- Nghiên cứu cải tiến hiệu năng dịch máy thông qua việc khai thác tài nguyên đơn ngữ và song ngữ tổng quát cũng như theo miền.

- Nghiên cứu thu thập và khai thác các nguồn tài nguyên đa ngữ (nhiều hơn một cặp ngôn ngữ).

- Nghiên cứu các vấn đề liên quan tới việc nâng cao chất lượng biểu diễn từ và biểu diễn ngữ nghĩa đa ngữ hướng tới các hệ thống dịch máy đa ngữ.

14. Các công trình công bố liên quan đến luận án:

[1] **Nguyễn Tiến Hà**, Nguyễn Thị Minh Huyền, Nguyễn Minh Hải (2018), "Xây dựng kho ngữ liệu du lịch song ngữ Việt - Anh dóng hàng mức câu cho dịch máy", *Tạp chí các công trình nghiên cứu phát triển công nghệ thông tin và truyền thông* Tập V-1, số 39, Bộ thông tin và truyền thông, tr. 9-16.

[2] **Nguyễn Tiến Hà**, Nguyễn Thị Minh Huyền (2019), "Tiền xử lý câu dài trong dịch máy nơ-ron", *Kỷ yếu Hội nghị quốc tế RIVF 2019 về Công nghệ Truyền thông và Điện toán*, DOI: 10.1109/RIVF.2019.8713737, tr. 1-6.

[3] **Nguyễn Tiến Hà**, Nguyễn Thị Minh Huyền (2019), "Xây dựng tự động từ điển Việt – Anh và ứng dụng trong lĩnh vực du lịch", *Kỷ yếu Hội nghị Quốc gia lần thứ VII về Nghiên cứu cơ bản và ứng dụng Công Nghệ thông tin (FAIR)*, tr. 568-576.

[4] **Nguyễn Tiến Hà**, Ngô Thế Quyền, Nguyễn Thị Minh Huyền, Hà Mỹ Linh (2019), "Trích rút thuật ngữ song ngữ Anh-Việt từ văn bản đơn ngữ tiếng Việt dựa vào luật", *Kỷ yếu Hội nghị quốc tế SoICT 2019 về Công nghệ thông tin và Truyền thông lần thứ 10*, tr. 56–62.

[5] Phạm Nghĩa Luân, **Nguyễn Tiến Hà**, Nguyễn Văn Vĩnh (2019), "Chữa lỗi ngữ pháp cho tiếng Việt sử dụng dịch máy", *Kỷ yếu Hội nghị quốc tế PACLING lần thứ XVI*, Hà Nội, Việt Nam, tr. 505-512.

[6] **Nguyễn Tiến Hà**, Đỗ Thanh Hà (2020), "Sinh chú giải tiếng Việt tự động cho ảnh", *Kỷ yếu Hội nghị quốc tế MAPR 2020 về Phân tích thông tin đa phương tiện và nhận dạng lần thứ 3*, 978-1-7281-6555-4/20/$31.00 ©2020 IEEE.

[7] **Nguyễn Tiến Hà**, Đỗ Thanh Hà, Nguyễn Vân Anh (2020), "Chú giải tiếng Việt cho ảnh dựa vào mạng nơ-ron", *đã được chấp nhận báo cáo nói tại Hội nghị quốc tế ICCCI 2020 về trí tuệ nhóm lần thứ 12*.

Hà Nội, ngày tháng năm 2020

Người hướng dẫn luận án **Nghiên cứu sinh**

TS. Nguyễn Thị Minh Huyền PGS.TS. Nguyễn Hữu Ngự Nguyễn Tiến Hà

INFORMATION ON DOCTORAL THESIS

1. Full name: Nguyen Tien Ha 2. Sex: Male

3. Date of birth: 04/08/1977 4. Place of birth: Vinh Phuc

5. Admission decision number: No. 4374/QĐ-KHTN-CTSV, dated on 03rd December, 2012 by Rector of VNU University of Science, VNU

6. Changes in academic process: Extension education time according to: Decision No. 741/QĐ-ĐHKHTN dated 31st March, 2016 and Decision No. 1034/QĐ-ĐHKHTN dated 25th April, 2017; Local return: Decision No. 5034/QĐ-ĐHKHTN dated 29th December, 2017 signed by the Rector of VNU University of Science.

7. Official thesis title: Building English-Vietnamese language resources for domain specific machine translation

8. Major: Mathematical Foundation for Computer Science Code: 9460117.02

10. Supervisors: Dr. Nguyen Thi Minh Huyen and Assoc. Prof. Dr. Nguyen Huu Ngu

11. Summary of the new findings of the thesis

- Proposition of a technique to improve the alignment tool XAlign for the Vietnamese-English language pair. Two datasets have been built using this tool:: a Vietnamese-English sentence-aligned bilingual corpus in the tourism domain of more than 20,000 pairs of sentences and a Vietnamese-English sentence-aligned bilingual corpus in the general domain of more than 270,000 pairs of sentences. The thesis has empirically proved that the exploitation of these corpuses significantly improves the quality of machine translation in the tourism domain.

- Proposition of methods for extracting bilingual terms from bilingual and monolingual corpora. By using these methods, we have got more than 40,000 term pairs including more than 1,000 pairs in the tourism domain, more than 500 pairs in the medical domain and the remaining pairs in the general domain.

- Proposition of a long-sentence pre-processing technique for enriching the training data and consequently improving the quality of machine translation systems.

- Implementation of a method for automatically generating Vietnamese captions of images based on English-Vietnamese machine translation. Proposition of a dictionary-based technique for handling unknown words to improve the quality of this machine translation system.

12. Practical applicability, if any:

Tools and bilingual resources built in the thesis, including bilingual corpora and bilingual terminologies, can be applied in improving the quality of English-Vietnamese translation systems. The results of the thesis related to generation of Vietnamese captions of images based on machine translation can also be applied for building annotated image datasets, which are useful for researches and applications in the field of artificial intelligence.

13. Further research directions, if any

In the future, in addition to research on improving the machine translation performance by exploiting bilingual and monolingual resources in the general domain, the problem of building and exploiting multilingual resources is also worth being paid much attention. The problems related to improving the quality of word representation and multilingual semantic representation also need to be further studied.

14. Thesis-related publications:

[1] **Ha Nguyen Tien**, Huyen Nguyen Thi Minh, Hai Nguyen Minh (2018), "Building a sentence-aligned Vietnamese–English bilingual corpus in the tourism domain for machine translation". *Journal of Research and Development on Information and Communication Technology*, Vol V-1, Number 39, Ministry of Information and Communications of Vietnam, pp. 9-16.

[2] **Ha Nguyen Tien,** Huyen Nguyen Thi Minh (2019), "Long Sentence Preprocessing in Neural Machine Translation". *In Proceedings of the 2019 IEEE-RIVF International Conference on Computing and Communication Technologies*, DOI: 10.1109/RIVF.2019.8713737, pp. 1-6.

[3] **Ha Nguyen Tien,** Huyen Nguyen Thi Minh (2019), "Automatic construction of the Vietnamese-English bilingual dictionary and application in tourism domain", *In Proceedings of the 7th National Conference on Fundamental and Applied Information Technology (FAIR)*, DOI: 10.15625/vap.2019.00073, pp. 568-576.

[4] **Ha Nguyen Tien**, Quyen Ngo The, Huyen Nguyen Thi Minh and Linh Ha My (2019), "Rule based English-Vietnamese bilingual terminology extraction from Vietnamese documents", *In Proceedings of The Tenth International Symposiumon Information and Communication Technology (SoICT 2019)*, pp. 56–62.

[5] Luan Nghia Pham, **Ha Nguyen Tien** and Vinh Van Nguyen (2019), "Grammatical error correction for Vietnamese using Machine Translation", *In Proceedings of 16th International Conference of the Pacific Association for Computational Linguistics (PACLING 2019)*, pp. 505-512.

[6] **Ha Nguyen Tien**, Thanh-Ha Do (2020), "Generating Vietnamese Language Caption Automatically for Scene Images", *In Proceedings of the Third International Conference on Multimedia Analysis and Pattern Recognition*. 978-1-7281-6555-4/20/$31.00 ©2020 IEEE.

[7] **Ha Nguyen Tien**, Thanh-Ha Do, Van-Anh Nguyen (2020), "Image Captioning in Vietnamese Language based on Deep Learning Network", *accepted for oral presentation at the conference and published in the Springer LNCS/LNAI proceedings.*

Date: ……

Supervisors **PhD Student**

Dr.Nguyen Thi Minh Huyen Assoc.Prof.Dr. Nguyen Huu Ngu Nguyen Tien Ha