ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

THÔNG TIN VỀ LUẬN ÁN TIẾN SĨ

1. Họ và tên nghiên cứu sinh: **Trần Mai Vũ** 2. Giới tính: **Nam**

3. Ngày sinh: **25/08/1984** ..... 4. Nơi sinh: **Thừa Thiên Huế**

5. Quyết định công nhận nghiên cứu sinh số:3205/QĐ-SĐH, ngày 08 tháng 11 năm 2010 của Đại học Quốc gia Hà Nội

6. Các thay đổi trong quá trình đào tạo: .....

7. Tên đề tài luận án: **Nghiên cứu nhận dạng thực thể có tên và thực thể biểu hiện trong văn bản và ứng dụng**

8. Chuyên ngành: Hệ thống thông tin ..... 9. Mã số: 62.48.05.01 .....

10. Cán bộ hướng dẫn khoa học: PGS.TS. Hà Quang Thụy và PGS.TS. Nguyễn Lê Minh

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

- Đề xuất mô hình kết hợp nhận dạng đồng thời thực thể và các thuộc tính liên quan đến thực thể, mô hình cho phép sử dụng nhiều loại đặc trưng khác nhau nhằm tăng cường tính ngữ nghĩa và hiệu quả của quá trình nhận dạng. Một tập dữ liệu với gần 10.000 câu đã được gán nhãn thực thể và thuộc tính cũng được xây dựng phục vụ cho việc huấn luyện và đánh giá. Kết quả của mô hình nhận dạng đạt 83,39 với độ đo F1.

- Xây dựng một hệ thống hỏi đáp tự động ứng dụng mô hình nhận dạng thực thể và thuộc tính đã được đề xuất. Các bước phân tích câu hỏi và trả lời câu hỏi đều cho thấy tầm quan trọng của mô hình nhận dạng thực thể đối với mô hình hỏi đáp. Kết quả của mô hình tương đối khả quan với độ đo F1 đạt 65,5.

- Góp phần mở rộng khung cấu trúc thực thể y sinh, thống nhất và tổng quát lại các định nghĩa về các thực thể y sinh có liên quan đến nhau như bệnh, hóa chất, gene, sinh vật, kiểu biểu hiện và bộ phân cơ thể. Đề xuất mô hình giải quyết bài toán nhận dạng thực thể kiểu biểu hiện và các thực thể liên quan, đây là loại thực thể mới trong y sinh với các tính chất phức tạp về mặt ngữ nghĩa. Mô hình giải quyết đạt kết quả khả quan với tất các thực thể có trong lược đồ nhận dạng.

- Đưa ra các so sánh, nhận định về vấn đề thích nghi miền dữ liệu đối với việc nhận dạng thực thể y sinh, các kết quả cho phép những nghiên cứu sau này về nhận dạng thực

thể kiểu biểu hiện có một khung nhìn tổng quát trong quá trình chọn lựa dữ liệu huấn luyện và đánh giá.

- Nâng cao chất lượng nhận dạng thực thể kiểu biểu hiện và thực thể y sinh liên quan bằng kỹ thuật lai ghép, kết hợp nhiều mô hình nhận dạng khác nhau. Luận án đề xuất 3 phương pháp lai ghép, kết hợp và đưa ra các đánh giá, nhận xét về các phương pháp này. Các kết quả đã chỉ ra được tính hiệu quả của các phương pháp lai ghép so với kỹ thuật nhận dạng đơn mô hình khi làm tăng kết quả lên 1,5% với độ đo F.

12. Khả năng ứng dụng trong thực tiễn: (*nếu có*).....

13. Những hướng nghiên cứu tiếp theo: (*nếu có*)

a. Mô hình nhận dạng thực thể tiếng Việt vẫn còn một số lớp nhận dạng có kết quả chưa cao do vấn đề mất cân bằng dữ liệu trong tập huấn luyện. Để giải quyết vấn đề này có thể áp dụng một số kỹ thuật làm giảm sự ảnh hưởng giữa các lớp có số lượng dữ liệu lớn đến các lớp có số lượng dữ liệu nhỏ hơn hay áp dụng một số kỹ thuật lựa chọn đặc trưng.

b. Áp dụng bài toán nhận dạng thực thể biểu hiện và các thực thể liên quan cho dữ liệu văn bản y sinh thực tế, bên cạnh đấy ứng dụng các phương pháp trích xuất quan hệ nhằm làm rõ sự tương tác giữa các thực thể với nhau.

c. Thử nghiệm phương pháp thích nghi miền với nhiều miền dữ liệu hơn để cho thấy sự tác động về mặt hiệu quả giữa các miền dữ liệu qua đấy đề xuất một mô hình cho phép nhận dạng được thực thể biểu hiện cho tất cả các loại bệnh di truyền.

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

- Hoang-Quynh Le, Mai-Vu Tran, Thanh Hai Dang, Nigel Collier (2015). The UET-CAM System in the BioCreAtIvE V CDR Task. In Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, 2015.
- Nigel Collier, Ferdinand Paster, Mai-Vu Tran (2014). The impact of near domain transfer on biomedical named entity recognitions *LOUHI 2014, EACL 2014*, Sweden, 2014.
- Nigel Collier, Mai-Vu Tran, Hoang-Quynh Le, Quang-Thuy Ha, Anika Oellrich, Dietrich Rebholz-Schuhmann (2013). Learning to Recognize Phenotype Candidates in the Auto-Immune Literature Using SVM Re-Ranking. *PLoS ONE* 8(10): e72965, October 2013.
- Mai-Vu Tran, Nigel Collier, Hoang-Quynh Le, Van-Thuy Phi and Thanh-Binh Pham (2013). Exploing a Probabilistic Earley Parser for Event Composition in Biomedical Texts, *BIONLP-ST*:130-134, 2013.

- Mai-Vu Tran, Duc-Trong Le (2013). vTools: Chunker and Part-of-Speech tools, *RIVF-VLSP 2013 Workshop*.
- Nigel Collier, Mai-Vu Tran, Hoang-Quynh Le, Anika Oellrich, Ai Kawazoe, Martin Hall-May, Dietrich Rebholz-Schuhmann (2012). A Hybrid Approach to Finding Phenotype Candidates in Genetic Texts, *COLING 2012:* 647-662.
- Mai-Vu Tran, Duc-Trong Le, Xuan-Tu Tran and Tien-Tung Nguyen (2012). A Model of Vietnamese Person Named Entity Question Answering System, *PACLIC 2012*, Bali, Indonesia, October 2012.
- Mai-Vu Tran, Minh-Hoang Nguyen, Sy-Quan Nguyen, Minh-Tien Nguyen, Xuan-Hieu Phan (2012). VnLoc: A Real–time News Event Extraction Framework for Vietnamese, *KSE'2012:*161-166, Da Nang, August 17-19, 2012.
- Huyen-Trang Pham, Tien-Thanh Vu, Mai-Vu Tran, Quang-Thuy Ha (2011). A Solution for Grouping Vietnamese Synonym Feature Words in Product Reviews. *APSCC 2011*: 503-508.
- Hoang-Quynh Le, Mai-Vu Tran, Nhat-Nam Bui, Nguyen-Cuong Phan, Quang-Thuy Ha (2011). An Integrated Approach Using Conditional Random Fields for Named Entity Recognition and Person Property Extraction in Vietnamese Text. *IALP 2011*:115-118.
- Mai-Vu Tran, Tien-Tung Nguyen, Thanh-Son Nguyen, Hoang-Quynh Le (2010). Automatic Named Entity Set Expansion Using Semantic Rules and Wrappers for Unary Relations. *IALP 2010*: 170-173.

Ngày 21 tháng 08 năm 2017 *Ngày 15 tháng 08 năm* 2017

Xác nhận của cán bộ hướng dẫn **Nghiên cứu sinh**

Thay mặt tập thể hướng dẫn

PGS.TS. Hà Quang Thụy Trần Mai Vũ

VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name: TRAN MAI VU ..... 2. Sex: MALE

3. Date of birth: 25/08/1984 ..... 4. Place of birth: THUA THIEN HUE..

5. Admission decision number: 3205/QĐ-SĐH Dated 08/11/2010 .....

6. Changes in academic process: ....

(List the forms of change and corresponding times)

7. Official thesis title: .....

8. Major: ..... 9. Code: .....

10. Supervisors: .....

(Full name, academic title and degree)

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

Main achieved science results:

- *Proposing the model for recognizing entities and their attributes simultaneously. This model allows using various features for enhancing the semantic characteristics and the model effectiveness. A dataset of 10.000 sentences was annotated with entities and their attributes. The results of proposed model reached 83.39 of F1.*
- *Developing a Vietnamese question answering system (QAS) which based on proposed model of recognizing entities and their attributes. Both steps of question analyzing and answering show the importance of NER in QAS. The results of QAS is positive, reached 65.5 of F1.*
- *Contributing to expanding biomedical entities structure frame, consisting and generalizing the definitions of related biomedical entities, such as disease, chemical, gene, organism, body part and phenotype.*
- *Proposing model to recognize phenotype and related entities, in which, phenotype is a semantically complex entity. The model reached positive results in all phenotype of our scheme.*

- Giving the comparisons, judgments about problem of domain adaptive for biomedical NER, the results give future biomedical NER researches a general view to select data for training and evaluating.

- Improving the effectiveness of biomedical NER system by building the hybrid model which is the combining of several different NER methods. Thesis proposed 3 hybrid approaches and evaluated them. The results showed the effectiveness of the hybrid approach than single NER techniques that increase the model results 1.5% of F1.

12. Practical applicability, if any: .....

13. Further research directions, if any: .....

14. Thesis-related publications: .....

- Hoang-Quynh Le, Mai-Vu Tran, Thanh Hai Dang, Nigel Collier (2015). The UET-CAM System in the BioCreAtIvE V CDR Task. In Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain, 2015.
- Nigel Collier, Ferdinand Paster, Mai-Vu Tran (2014). The impact of near domain transfer on biomedical named entity recognitions *LOUHI 2014, EACL 2014*, Sweden, 2014.
- Nigel Collier, Mai-Vu Tran, Hoang-Quynh Le, Quang-Thuy Ha, Anika Oellrich, Dietrich Rebholz-Schuhmann (2013). Learning to Recognize Phenotype Candidates in the Auto-Immune Literature Using SVM Re-Ranking. *PLoS ONE* 8(10): e72965, October 2013.
- Mai-Vu Tran, Nigel Collier, Hoang-Quynh Le, Van-Thuy Phi and Thanh-Binh Pham (2013). Exploing a Probabilistic Earley Parser for Event Composition in Biomedical Texts, *BIONLP-ST*:130-134, 2013.
- Mai-Vu Tran, Duc-Trong Le (2013). vTools: Chunker and Part-of-Speech tools, *RIVF-VLSP 2013 Workshop*.
- Nigel Collier, Mai-Vu Tran, Hoang-Quynh Le, Anika Oellrich, Ai Kawazoe, Martin Hall-May, Dietrich Rebholz-Schuhmann (2012). A Hybrid Approach to Finding Phenotype Candidates in Genetic Texts, *COLING 2012:* 647-662.
- Mai-Vu Tran, Duc-Trong Le, Xuan-Tu Tran and Tien-Tung Nguyen (2012). A Model of Vietnamese Person Named Entity Question Answering System, *PACLIC 2012*, Bali, Indonesia, October 2012.
- Mai-Vu Tran, Minh-Hoang Nguyen, Sy-Quan Nguyen, Minh-Tien Nguyen, Xuan-Hieu Phan (2012). VnLoc: A Real–time News Event Extraction Framework for Vietnamese, *KSE'2012:*161-166, Da Nang, August 17-19, 2012.

- Huyen-Trang Pham, Tien-Thanh Vu, Mai-Vu Tran, Quang-Thuy Ha (2011). A Solution for Grouping Vietnamese Synonym Feature Words in Product Reviews. *APSCC 2011*: 503-508.
- Hoang-Quynh Le, Mai-Vu Tran, Nhat-Nam Bui, Nguyen-Cuong Phan, Quang-Thuy Ha (2011). An Integrated Approach Using Conditional Random Fields for Named Entity Recognition and Person Property Extraction in Vietnamese Text. *IALP 2011*:115-118.
- Mai-Vu Tran, Tien-Tung Nguyen, Thanh-Son Nguyen, Hoang-Quynh Le (2010). Automatic Named Entity Set Expansion Using Semantic Rules and Wrappers for Unary Relations. *IALP 2010*: 170-173.

Supervisors **Author**

Date: 21/08/2017
Signature: ……
Full name: Assoc. Prof. Ha Quang Thuy

Date: 15/08/2017
Signature: ……
Full name: Tran Mai Vu