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: **Lê Quang Hùng** 2. Giới tính: Nam

3. Ngày sinh:10/10/1981 4. Nơi sinh: Quảng Bình

5. Quyết định công nhận NCS số 1142/QĐ-CTSV ngày 28/12/2011 của Hiệu trưởng Trường Đại học Công nghệ.

6. Các thay đổi trong quá trình đào tạo: Thay đổi tập thể cán bộ hướng dẫn NCS theo Quyết định số 48/QĐ-ĐT ngày 25 tháng 01 năm 2013 của Hiệu trưởng Trường Đại học Công nghệ.

7. Tên đề tài luận án: Khai phá tri thức song ngữ và ứng dụng trong dịch máy Anh-Việt

8. Chuyên ngành: Khoa học máy tính 9. Mã số: 62 48 01 01

10. Cán bộ hướng dẫn khoa học:

- Hướng dẫn chính: **PGS.TS. Lê Anh Cường**
- Hướng dẫn phụ: **PGS.TS. Huỳnh Văn Nam**

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

- Trích chọn thêm dữ liệu song ngữ: chúng tôi đề xuất một số phương pháp để xây dựng ngữ liệu song ngữ cho dịch máy thống kê (Statistical Machine Translation - SMT) từ Web và sách điện tử song ngữ. Đối với nguồn từ Web, chúng tôi đề xuất hai phương pháp thiết kế các đặc trưng dựa trên nội dung: sử dụng cognate và sử dụng các phân đoạn dịch. Đối với nguồn từ sách điện tử, chúng tôi đề xuất phương pháp dựa trên nội dung, sử dụng một số mẫu liên kết giữa các khối văn bản trong hai ngôn ngữ để rút trích các câu song ngữ.
- Cải tiến mô hình gióng hàng dựa vào ràng buộc: chúng tôi đề xuất một số cải tiến mô hình IBM theo cách tiếp cận dựa trên ràng buộc, bao gồm: ràng buộc neo, ràng buộc về vị trí của từ, ràng buộc về từ loại và ràng buộc về cụm từ. Các ràng buộc này sau đó được sử dụng để ước lượng các tham số của mô hình trong thuật toán cực đại kỳ vọng.
- Xác định các cụm song ngữ Anh – Việt: chúng tôi đề xuất phương pháp xác định cụm từ song ngữ cho dịch máy thống kê. Trước hết, chúng tôi sẽ sử dụng tập các mẫu cú pháp ở một ngôn ngữ để phát hiện cụm từ nguồn. Sau đó, chúng tôi tìm bản dịch của cụm từ nguồn sử dụng mô hình gióng hàng từ ràng buộc.

12. Khả năng ứng dụng trong thực tiễn:

Các kết quả của luận án nhằm tăng chất lượng dịch cho hệ dịch máy Anh – Việt. Các kết quả đã được đăng ở tạp chí và hội nghị quốc tế, vì vậy có đóng góp cho khoa học trong lĩnh vực xử lý ngôn ngữ tự nhiên và có ý nghĩa trong việc phát triển các hệ dịch máy, có khả năng sử dụng trong thực tế.

13. Những hướng nghiên cứu tiếp theo:

- Thứ nhất, đối với việc xây dựng ngữ liệu song ngữ, chúng tôi dự định sẽ tiếp tục khai thác nguồn từ Web với các thành phần song ngữ khác như đoạn, câu và cụm từ. Công việc này cũng sẽ rất có ý nghĩa trong trường hợp chất lượng dịch giữa các trang web song ngữ không tốt.
- Thứ hai, với bài toán gióng hàng từ cho SMT, chúng tôi sẽ xử lý các ngoại lệ xảy ra đối với các ràng buộc như đã đề xuất.
- Thứ ba, đối với bài toán xác định cụm từ song ngữ cho SMT, chúng tôi sẽ sử dụng nhiều hơn nữa các mẫu cú pháp để tăng số lượng các cụm từ được rút trích. Ngoài ra, chúng tôi dự định sử dụng phương pháp đã đề xuất để xây dựng một từ điển cụm từ song ngữ cho cặp ngôn ngữ Anh-Việt và sẽ cải thiện chất lượng của SMT Anh-Việt bằng cách sử dụng từ điển này.

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

[1] Le Quang Hung and Le Anh Cuong (2010), "Extracting parallel texts from the web", *Proceedings of the Second International Conference on Knowledge and Systems Engineering, IEEE Computer Society*, pages 147-151.

[2] Le Quang Hung and Le Anh Cuong (2012), "Improving Word Alignment for Statistical Machine Translation Based on Constraints", *Asian Language Processing (IALP), International Conference on, IEEE Computer Society*, pages 113-116.

[3] Le Quang Hung and Le Anh Cuong (2012), "Statistical Word Alignment with Part-of-Speech Constraint", *Kỷ yếu hội thảo Quốc gia lần thứ XV "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"*, trang 410-416.

[4] Quang-Hung LE, Duy-Cuong NGUYEN, Duc-Hong PHAM, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Paragraph Alignment for English-Vietnamese Parallel E-Books", In *Knowledge and Systems Engineering, Springer International Publishing*, pages 251-259.

[5] Quang-Hung LE, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Parallel phrase extraction from English-Vietnamese parallel corpora", In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, pages 175-179.

[6] Le Quang Hung and Le Anh Cuong (2013), "An effective method to sentence alignment for the English-Vietnamese parallel e-book", *Kỷ yếu hội thảo Quốc gia lần thứ XVI "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông*", trang 12-16.

[7] Le Quang Hung (2014), "A new approach to extract parallel corpus", *Tạp chí khoa học Trường Đại học Quy Nhơn*, Số 4 - Tập VIII, trang 12-24.

[8] Quang-Hung LE and Anh-Cuong LE (2014), "Syntactic pattern based Word Alignment for Statistical Machine Translation", *The International Journal of Knowledge and Systems Science (IJKSS) IGI Global Publishing*, Volume 5 Issue 3, pages 36-45.

Ngày 09 tháng 09 năm 2015	*Ngày 09 tháng 09 năm 2015*
Xác nhận của cán bộ hướng dẫn	**Nghiên cứu sinh**
(Kí và ghi rõ họ tên)	*(Kí và ghi rõ họ tên)*
PGS.TS. Lê Anh Cường	**Lê Quang Hùng**