TRANG THÔNG TIN LUẬN ÁN TIẾN SĨ

- Tên đề tài luận án tiến sĩ: **Nghiên cứu các phương pháp học máy cho trích xuất thông tin tự động từ văn bản**
- Chuyên ngành: **Hệ thống thông tin**
- Mã số: **9.48.01.04**
- Họ và tên NCS: **Nguyễn Thị Thanh Thủy**
- Người hướng dẫn khoa học:
    1. GS.TS. Từ Minh Phương
    2. PGS.TS. Ngô Xuân Bách
- Cơ sở đào tạo: **Học viện Công nghệ Bưu chính Viễn thông**

NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN:

1. Đề xuất giải pháp nâng cao hiệu quả cho trích xuất khía cạnh và phân loại quan điểm trong ngôn ngữ tiếng Việt bằng cách khai thác nguồn dữ liệu đã được gán nhãn sẵn từ ngôn ngữ khác. Việc xác định các loại khía cạnh và phân loại quan điểm được thực hiện theo từng câu thay vì toàn bộ bài đánh giá sẽ thực tế hơn và có thể áp dụng trong nhiều ứng dụng thế giới thực. Phương pháp đề xuất khá tổng quát và linh hoạt do không phụ thuộc vào ngôn ngữ và các thuật toán học máy, giúp giải quyết khó khăn do việc thiếu tài nguyên dữ liệu huấn luyện trong một số ngôn ngữ có ít tài nguyên cho bài toán này (như tiếng Việt).
2. Đề xuất phương pháp trích xuất thông tin sử dụng học máy truyền thống và học sâu cho văn bản pháp quy tiếng Việt. Các thông tin được trích xuất bao gồm thực thể tham chiếu và mối quan hệ giữa các thực thể văn bản pháp quy. Ngoài việc sử dụng các phương pháp học máy truyền thống, nghiên cứu còn sử dụng các phương pháp học sâu, đồng thời kết hợp lợi thế của mô hình học sâu và các đặc trưng được thiết kế thủ công (theo phương pháp học máy truyền thống).
3. Đề xuất phương pháp trích xuất kết hợp thực thể và quan hệ trong văn bản pháp quy tiếng Việt sử dụng mô hình dựa trên học sâu. Mô hình trích xuất kết hợp sử dụng kiến trúc bộ mã hóa-giải mã dựa trên Transformer với cơ chế giải mã song song không tự hồi quy (non-autoregressive decoding mechanism) để trích xuất đồng thời các thực thể tham chiếu và quan hệ trong văn bản pháp quy (khác với đề xuất thứ hai, thực hiện trích xuất các thông tin này theo cách tuần tự).

CÁC ỨNG DỤNG, KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HOẶC NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU:

1. Kết quả nghiên cứu trong đề xuất thứ nhất cho thấy, với việc sử dụng thêm dữ liệu (đã được gán nhãn) dịch từ ngôn ngữ khác (như tiếng Anh), phương pháp đề xuất đã cải thiện hiệu năng trong cả hai nhiệm vụ trích xuất khía cạnh và phân loại quan điểm tiếng Việt.

Từ kết quả có được, hướng phát triển nghiên cứu tiếp theo có thể tập trung vào các phương pháp tiên tiến sử dụng dữ liệu có sẵn từ các ngôn ngữ khác nhằm cải thiện các nhiệm vụ NLP cho các ngôn ngữ ít tài nguyên (như tiếng Việt). Ví dụ, có thể thực hiện: (1) Tận dụng nhiều bộ dữ liệu của cùng một nhiệm vụ NLP từ các ngôn ngữ khác nhau để giải quyết nhiệm vụ đó cho tiếng Việt; (2) Tận dụng nhiều bộ dữ liệu của các nhiệm vụ NLP khác nhau từ một ngôn ngữ khác để giải quyết các nhiệm vụ đó cho tiếng Việt (nghĩa là học đa tác vụ).

2. Kết quả nghiên cứu trong các đề xuất thứ hai và thứ ba cho thấy, các mô hình đề xuất có thể trích xuất chính xác riêng rẽ cũng như đồng thời cả hai thông tin thực thể tham chiếu và quan hệ giữa các thực thể trong văn bản pháp quy tiếng Việt, đặc biệt cho kết quả tốt hơn so với các mô hình cơ sở khác trong các câu văn bản pháp quy phức tạp có nhiều thực thể tham chiếu. Một số hướng nghiên cứu có thể phát triển: (1) Có thể sử dụng kết quả trích xuất thực thể tham chiếu và phân loại quan hệ giữa các thực thể văn bản pháp quy cho các tác vụ xử lý văn bản pháp quy khác như truy xuất thông tin pháp quy, tóm tắt văn bản pháp quy và trả lời câu hỏi trong lĩnh vực pháp quy; (2) Có thể xây dựng các ứng dụng xử lý văn bản pháp quy nhằm giúp người dùng đọc, hiểu và truy xuất thông tin cần thiết từ văn bản pháp quy. Những việc này rất có ý nghĩa trong thực tiễn.

Xác nhận của người hướng dẫn khoa học **Nghiên cứu sinh**

GS.TS. Từ Minh Phương **Nguyễn Thị Thanh Thủy**

PGS.TS. Ngô Xuân Bách