THÔNG TIN LUẬN ÁN

Tên luận án:

MÔ HÌNH TÌM KIẾM VĂN BẢN TIẾNG VIỆT DỰA TRÊN NGỮ NGHĨA

Chuyên ngành:	Khoa học máy tính
Mã số:	62.48.01.01
Họ tên NCS:	Đỗ Thị Thanh Tuyền
Hướng dẫn khoa học:	PGS. TS. Nguyễn Tuấn Đăng
	PGS. TS. Vũ Đức Lung
Cơ sở đào tạo:	Trường ĐH CNTT – ĐHQG-HCM

1. TÓM TẮT

Truy hồi văn bản theo ngữ nghĩa là một trong những hướng nghiên cứu quan trọng nhằm khai thác tài liệu văn bản trong điều kiện khối lượng văn bản ngày càng lớn. Bài toán truy hồi văn bản theo ngữ nghĩa có thể được giải quyết theo các hướng tiếp cận khác nhau như phân tích ngữ nghĩa tiềm ẩn, mô hình chủ đề, ngữ nghĩa phân bố với vector ngữ nghĩa được xác định từ mạng nơron học sâu và phân tích ngữ nghĩa câu theo ngữ nghĩa học tính toán. Luận án giải quyết bài toán truy hồi văn bản theo hướng tiếp cận phân tích ngữ nghĩa câu với mục tiêu so khớp văn bản và cụm từ truy vấn ở mức ngữ nghĩa thông qua biểu diễn ngữ nghĩa của chúng chứ không phải ở mức từ vựng và nghĩa từ vựng.

Để hoàn thành mục tiêu, cách tiếp cận của luận án là đề xuất một ngôn ngữ hình thức để biểu diễn ngữ nghĩa của cụm từ dựa trên dạng biểu diễn Universal Dependency, sử dụng một cơ sở tri thức ngữ nghĩa từ vựng tiếng Việt (VLO – Vietnamese Lexicon Ontology) để biểu diễn nghĩa từ vựng và kiểm tra các ràng buộc ngữ nghĩa, nghiên cứu phương pháp gán nhãn nghĩa trong VLO cho từ vựng của cụm từ và nghiên cứu một phương pháp biến đổi cây cú pháp phụ thuộc thành đồ thị phụ thuộc theo dạng biểu diễn Universal Dependency dựa trên nghĩa từ vựng và ràng buộc ngữ nghĩa. Kế đến, nghiên cứu một mô hình truy hồi văn bản phù hợp với biểu diễn ngữ nghĩa của văn bản sử dụng hai hệ thống chỉ mục (lớp nghĩa SCI và quan hệ nghĩa SRI) để quá trình tính toán độ liên quan theo độ đo được xây dựng dựa trên độ đo khoảng cách Jaccard-Tanimoto. Kết quả truy hồi văn bản theo cách

tiếp cận của luận án cho kết quả tốt hơn mô hình xác suất với công thức xếp hạng BM25 cho cả hai trường hợp mở rộng và không mở rộng truy vấn, mô hình vector với công thức xếp hạng TF.IDF, mô hình sử dụng vector ngữ nghĩa và mô hình chủ đề với Latent Dirichlet Allocation (LDA).

Các kết quả nghiên cứu của luận án đều được đăng trong các kỷ yếu hội nghị quốc tế và tạp chí quốc tế được lập chỉ mục bởi các tổ chức có uy tín như ACM Digital Library, IEEE, DBPL, v.v.

2. CÁC ĐÓNG GÓP CHÍNH CỦA LUẬN ÁN

Luận án có các đóng góp chính như sau:

1) Đề xuất ngôn ngữ hình thức để biểu diễn ngữ nghĩa cụm từ theo dạng biểu diễn Universal Dependency.
2) Nghiên cứu phương pháp phân tích ngữ nghĩa cụm từ dựa trên luật để biến đổi kết quả phân tích cú pháp phụ thuộc từ mô hình dùng các đặc trưng được rút trích từ mạng nơron Bidirectional LSTM.
3) Xây dựng VLO để cung cấp các nhãn nghĩa và ràng buộc ngữ nghĩa làm cơ sở cho việc biến đổi các quan hệ phụ thuộc.
4) Đề xuất mô hình truy hồi văn bản dựa trên ngữ nghĩa với hai hệ thống chỉ mục (lớp nghĩa SCI và quan hệ nghĩa SRI), hàm xếp hạng được phát triển từ độ đo khoảng cách Jaccard-Tanimoto có tác dụng xếp hạng tốt hơn hàm xếp hạng BM25.

3. NHỮNG VẤN ĐỀ CẦN TIẾP TỤC NGHIÊN CỨU

Luận án đã nghiên cứu mô hình truy hồi văn bản có kết quả xếp hạng tốt và phương pháp phân tích ngữ nghĩa cụm từ dựa trên kết quả phân tích cú pháp phụ thuộc phù hợp với yêu cầu. Tuy nhiên, để kết quả tốt hơn nữa cần phải:

- Bổ sung thêm các lớp nghĩa và ràng buộc ngữ nghĩa cho VLO và phát triển dependency graphbank cho tiếng Việt để phát triển mô hình phân tích phụ thuộc trong câu tiếng Việt tốt hơn.
- Nghiên cứu xây dựng VLO tự động để giảm chi phí.
- Nghiên cứu huấn luyện mô hình phân tích phụ thuộc trực tiếp trên dependency graphbank tiếng Việt.

- Nghiên cứu áp dụng các kết quả từ hướng tiếp cận mạng nơron học sâu để tăng hiệu quả của các thành phần trong mô hình.

CÁN BỘ HƯỚNG DẪN 1	CÁN BỘ HƯỚNG DẪN 2	NGHIÊN CỨU SINH
PGS. TS. Nguyễn Tuấn Đăng	PGS. TS. Vũ Đức Lung	Đỗ Thị Thanh Tuyền