TÓM TẮT KHOÁ LUẬN TỐT NGHIỆP NĂM HỌC 2024

XÂY DỰNG DỮ LIỆU VÀ TRÍCH XUẤT QUAN HỆ TRONG VĂN BẢN Y TẾ TIẾNG VIỆT

Họ và tên sinh viên: Nguyễn Chí Dũng

Ngày sinh: 05/10/2002 Mã SV: 20002040

Khóa: QH.2020 Khoa: Toán-Cơ-Tin học

Họ và tên cán bộ hướng dẫn: Nguyễn Thị Minh Huyền, Nguyễn Hải Vinh

Tóm tắt nội dung khoá luận tốt nghiệp:

Với mong muốn thu hẹp khoảng cách giữa nguồn dữ liệu trích xuất quan hệ thực thể y tế Tiếng Việt và Tiếng Anh, góp phần thúc đẩy ứng dụng công nghệ thông tin vào lĩnh vực y tế tại Việt Nam, mang lại lợi ích thiết thực cho cộng đồng, khóa luận này tập trung vào việc xây dựng dữ liệu trích xuất quan hệ thực thể y tế Tiếng Việt và áp dụng mô hình PhoBERT phân loại các quan hệ được định nghĩa trong bộ dữ liệu được xây dựng. Khóa luận giới thiệu về hệ thống y tế hợp nhất UMLS và các khái niệm trong nhận diện thực thể và trích xuất quan hệ, sau sau đó đi sâu phân tích mô hình PhoBERT, so sánh hiệu suất của nó với mô hình XLM-R và áp dụng tren bộ dữ liệu VSLP 2020 Relation Extraction. Tiếp theo, khóa luận trình bày quy trình xây dựng bộ nhãn thực thể và quan hệ y tế Tiếng Việt dựa trên quy chuẩn quốc tế, thu thập dữ liệu từ Wikipedia và gán nhãn bằng công cụ INCEpTION. Kết quả thực nghiệm cho thấy mô hình PhoBERT đạt hiệu suất cao trên bộ dữ liệu mẫu, với Macro-Average F-score là 0.67 và Micro-Average F-score là 0.85 trên tập huấn luyện. Trên tập phát triển, kết quả thu được Macro-Average F-score là 0.65 và Micro-Average F-score là 0.78.

Từ khoá: Xây dựng dữ liệu, trích xuất quan hệ, RE