

Nhận diện thực thể trong văn bản y học tiếng Việt

Nhận diện thực thể (NER) là một thành phần quan trọng trong việc xử lý văn bản y tế, đặc biệt là trong việc trích xuất thông tin lâm sàng từ các báo cáo bệnh lý. Trong bài báo này, chúng tôi trình bày một hệ thống NER được thiết kế riêng cho văn bản y tế tiếng Việt, tập trung vào lĩnh vực bệnh học. Hệ thống của chúng tôi xác định và phân loại các thực thể y học chính như “person”, “disease”, “symptom”, “part” and “treatment”. Chúng tôi sử dụng một mô hình tiền huấn luyện dựa trên Electra được tối ưu hóa cho tiếng Việt, đạt được điểm F1 là 90% trên tập dữ liệu bệnh lý chuẩn. Kết quả cho thấy tính hiệu quả của phương pháp của chúng tôi trong việc xử lý các thuật ngữ y tế phức tạp trong tiếng Việt, cung cấp một giải pháp mạnh mẽ cho việc trích xuất thông tin lâm sàng. Nghiên cứu trong tương lai nhằm mở rộng khả năng của mô hình để bao phủ thêm các thực thể y tế và cải thiện hiệu suất trên các thuật ngữ hiếm hoặc mơ hồ.

Từ khóa

NER Vietnamese model, Electra, thực thể y học, rút trích thông tin văn bản, bệnh học

Thông tin các tác giả

1/ Nguyễn Thị Xuân Hiền, học viên bậc thạc sĩ trường Đại Học Khoa Học Tự Nhiên
Email: ntchien98@gmail.com
2/ Phạm Trần Nhật Minh, học viên bậc thạc sĩ trường Đại Học Khoa Học Tự Nhiên
Email: minhpham@gmail.com

Auteurs: M. PHẠM, Minh (Trường Đại Học Khoa Học Tự Nhiên); HIỀN, Nguyễn Thị Xuân (Trường Đại học Khoa Học Tự Nhiên)

Classification de Session: Tiểu ban 1

Classification de thématique: Tiểu ban khoa học tự nhiên, công nghệ