

XÂY DỰNG MÔ HÌNH ĐÁNH GIÁ NGỮ LIỆU

Nhu cầu phát triển các ứng dụng xử lý ngôn ngữ tự nhiên tiếng Việt đặt ra yêu cầu cấp thiết phải có cơ sở dữ liệu ngữ liệu chất lượng cao, tuy nhiên hiện nay chưa có khung tham chiếu thống nhất cho việc đánh giá loại tài nguyên này. Nghiên cứu này đề xuất một mô hình đánh giá ngữ liệu tiếng Việt, kết hợp phương pháp định lượng và định tính, với sự hỗ trợ từ WordSketch để phân tích collocation. Mô hình được xây dựng cho thấy tính khả thi trong việc tích hợp các tiêu chí định lượng và định tính vào một quy trình thống nhất, có khả năng phản ánh một cách hệ thống chất lượng và hạn chế của ngữ liệu. Kết quả thực nghiệm bước đầu trên bộ dữ liệu 10.000 câu gán nhãn chứng minh mô hình có tiềm năng áp dụng cho tiếng Việt. Nghiên cứu này góp phần hình thành khung tham chiếu đánh giá, mở ra hướng cải thiện và phát triển ngữ liệu tiếng Việt phục vụ nghiên cứu và ứng dụng NLP trong tương lai. Bộ mã nguồn mở được công bố tại: https://github.com/vuthithi24/wordsketch_vietnamese.git.

Từ khóa

quality corpora, corpus evaluation, WordSketch, collocation extraction, Vietnamese NLP

Thông tin các tác giả

1/ Vũ Thi Thi: CN., đang công tác tại Trường Đại học Khoa học tự nhiên, ĐHQG-HCM, 227 Nguyễn Văn Cù, Phường Chợ Quán, TP. HCM, email: vtthi@hcmus.edu.vn

2/ Đinh Điền: PGS. TS., đang giảng dạy tại Trường Đại học Khoa học tự nhiên, ĐHQG-HCM, 227 Nguyễn Văn Cù, Phường Chợ Quán, TP. HCM, email: ddiem@fit.hcmus.edu.vn

Author: VU, Thi-Thi

Track Classification: Tiểu ban 1: Những tiến bộ và thành tựu mới trong lĩnh vực Ngôn ngữ học Tính toán