

GÁN NHÃN NGỮ NGHĨA TỰ ĐỘNG CHO TỪ ĐIỂN OALD ANH-ANH-VIỆT DỰA TRÊN MẠNG TỪ WORDNET

Nghiên cứu này đề xuất phương pháp gán nhãn ngữ nghĩa tự động cho các định nghĩa trong từ điển “Oxford Advanced Learner’s Dictionary with Vietnamese translation” (OALD, ấn bản lần thứ 8) dựa trên “synset_id” của WordNet. Mô hình ghép cặp các định nghĩa tương đồng, sử dụng phương pháp “sentence embedding” : “Word2Vec”, “GloVe”, “FastText” và “Sentence-BERT”. Các thử nghiệm được thực hiện trên hai tập dữ liệu chuẩn là “SICK” và “Quora Question Pairs” nhằm đánh giá độ chính xác và tìm ra phương pháp tính độ tương đồng tốt nhất. Kết quả cho thấy mô hình “SBERT” đạt độ chính xác cao nhất và được sử dụng trong quá trình gán nhãn ngữ nghĩa tự động. Việc liên kết giữa từ điển OALD với WordNet mở ra tiềm năng kết nối giữa những từ điển đa ngôn ngữ thông qua “synset_id”.

Từ khóa

OALD, WordNet, SBERT, Gán nhãn ngữ nghĩa tự động, Rút trích đặc trưng văn bản

Thông tin các tác giả

- 1/ Phan Văn Bá Hải: Sinh viên trường Đại học Khoa học Tự nhiên - ĐHQG HCM, 227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM, email: phanvanbahai@gmail.com
- 2/ Đỗ Quốc Trí: Sinh viên trường Đại học Khoa học Tự nhiên - ĐHQG HCM, 227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM, email: doquoctri2003@gmail.com
- 3/ Phan Thị Mỹ Trang: ThS. NCS. trường Đại học Khoa học xã hội và Nhân văn - ĐHQG HCM, 10-12 Đinh Tiên Hoàng, Bến Nghé, Quận 1, TP.HCM; đang giảng dạy tại trường Đại học Công nghệ Sài Gòn, 180 Cao Lỗ, Phường 4, Quận 8, TP.HCM, email: mytrang779@gmail.com
- 4/ Đinh Điền: PGS. TS. đang công tác tại trường Đại học Khoa học Tự nhiên - ĐHQG HCM, 227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM, email: ddien@fit.hcmus.edu.vn

Auteurs: M. HẢI, Phan Văn Bá (Trường Đại học Khoa học Tự nhiên - ĐHQG HCM); ĐỖ QUỐC, Trí (Trường Đại học Khoa học Tự nhiên TP.HCM)

Co-auteurs: TRANG, Phan Thị Mỹ (Trường Đại học Khoa học xã hội và Nhân văn - ĐHQG TP.HCM; Trường Đại học Công nghệ Sài Gòn); ĐIỀN, Đinh (Trường Đại học Khoa học Tự nhiên, ĐHQG TP.HCM)

Classification de Session: Tiểu ban 1

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