

ỨNG DỤNG MÔ HÌNH DỊCH MÁY THỐNG KÊ TRONG BÀI TOÁN DỊCH VĂN BẢN CHỮ QUỐC NGŨ VIỆT NAM SANG CHỮ HÁN NÔM

Nếu chữ Quốc Ngữ là chữ viết chính thức hiện nay và được xem như một minh chứng cho độc lập và chủ quyền của quốc gia này, thì chữ Hán Nôm lại được xem như chữ viết đầu tiên của riêng người Việt sau gần 1000 năm Bắc thuộc và là nền tảng mở đường cho chữ Quốc Ngữ sau này. Chúng tôi mong muốn tạo điều kiện cho người Việt Nam hiện đại tiếp cận đến giá trị tinh thần xa xưa ấy dễ dàng hơn thông qua việc cung cấp hướng phát triển cho bài toán dịch ngược từ chữ Quốc Ngữ sang chữ Hán Nôm. Bài báo này ứng dụng và khảo sát mô hình dịch máy thống kê Moses trong bài toán dịch ngược. Ngữ liệu huấn luyện bao gồm 23K cặp câu Quốc Ngữ - Hán Nôm song song được thu thập từ các văn bản truyện, thơ, văn, bản tưng... trên trang web nomfoundatinon.org. Kết quả cho thấy mô hình baseline của Moses đem lại được kết quả ổn định hơn so với một mô hình dịch ngược state-of-the-art (SOTA) khác hiện nay.

Thông tin các tác giả

1/Ngô Thái Phụng: Sinh viên đang học tại Trường Đại học Khoa học tự nhiên, ĐHQG TP. HCM, 227 Nguyễn Văn Cù, Phường 4, Quận 5, TP. HCM, email: ntphung21@apcs.fitus.edu.vn
2/Đặng Minh Nhật: Sinh viên đang học tại Trường Đại học Khoa học tự nhiên, ĐHQG TP. HCM, 227 Nguyễn Văn Cù, Phường 4, Quận 5, TP. HCM, email: dmnhut22@apcs.edu.vn

Từ khóa

dịch ngược, chữ Quốc Ngữ, chữ Hán Nôm, dịch máy thống kê, Moses SMT, xử lý ngôn ngữ tự nhiên

Auteur: PHỤNG, Ngô Thái (Trường đại học Khoa học tự nhiên, ĐHQG TP.HCM)

Co-auteur: M. ĐẶNG, Minh Nhật (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