VII-O-11

KHẢO SÁT YẾU TỐ RANH GIỚI TỪ TRONG DỊCH THỐNG KÊ HOA – VIỆT

Trần Thanh Phước[(1)], Đinh Điền[(2)]

(1) Đại Học Công Nghiệp Thực PhẩmTp.HCM

(2) Khoa Công nghệ Thông tin, Trường ĐH KHTN, ĐHQG-HCM

Tóm tắt

Trong các ngôn ngữ đơn lập như tiếng Hoa và tiếng Việt, các từ không được phân biệt với nhau bởi khoảng trắng, một từ có thể bao gồm một hoặc nhiều từ chính tả. Việc có nên phân đoạn từ hay không trước khi cho qua hệ thống huấn luyện và dịch là vấn đề cần được xem xét. Trong bài báo này, chúng tôi sẽ tiến hành khảo sát ảnh hưởng của yếu tố ranh giới từ đến kết quả dịch thống kê Hoa-Việt. Kết quả thực nghiệm của bài báo sẽ làm cơ sở cho các hướng nghiên cứu cải tiến phân đoạn từ tiếp theo nhằm tăng hiệu suất dịch.

Chúng tôi đã khảo sát trên hai trường hợp sau: không phân đoạn từ và phân đoạn từ trên kho ngữ liệu 8.000 và 12.000 cặp câu. Dựa trên kết quả thực nghiệm, chúng tôi nhận thấy rằng: ngữ liệu chưa phân đoạn từ hoặc được phân đoạn từ đều có những ưu và khuyết điểm riêng. Một hướng cải tiến mà bài báo đề xuất là tích hợp các ưu điểm của hai phương pháp này vào hệ thống dịch máy.

Email liên hệ:***phuoctt@cntp.edu.vn***