Phục Hồi Thanh Điệu Tự Động Cho Tiếng Việt Sử Dụng Mạng Neuron và Mô Hình Ngôn Ngữ

***Đào Tuấn An*[1], *Trương Hưng Thịnh*[1], *Phan Thị Phương Uyên*[1]**
[1]Khoa Công nghệ thông tin,
Trường Đại học Khoa học Tự Nhiên, ĐHQG-HCM
dtan@apcs.vn, ththinh@apcs.vn

Tóm tắt

Do sai sót khi nhập liệu hoặc để thuận tiện, một lượng lớn tài liệu tiếng Việt không dấu trên mạng Internet đã bị mất đi ý nghĩa ban đầu của nó. Mặc dù tiếng Việt sử dụng bảng chữ cái Latin như một số ngôn ngữ phổ biến như tiếng Anh, tiếng Pháp, nhưng nó cũng cấu thành từ thanh điệu. Một từ với những thanh điệu khác nhau sẽ mang những ý nghĩa rất khác nhau. Vì vậy, việc phục hồi thanh điệu tự động cho tiếng Việt có rất nhiều ứng dụng thực tiễn như kiểm tra lỗi chính tả, truy vấn thông tin. Trong nghiên cứu này, chúng tôi trình bày một phương pháp học sâu kết hợp mô hình sequence-to-sequence và mô hình ngôn ngữ để phục hồi những thanh điệu đã mất trong văn bản tiếng Việt. Kết quả thực nghiệm thu được cho thấy sự cải tiến so với các phương pháp trước đó. Phương pháp đề xuất có thể được mở rộng sang các ngôn ngữ có thanh điệu khác như tiếng Pháp, tiếng Ro-ma-ni-a, tiếng Đức và các tiếng khác.

Từ khóa: phục hồi thanh điệu, mô hình ngôn ngữ, học sâu, sequence-to-sequence.

Automatic Vietnamese Diacritic Restoration Using Neural Network and Language Model

***Dao Tuan An*[1], *Truong Hung Thinh*[1], *Phan Thi Phuong Uyen*[1]**
[1]Faculty of Information Technology, University of Science, VNU-HCM
dtan@apcs.vn, ththinh@apcs.vn

Abstract

Due to errors when inputting text or convenience reasons, there is a large amount of accent-less Vietnamese text on the Internet which lose its original meaning. Although the Vietnamese language uses the Latin alphabet like some popular languages such as English, French, it also contains diacritics. Different diacritics apply to the same word lead to a significant change in term of its semantic meaning. Therefore, automatic restoration of diacritic has many practical applications such as Spelling Checking, Information Retrieval. We present a deep learning approach which combines sequence-to-sequence model with a language model to recover missing diacritics in Vietnamese text. Experiments show an improvement over existing results. The proposed method could be extended to other languages which contain accented text such as French, Romanian, German, and so on.

Key words: diacritic restoration, language model, deep learning, sequence-to-sequence.