

Tự động trích chọn thực thể tên người trong văn bản tiếng Việt

Học viên: Đào Minh Quyền
Đơn vị công tác: Công ty TNHH Hoya Glass Disk Việt Nam
Email: quyendm82@gmail.com

GVHD: TS. Nguyễn Trí Thành
Đơn vị công tác: Trường Đại học Công nghệ, ĐHQG – Hà Nội
Email: ntthanh@vnu.edu.vn

Từ khóa: Trích chọn thông tin, nhận dạng thực thể, xử lý ngôn ngữ tự nhiên

1. GIỚI THIỆU BÀI TOÁN

Hệ thống trích chọn tên người là bài toán cơ bản và có ý nghĩa trong việc xây dựng các hệ thống xử lý ngôn ngữ tự nhiên như xây dựng các ontology làm tiền đề cho các Semantic web, ứng dụng trong hệ thống dịch tự động hay tóm tắt văn bản, ... Luận văn tập trung vào việc xây dựng hệ thống tự động trích chọn tên người trong văn bản Tiếng Việt.

2. NỘI DUNG LUẬN VĂN

2.1 Xây dựng hệ thống tự động trích chọn

Bao gồm 3 bước:

Bước 1: Đánh giá toàn cục

Dựa trên từ điển họ trong tiếng Việt & đặc điểm ngôn ngữ, thực hiện trích chọn tên người đầy đủ. Lưu danh sách tên người và danh sách từ đứng trước tên người (từ tổ - prefix) ra 2 file

Bước 2: Gán nhãn

Gán nhãn cho từ trong văn bản đầu vào. Nếu là tên người gán nhãn P, các trường hợp khác gán nhãn O

Bước 3: Kiểm tra prefix để làm tăng độ chính xác

Dựa vào prefix để kiểm tra lại các từ đã được gán nhãn P.

2.2 Cài đặt hệ thống theo mô hình tự động trích chọn

Cài đặt thực nghiệm hệ thống. Trích chọn với độ chính xác gần 92%

2.3 Sử dụng kết quả của hệ thống tự động trích chọn làm dữ liệu huấn luyện cho CRF++

Mục đích: Đánh giá độ chính xác của kết quả trích chọn của hệ thống tự động trích chọn.

Cài đặt: Bổ sung thêm 1 số thuộc tính nổi bật của thực thể tên người trong văn bản tiếng Việt vào quá trình trích chọn của CRF ++. Luận văn đã cài đặt thành công mô hình với độ chính xác khi trích chọn là hơn 92%

3. KẾT LUẬN

Từ việc nghiên cứu bài toán trích chọn thực thể, luận văn đã đề xuất giải pháp và thực hiện cài đặt thử nghiệm hệ thống tự động trích chọn tên người trong văn bản tiếng Việt. Qua thực nghiệm cho thấy, mô hình là khả thi và có thể áp dụng được.

Về mặt nội dung, khóa luận đã đạt được những kết quả sau:

- Giới thiệu bài toán trích chọn thực thể tên người và một số nghiên cứu liên quan trong tiếng Việt
- Đề xuất mô hình giải quyết bài toán, tiến hành cài đặt thử nghiệm bước đầu cho được kết quả khả quan. Trích chọn với độ chính xác hơn 92% và không cần có dữ liệu huấn luyện ban đầu.

Tuy nhiên, do hạn chế về mặt thời gian và kiến thức, luận văn còn một số hạn chế sau:

- Mới chỉ dừng lại ở việc trích chọn tên người, vẫn còn một số trường hợp trích chọn sai / thiếu.
- Chưa tích hợp thêm được các thông tin liên quan về người đó như chức vụ, quê quán, ...

Về định hướng nghiên cứu, trong thời gian tới luận văn sẽ phát triển theo hướng như:

- Trích chọn thêm các đặc trưng liên quan hay phân lớp của người đó trong văn bản
- Giải quyết tốt hơn các trường hợp nhập nhằng để làm tăng độ chính xác của giải thuật.