ỨNG DỤNG CRF NHẬN DẠNG THỰC THỂ ĐỊNH DANH TRONG VĂN BẢN TIẾNG VIỆT

APPLICATION OF CRF FOR NAMED ENTITY RECOGNITION IN VIETNAMESE DOCUMENTS

Tác giả: *Võ Trung Hùng*, Lâm Tùng Giang, Trần Thị Liên*

Tóm tắt bằng tiếng Việt:

Nhận dạng các thực thể định danh là một lĩnh vực đang nhận được sự quan tâm rộng rãi của các nhà nghiên cứu. Đã có nhiều kết quả trong lĩnh vực này trong một số ngôn ngữ như Anh, Pháp, Trung Quốc,… nhưng với Tiếng Việt thì còn hạn chế. Mục đích nghiên cứu để xây dựng một hệ thống nhận dạng thực thể cho phép nhận dạng các thực thể có tên trong văn bản Tiếng Việt như tên người, địa điểm, tổ chức, thời gian,… được phát triển dựa trên công cụ CRF++. Nhiệm vụ chính là xây dựng một tập dữ liệu tốt, đầy đủ, chính xác nhằm hỗ trợ cho việc nhận dạng thực thể và xây dựng một hệ thống huấn luyện, kiểm thử và ứng dụng. Hệ thống nhận dạng thực thể đã thực nghiệm trên 300 bài báo với nhiều lĩnh vực khác nhau và hoạt động có tính khả thi cao với độ đo F1 trung bình qua 10 lần thực nghiệm đạt 84,8%.

Từ khóa: Nhận dạng thực thể có tên; mô hình CRF; công cụ CRF++; tên các thực thể trong Tiếng Việt; hệ thống nhận dạng thực thể

Tóm tắt bằng tiếng Anh:

Named Entity Recognition, a subfield of Information Extration, is getting wide attention. Researches with English, French or Chinese produce good results but there are not many works with Vietnamese. The purpose of this study is building a named entity recognition system allowing identification of named entities such as person name, location, organization, time in Vietnamese texts by using CRF + + tool. The main task is creating tools and training data for building a named entity recognition model to facilitate the identification of the entities in vienamese documents. The Entity Recognition system was evaluated 10 times on over 300 papers and gives the average F1 measure of 84,8%.

Key words: Name Entity Recognition; CRF model; CRF++ Toolkit; Name entity in Vietnamese; entity recognition system