GIẢI PHÁP TRÍCH RÚT VÀ PHÂN LOẠI CÁC THỰC THỂ DANH TỪ RIÊNG CHO KHO NGỮ LIỆU PHỤC VỤ XỬ LÝ NGÔN NGỮ TỰ NHIÊN

EXTRACTION AND CLASSICATION OF NAMED ENTITIES FROM CORPORA IN NATURAL LANGUAGE PROCESSING

Tác giả: *Đặng Đai Tho**, *Huỳnh Công Pháp*, *Doãn Hằng Hiệu*

Tóm tắt bằng tiếng Việt:

Trích rút và phân loại thực thể danh từ riêng cho các kho ngữ liệu phục vụ xử lý ngôn ngữ tự nhiên là bước quan trọng và là tiền đề cho việc mở rộng cũng như xây dựng các kho ngữ liệu theo hướng ngữ nghĩa. Việc nghiên cứu trích rút và phân loại thông tin đã được thực hiện với nhiều ngôn ngữ. Tuy nhiên, đến nay vẫn chưa có công trình nào nghiên cứu trích rút và phân loại thực thể danh từ riêng trên các kho ngữ liệu phục vụ xử lý ngôn ngữ tự nhiên. Hơn nữa, các phương pháp trích rút và phân loại thông tin đã sử dụng như nêu ở trên đều có những nhược điểm riêng của nó.
Trong bài báo này, chúng tôi đề xuất giải pháp kết hợp thuật toán so khớp tối đa với phân tích quan hệ ngữ cảnh giữa các thành tố trong văn bản để trích rút và phân loại các thực thể danh từ riêng cho kho ngữ liệu phục vụ xử lý ngôn ngữ tự nhiên. Giải pháp này bước đầu đã mang lại kết quả rất đáng khích lệ.

Từ khóa: Trích rút thông tin; phân loại thông tin; kho ngữ liệu; trích rút tên riêng; phân loại tên riêng.

Tóm tắt bằng tiếng Anh:

Extraction and classification of named entities from corpora in Natural Language Processing (NLP) is an important initial step for extending and building semantic oriented corpora. Though there have been many researches on the extraction and classification of information from internet resources in foreign languages, no research has dealt with corpora in NLP. Moreover, information extraction and classification methods currently used such as rule based, machine learning or hidden Markov have shown some drawbacks. In this paper, we propose a solution combining Maximum Matching method and contextual relation analysis of entities in the text for extracting and classifying named entities from corpora in NLP. In the first stage of our research, this proposed solution has given positive results.

Key words: Information extraction; information classification; named entity extraction; named entity classification; corpora;