HỌC DỰA TRÊN SỰ BIẾN ĐỔI TRẠNG THÁI VÀ CÁC ỨNG DỤNG TRONG LĨNH VỰC XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Nguyễn Tấn Đạt, Lê Minh Sơn
Khoa CNTT, Trường Đại học Khoa học Tự Nhiên - ĐHQG tp.HCM

Tóm tắt:

Học dựa trên sự biến đổi trạng thái (TBL) là một phương pháp máy học dạng thống kê được phát triển trong những năm gần đây bởi tiến sĩ Eric Brill. Phương pháp này đã và đang đạt được nhiều thành công trong các ứng dụng của lĩnh vực xử lý ngôn ngữ tự nhiên.

Phương pháp học TBL sử dụng một ngữ liệu lớn đã được xác định thông tin đầy đủ bằng tay và một tập hợp mẫu luật chuyển đổi trạng thái, từ đó học và sinh ra bộ luật nhằm giải quyết vấn đề đặt ra. Phương pháp học này đã đạt được những kết quả đáng kinh ngạc khi giải các bài toán trong lĩnh vực xử lý ngôn ngữ như: đánh nhãn từ loại, xác định vị trí bổ nghĩa của ngữ giới từ, bắt lỗi chính tả, khữ nhập nhằng ngữ nghĩa, .

Ưu điểm của phương pháp này là tạo ra một bộ luật dễ hiểu, thời gian học cũng như thực hiện là tuyến tính, giải quyết tốt những tình huống có xác suất thấp hoặc chưa gặp trong quá trình học, . Khuyết điểm quan trọng nhất của TBL là thời gian huấn luyện rất lâu, bên cạnh đó số lượng luật rút ra đôi khi nhiều một cách không cần thiết. Hiện nay, các ứng dụng của TBL vẫn đang được phát triển cũng phương pháp học này cũng nhận được nhiều nghiên cứu cải tiến đáng kể như: Fast TBL, μ TBL, Lazy TBL, .

TRANSFORMATION-BASED LEARNING AND ITS APPLICATION IN NATURAL LANGUAGE PROCESSING

Nguyen Tan Dat, Le Minh Son
Department of Information Technology, University of Natural Sciences - VNU.HCM

Abstract:

Transformation-based learning (TBL) is a statistic machine leaning method developed in recent years by Dr. Eric Brill. This method has been obtaining many successes in applications in natural language processing.

TBL uses a large hand-annotated corpus and a set of transformation rule templates, from which it learns and generates a set of transformation rules to solve a specific problem. This learning method has acquired amazing results in solving problems in natural language processing, such as: POS tagger, PP attachment, spelling checking, WSD .

The advantages of this method are that it generates a set of easy-to-understand rules, its training and performing time is linear, and it implements in situations having statistic value low or zero quite well, The most important disvantages of TBL are its very long training time, and its large set of generated rules, some in which may be unnecessary. At the moment, applications of TBL are being developed, and there are also many developments and modifications from TBL, e.g.: Fast TBL, μ TBL, Lazy TBL .