Báo cáo Project 2

1. Đặt vấn đề

* Tổng quan
* Gán nhãn từ loại là gì?

Gán nhãn từ loại là việc xác định các chức năng ngữ pháp của từ trong câu. Đây là bước cơ bản trước khi phân tích sâu văn phạm hay các vấn đề xử lý ngôn ngữ phức tạp khác.

Thông thường, một từ có thể có nhiều chức năng ngữ pháp, ví dụ : “the book is interesting” và “We book a room” .

* Các phương pháp có thể sử dụng để gán nhãn từ loại.

Một số hướng tiệp cận chính trong gán nhãn từ loại tiếng Anh[Đinh Điền] bao gồm: gắn nhãn dựa trên mô hình Markov ẩn (MHH); các mô hình dựa trên bộ nhớ(Daelemans, 1996); mô hình dựa trên luật (Transformation Base Learning, Brill, 1995); Maximum Entropy; cây quyết định (Schmid, 1994a); mạng nơ-ron(Schmid, 1994b).

1. Nội dung nghiên cứu

* Mô hình Markov ẩn
* Áp dụng HMM và thuật toán Viterbi vào gán nhãn từ loại
* Dữ liệu
* Thuật toán Viterbi
* Thực nghiệm
* Phương pháp thực nghiệm
* Kết quả thực nghiệm
* Nhận xét.

Có 6 trong tổng số 45 nhãn có giá trị F1 nhỏ hơn 70. Trong đó LS là thấp nhất (2.86)

* Phân tích nhãn NNPS (60.16% phân loại đúng) có 22.15% là phân loại nhầm thành NNP, 8.15% NNS và còn lại là các nhãn khác.
* Cả NNP và NNPS đều có khả năng đứng sau nhiều loại nhãn giống nhau.
* Có 34.68% từ có nhãn là NNPS có khả năng là nhãn NNP 17.99% từ có nhãn là NNPS có khả năng là nhãn NNS. 11.20% số từ có nhãn là NNPS có xác suất là nhãn NNP và NNPS tương đương nhau.
* Chương trình demo.
* Giao diện
* Các lớp quan trọng

1. Tổng kết.