1. Giới thiệu bài toán POS Tagging

Gán nhãn từ loại (Part of Speech Tagging - POS Tagging) là một trong những bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên. Nó xác định và gán nhãn từ loại, chẳng hạn như: Sông - danh từ(NN), đi - động từ(V), đẹp -  tính từ (JJ), Họ - đại từ(PRP), và - giới từ(IN) cho từng từ trong câu. Giúp máy tính có thể hiểu rõ hơn về cấu trúc ngữ pháp và ngữ nghĩa của câu. Việc gán nhãn từ loại chính xác đóng vai trò rất quan trọng, vì nó ảnh hưởng trực tiếp đến hiệu quả của nhiều ứng dụng như dịch máy, phân tích cảm xúc hay tóm tắt văn bản.

2.  Mô hình Markov ẩn

Mô hình Markov ẩn (Hidden Markov Model - HMM) là một mô hình thống kê mô tả hệ thống có các trạng thái ẩn mà ta không thể quan sát trực tiếp. Thay vào đó, ta chỉ thấy các đầu ra (quan sát) được sinh ra từ những trạng thái này. Nhiệm vụ của HMM là suy ra các trạng thái ẩn dựa trên chuỗi quan sát.

Khác với mô hình Markov thông thường – nơi các trạng thái có thể được quan sát trực tiếp – HMM bổ sung thêm lớp “ẩn” và xác suất phát ra đầu ra từ mỗi trạng thái. Nhờ đó, HMM được ứng dụng rộng rãi trong các bài toán như nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên và nhận dạng mẫu.

3. Giải thuật Viterbi

Giải thuật Viterbi là một thuật toán lập trình động được sử dụng để tìm chuỗi trạng thái ẩn có khả năng xảy ra nhất dựa trên một chuỗi sự kiện quan sát được. Nó tìm kiếm chuỗi nhãn tối ưu bằng cách xem xét tất cả các khả năng và chọn ra chuỗi có xác suất cao nhất. Ứng dụng của thuật toán này rất đa dạng, từ giải mã tín hiệu trong viễn thông đến nhận dạng giọng nói và xử lý ngôn ngữ tự nhiên.

4. Ứng dụng HMM trong gán nhãn từ loại.

Trong bài toán gán nhãn từ loại, HMM được sử dụng để xác định chuỗi nhãn từ loại (các trạng thái ẩn) có khả năng xảy ra nhất cho một chuỗi từ quan sát được trong câu.

Thuật toán Viterbi thường được sử dụng để tìm ra chuỗi nhãn từ loại tối ưu nhất cho một câu, bằng cách sử dụng các xác suất này và quy hoạch động để chọn ra chuỗi có xác suất tích lũy cao nhất.

Kết luận: HMM là một trong những mô hình cơ bản và kinh điển trong xử lý ngôn ngữ tự nhiên.

Nó cung cấp cách tiếp cận thống kê rõ ràng cho bài toán gán nhãn từ loại, và giải thuật Viterbi giúp tìm chuỗi nhãn tối ưu một cách hiệu quả.

Mặc dù đã được thay thế bởi các mô hình học sâu hiện đại, HMM vẫn là nền tảng quan trọng giúp hiểu cơ chế xác suất trong NLP.