I. Lý thuyết

1.1 Mô hình Markov ẩn (Hidden Markov Model - HMM)

Mô hình Markov ẩn (HMM) là một mô hình xác suất mô tả một quá trình có các trạng thái ẩn (không quan sát được trực tiếp) và các quan sát (dữ liệu đầu ra) phụ thuộc vào các trạng thái đó. Trong bài toán gán nhãn từ loại (POS tagging), các 'trạng thái ẩn' là các nhãn từ loại (danh từ, động từ, giới từ, v.v.), còn 'quan sát' là các từ xuất hiện trong câu. HMM được đặc trưng bởi ba thành phần chính:

- Tập trạng thái (Tags): ví dụ {N, V, P, E,...}
- Ma trận xác suất chuyển trạng thái (Transition): P(tag_i | tag_{i-1})
- Ma trận xác suất phát xạ (Emission): P(word | tag)

Trong POS tagging, nhiệm vụ của HMM là tìm chuỗi nhãn (tags) có xác suất cao nhất cho chuỗi từ đã cho.

1.2 Ứng dụng của HMM trong gán nhãn từ loại

HMM được sử dụng để gán nhãn từ loại cho câu bằng cách xem quá trình sinh câu là một chuỗi các trạng thái (từ loại) sinh ra các từ. Mục tiêu là tìm chuỗi nhãn T = (t1, t2, ..., tn) sao cho P(T|W) lớn nhất với W là chuỗi các từ quan sát.

Theo định lý Bayes, ta có:

$$P(T|W) \propto P(W|T) * P(T)$$

Trong đó:

- P(T): Xác suất chuỗi nhãn (dựa trên ma trận chuyển trạng thái)
- P(W|T): Xác suất sinh các từ từ chuỗi nhãn (dựa trên ma trận phát xạ)

1.3 Giải thuật Viterbi

Giải thuật Viterbi được sử dụng để tìm chuỗi nhãn T tối ưu có xác suất lớn nhất. Thuật toán hoạt động dựa trên nguyên tắc quy hoạch động, lưu lại xác suất cao nhất tại mỗi bước và truy vết ngược để tìm ra chuỗi nhãn tối ưu.

Các bước chính:

1. Khởi tạo: Tính xác suất cho từng nhãn tại vị trí đầu tiên.
2. Đệ quy: Với mỗi từ tiếp theo, chọn nhãn có xác suất cao nhất dựa trên nhãn trước đó.
3. Kết thúc: Chọn chuỗi nhãn có xác suất lớn nhất tại từ cuối cùng.
4. Truy vết ngược để lấy chuỗi nhãn tối ưu.

II. Đánh giá mô hình

2.1 Kết quả thực nghiệm

Mô hình được huấn luyện trên tập dữ liệu mẫu gồm 5 câu tiếng Việt đơn giản. Sau khi huấn luyện, mô hình được kiểm tra với 3 câu thử nghiệm:

1. Tôi đang học tại trường đại_học → Dự đoán: 100% chính xác
2. Bạn ăn cơm → Dự đoán: 100% chính xác
3. Tôi học đại_học → Dự đoán: 100% chính xác

Độ chính xác tổng thể: 12/12 từ đúng (Accuracy = 1.0000)

2.2 Phân tích kết quả

Kết quả cho thấy mô hình đạt độ chính xác tuyệt đối trên tập thử nghiệm nhỏ. Điều này là do dữ liệu huấn luyện và dữ liệu kiểm tra có sự tương đồng cao. Tuy nhiên, trong thực tế, nếu áp dụng với các câu phức tạp hơn hoặc chứa từ chưa thấy, độ chính xác sẽ giảm đáng kể do mô hình HMM chỉ dựa vào xác suất chuyển và phát đơn giản.

2.3 Nguyên nhân lỗi tiềm ẩn

- Dữ liệu huấn luyện nhỏ, chưa đủ bao quát.
- Từ chưa thấy (out-of-vocabulary) được gán xác suất rất nhỏ.
- Một số từ có thể mang nhiều loại từ khác nhau tùy ngữ cảnh (đa nghĩa).

2.4 Hướng cải thiện

- Mở rộng tập dữ liệu huấn luyện với nhiều cấu trúc câu khác nhau.
- Áp dụng smoothing nâng cao (như Good-Turing, Kneser-Ney).
- Kết hợp thêm đặc trưng ngữ cảnh hoặc sử dụng các mô hình hiện đại hơn như CRF hoặc BiLSTM-CRF.