**Trường đại học Khoa học Tự nhiên**

**Khoa Công nghệ Thông tin**



**BÁO CÁO ĐỒ ÁN MÔ HÌNH MARKOV ẨN (Phần 3: Vận dụng)**

**Môn**: Toán ứng dụng & thống kê

Nhóm sinh viên thực hiện:

20120018 – Trần Kiều Minh Lâm

20120067 – Nguyễn Phạm Bá Duy

20120316 – Nguyễn Hoàng Lâm

Giảng viên hướng dẫn: Nguyễn Đình Thúc, Nguyễn Văn Quang Huy, Võ Nam Thục Đoan

**Năm học:** 2021 – 2022

1. **Mô tả bài toán (đầu vào và đầu ra kỳ vọng)**

**Loại bài toán:** gán nhãn từ loại trong tiếng Anh (Part-of-speech tagging in English).

**Nội dung:** vận dụng mô hình Markov ẩn để gán nhãn từ loại của các từ trong câu (tiếng Anh). Ví dụ: Peter can talk to Mary → Noun Modal Verb Prep Noun.

**Tham số đầu vào**: tập dữ liệu (dataset) về các câu trong tiếng anh và đã được gắn nhãn loại từ sẵn.

**Kết quả đầu ra**: các tham số cần trong mô hình Markov ẩn. Để từ đó, có thể dùng các tham số và thuật toán Viterbi để gắn nhãn cho các câu trong tiếng Anh.

1. **Mô tả các thành phần của mô hình (tập các trạng thái ẩn, các quan sát có thể, v.v.)**

* Không gian mẫu của trạng thái ẩn (state space): là các loại từ trong câu (danh từ, tính từ, động từ, đại từ, giới từ, trạng từ, từ hạn định, liên từ, số, dấu câu, từ nước ngoài / viết tắt, tiểu từ - PRT)
* Không gian mẫu của trạng thái quan sát được (observation space): là từ trong tiếng Anh.
* Chuỗi trạng thái ẩn (state sequence): là các loại từ gắn nhãn trong câu.
* Chuỗi trạng thái quan sát được (observation sequence): là các từ trong câu.
* Ma trận xác suất ban đầu của trạng thái ẩn (starting state probs): là xác suất của một loại từ là từ đầu tiên của câu.
* Ma trận xác suất chuyển trạng thái ẩn (transition probs): là xác suất để loại từ này xuất hiện ngay sau loại từ kia.
* Ma trận xác suất đầu ra trạng thái quan sát được (emission probs): là xác suất để một loại từ gắn vào một từ.

**Các giả thiết của mô hình Markov ẩn có phù hợp với tình huống này hay không?**

Các giả thiết trong mô hình Markov ẩn đều hợp lý trong bài toán này vì:

* Các trạng thái ẩn (loại từ - tag) chỉ nằm trong không gian mẫu gồm 12 loại từ.
* Các từ quan sát được chỉ nằm trong không gian mẫu dataset.
* Mỗi loại từ thường thì chỉ phụ thuộc vào một loại từ liền trước nó. Ví dụ sau tính từ thường là danh từ.
* Mỗi từ trong câu chỉ có thể là một loại từ. Ví dụ như từ “can” không thể vừa là noun vừa là verb.

1. **Tập dữ liệu đã dùng là gì? Bạn đã áp dụng các bước tiền xử lý nào?**

* Tập dữ liệu sử dụng là Natural Language Toolkit (nlkt) trong Python. Cụ thể là treebank corpus và universal tagset.

**Các bước tiền xử lý:**

* Chia tập dữ liệu ra làm 2 phần. Một phần dùng để tính toán các ma trận xác suất (training) và một phần để đánh giá kết quả sau khi training (đánh giá mô hình) với tỷ lệ chia là 80:20.

1. **Bạn đã đánh giá mô hình như thế nào? Kết quả đánh giá sau cùng là gì?**

Để đánh giá mô hình sau khi train, ta sử dụng tập test đã tách ra từ trước và chạy thử mô hình. Lấy kết quả đạt được so sánh với tag của tập test và tính độ chính xác của mô hình.

Độ chính xác = số tag được gắn đúng / tổng số tag trong câu \* 100 (%)

Dưới đây là kết quả của 10 câu lấy ngẫu nhiên trong tập Test.

Ảnh có chứa văn bản

Mô tả được tạo tự động

Thời gian chạy của mô hình là 17.33 giây, với độ chính xác là 87.67%, cho thấy mô hình có kết quả tốt. Tuy nhiên thời gian vẫn còn khá chậm, một câu tương đương 1,7 giây.

Đây là kết quả của 100 câu lấy ngẫu nhiên trong tập test:

Không có mô tả.

Thời gian chạy của mô hình là 198.12 giây, với độ chính xác là 88.99%. Thời gian chạy của 100 câu cũng cao hơn khoảng 10 lần so với thời gian chạy của 10 câu. Về độ chính xác của mô hình vẫn ở mức tốt.

Ước lượng thời gian để chạy toàn bộ tập Test (bao gồm 20,000 câu) là khoảng 10 tiếng. Do đó nhóm lấy một mẫu nhỏ hơn làm đại diện.

1. **Nhận xét về kết quả và đề xuất một số hướng cải tiến.**

**Nhận xét:**

* Độ chính xác: cao, khoảng 88%
* Thời gian: chậm, khoảng 1,8 giây cho 1 câu.
* So với Deep Learning, mô hình này vẫn có kết quả thấp hơn (88% < 97%).
* Mặc dù với độ chính xác không cao bằng Deep Learning nhưng đây là mô hình đơn giản hơn nhiều, dễ cài đặt.

**Hướng cải tiến:**

* Các xác suất (emissions) được tính toán mỗi lần duyệt qua 1 từ trong câu, từ đó dẫn tới hiệu năng kém. Chúng ta có thể tạo một ma trận lưu lại các kết quả thay vì tính lại mỗi lần lặp.
* Ngoài ra chúng ta có thể thêm vào các patterns. Ví dụ khi một từ kết thúc bằng “ing”, “ed”, … thì từ đó sẽ thường là VERB.

|  |  |
| --- | --- |
| Số từ 0-9 | NUMBER |
| -ing | VERB |
| -ed | VERB |
| -es | VERB |
| ‘s | NOUN |
| s | NOUN |

* Để có thể áp dụng các trường hợp đặc biệt trên ta sẽ sử dụng RegEx.

**Tài liệu tham khảo**

<https://www.mygreatlearning.com/blog/pos-tagging/>