**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**



**BÁO CÁO ĐỒ ÁN MÔN HỌC**

CS221 - XỬ LÝ NGÔN NGỮ TỰ NHIÊN

GÁN NHÃN TỪ LOẠI CHO CÂU TIẾNG VIỆT

ÁP DỤNG MÔ HÌNH HIDDEN MARKOV

GVHD: ThS. Nguyễn Trọng Chỉnh

Nhóm sinh viên:

Phan Thành Nhân - 19521944

Hồ Thịnh - 19522274

MỤC LỤC

1. Giới thiệu bài toán:

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: trong câu “con ngựa đá đá con ngựa đá”, cùng một từ “đá” nhưng từ thứ nhất và thứ ba giữ chức năng ngữ pháp là danh từ nhưng từ thứ hai lại là động từ trong câu.

Ví dụ: ‘Anh ấy đang ở nhà’→ (‘Anh’, N), (‘ấy’, A), (‘đang’, R), (‘ở’, V), (‘nhà’, N)

Trong báo cáo này, nhóm sẽ trình bày về những công đoạn nhóm đã thực hiện để tạo ra một bộ gán nhãn từ loại dựa trên mô hình Hidden Markov cũng như giới thiệu về những kĩ thuật mà nhóm sử dụng.

1. Thống kê dataset:

Trong báo cáo này, nhóm đã thu thập khoảng 45 câu trong trích đoạn của ‘ ‘ để làm ngữ liệu. Ngữ liệu bao gồm các từ tiếng Việt, dấu câu, chữ số và các tên riêng tiếng Anh trong đó có các từ đơn, từ ghép 2-3 âm tiết. Theo thống kê của nhóm, trong ngữ liệu câu dài nhất chứa 117 từ, và ngắn nhất là 4 từ.

Các nhãn dùng để gán nhãn từ loại ứng với ngữ liệu nói trên gồm 14 nhãn như sau N, V, A, R, P, E, CH, C, Cc, Np, L, T, M, X được quy ước như các nhãn như trong Pen Treebank POS Tagging. Trong đó số lượng từ được gán nhãn V, N chiếm nhiều nhất.

1. Tiền xử lý:

Ở bước tiền xử lý, vì đây là một đoạn trích mà nhóm thu thập được thông qua internet nên nhóm đã kiểm tra lại các dấu câu, khoảng trắng xem có bị thừa hay không để loại bỏ. Tiếp đến, nhóm xem các dấu câu trong đoạn ngữ liệu như là 1 từ trong câu và thêm khoảng trắng giữa chúng và các từ thông thường trong câu để tiện cho việc xử lý. Cuối cùng, nhóm đã xuống dòng từng câu trong ngữ liệu để tiện cho việc xử lý ở các bước sau.

1. Tách từ:

Tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép… có trong câu. Đối với xử lý ngôn ngữ, để có thể xác định cấu trúc ngữ pháp của câu, xác định từ loại của một từ trong câu, yêu cầu nhất thiết đặt ra là phải xác định được đâu là từ trong câu. Vấn đề này tưởng chừng đơn giản với con người nhưng đối với máy tính, đây là bài toán rất khó giải quyết.

Có nhiều phương pháp để thực hiện tách từ như: Maximum Matching, Hidden Markov Model (HMM), Transformation based Learning (TBL),... Trong báo cáo này nhóm sử dụng một phương pháp đơn giản đó là Maximum Matching.

1. Maximum Matching:

Ý tưởng của phương pháp so khớp tối đa (Maximum

Matching), hay còn được gọi là so khớp tối đa từ trái qua phải (From

Left to Right Maximum Matching) là duyệt một câu vào từ trái qua

phải và chọn cụm từ dài nhất có mặt trong một từ điển từ vựng đã

cho. Quá trình này tiếp tục được lặp lại cho đến khi cụm từ tìm được

có độ dài giảm dần cho đến hết câu.

Thuật toán có độ phức tạp O(nV) với n là số âm tiết trong chuỗi, V là số từ trong từ điển.

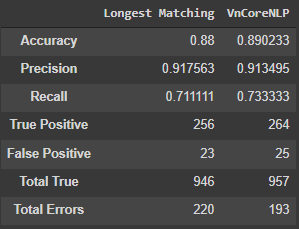
Ưu điểm của phương pháp so khớp tối đa là đơn giản, dễ hiểu và chạy nhanh. Hơn nữa phương pháp chỉ cần một tệp từ điển đầy đủ là có thể tiến hành phân đoạn các văn bản, hoàn toàn không phải trải qua huấn luyện như các phương pháp sẽ trình bày tiếp theo. Nhược điểm của phương pháp này là thuật toán gặp phải nhiều nhập nhằng và phụ thuộc vào từ điển.

1. Các vấn đề gặp phải:

Với từ điển đã được tạo sẵn, chúng ta có thể dễ dàng tách các từ ghép trong câu. Tuy nhiên cách tách từ trái sang phải đôi lúc cũng có những lỗi về mặt ngữ nghĩa. VD: ‘Tôi tập thể dục hằng ngày’ → ‘Tôi’, ‘tập thể’, ‘dục’, ‘hằng ngày’. Để giải quyết vấn đề này, chúng ta tiến hành chạy thêm tách từ phải sang trái để so sánh kết quả , giúp tăng độ chính xác.

1. Đánh giá kết quả tách từ đạt được:

Để đánh giá kết quả tách từ thông qua các phương pháp như Maximum Matching như đã giới thiệu ở trên hay sử dụng thư viện VNCoreNLP, nhóm đã chuẩn bị một ngữ liệu đã được tách từ thủ công gần đúng hoàn toàn và một tập từ vựng nhóm đã thu thập qua mạng để tách từ và đối sánh. Sau khi tách từ thông qua 2 phương pháp nói trên và đối sánh kết quả với kết quả tách tay, nhóm nhận được kết quả như bảng dưới đấy.

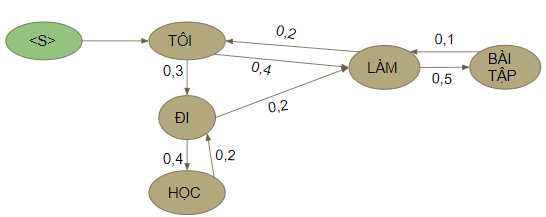


Kết quả cho thấy kết quả giữa 2 phương pháp này chênh lệch không quá nhiều, một phần là do tập từ vựng bao gồm khoảng 55000 từ mà nhóm đã thu thập khá đầy đủ nên việc tách từ bằng phương pháp Maximum Matching đạt được kết quả tốt. So với kết quả khi sử dụng thư viện VNCoreNLP thì kết quả của phương pháp Maximum Matching bị hạn chế khi các từ trong từ điển không có ví dụ như các từ ghép hay tên riêng tiếng Anh sẽ dẫn đến các trường hợp sai nhiều hơn.

1. Gán nhãn từ loại:
2. Mô hình Hidden Markov (HMM):

Trước khi nói về mô hình Hidden Markov chúng ta sẽ tìm hiểu về Markov Chains. Markov Chains là một mô hình cho chúng ta biết điều gì đó về xác suất của chuỗi các trạng thái ngẫu nhiên / biến. Chuỗi Markov đưa ra một giả định rất mạnh mẽ rằng nếu chúng ta muốn dự đoán tương lai trong chuỗi, tất cả những gì quan trọng là trạng thái hiện tại . Tất cả các trạng thái trước trạng thái hiện tại không có tác động đến tương lai ngoại trừ thông qua trạng thái hiện tại. Mô hình Markov Chains hoạt động dựa trên các thành phần như sau: tập gồm N trạng thái, ma trận chuyển đổi xác suất A (mỗi phần tử là xác suất chuyển đổi từ trạng thái hiện tại sang một trạng thái khác) và phân phối xác suất ban đầu.

Ví dụ Markov Chains:



Đôi khi, những gì chúng ta muốn dự đoán là một chuỗi các trạng thái không thể quan sát trực tiếp trong môi trường. Mặc dù chúng ta được cung cấp một chuỗi các trạng thái khác có thể quan sát được trong môi trường và các trạng thái ẩn này có một số phụ thuộc vào các trạng thái có thể quan sát được. Chúng ta có thể có các từ trong một câu là Trạng thái có thể quan sát (được cung cấp cho chúng tôi trong dữ liệu) nhưng Thẻ POS của chúng là trạng thái Ẩn và do đó chúng tôi sử dụng HMM để ước tính thẻ POS . Cần lưu ý rằng chúng tôi gọi các trạng thái Có thể quan sát là 'Quan sát' & Trạng thái ẩn là 'Trạng thái'.

Mô hình Markov ẩn (tiếng Anh là Hidden Markov Model - HMM) là mô hình thống kê trong đó hệ thống được mô hình hóa được cho là một quá trình Markov với các tham số không biết trước và nhiệm vụ là xác định các tham số ẩn từ các tham số quan sát được, dựa trên sự thừa nhận này. Các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp, ví dụ cho các ứng dụng nhận dạng mẫu. Trong một mô hình Markov điển hình, trạng thái được quan sát trực tiếp bởi người quan sát, và vì vậy các xác suất chuyển tiếp trạng thái là các tham số duy nhất. Mô hình Markov ẩn thêm vào các đầu ra: mỗi trạng thái có xác suất phân bổ trên các biểu hiện đầu ra có thể. Vì vậy, nhìn vào dãy của các biểu hiện được sinh ra bởi HMM không trực tiếp chỉ ra dãy các trạng thái.

1. Thuật toán Viterbi:

Để cải thiện hiệu suất của bài toán cũng như thời gian tính toán,

1. Đánh giá kết quả:

Tham khảo

* Tạp chí Khoa Học & Công Nghệ: Gán nhãn từ loại Tiếng Việt sử dụng mô hình Markov ẩn - Phạm Thị Minh Thu, Đào Thị Thuý Quỳnh.

<http://tailieudientu.lrc.tnu.edu.vn/Upload/Collection/brief/brief_36031_39590_171201315202339.pdf>

* Github Vietnamese POS tagging

<https://github.com/18520339/vietnamese-pos-tagging>