GÁN NHÃN TỪ LOẠI TIẾNG VIỆT BẰNG MÔ HÌNH TRƯỜNG NGẪU NHIÊN CÓ ĐIỀU KIỆN (CRF)

Nguyễn Văn Tú*, Lư Quang Hùng, Cao Văn Huấn*****
*** Trường Cao đẳng Kinh tế Thành phố Hồ Chí Minh**
**** Trường Đại học Hùng Vương Thành phố Hồ Chí Minh**
***** Đại học Y Dược Thành phố Hồ Chí Minh**
Email: caovanhuan@ump.edu.vn

Tóm tắt: *Xử lý ngôn ngữ tự nhiên (NLP) là lĩnh vực thu hút sự quan tâm nghiên cứu của nhiều chuyên gia trong và ngoài nước. Một trong những nhiệm vụ quan trọng của NLP là gán nhãn từ loại, tức là xác định từ loại của các từ trong văn bản. Việc gán nhãn từ loại hỗ trợ hiệu quả cho các công việc phân tích cú pháp, kiểm tra lỗi chính tả, phân loại văn bản, trích xuất thông tin, và xử lý tính đa nghĩa của từ. Trong nghiên cứu này, mô hình Conditional Random Fields (CRF) đã được áp dụng để thực hiện việc gán nhãn từ loại cho văn bản tiếng Việt. Qua thử nghiệm thực tế, kết quả đạt được có độ chính xác F1 là 96.21%. Điều này khẳng định tính hiệu quả và độ tin cậy cao của phương pháp đề xuất, đồng thời gợi mở hướng ứng dụng vào các công cụ xử lý ngôn ngữ tự nhiên cũng như hỗ trợ các ứng dụng giáo dục và đào tạo tiếng Việt.*

Từ khóa*: gán nhãn từ loại, Conditional Random Fields (CRF), Xử lý ngôn ngữ tự nhiên, Tiếng Việt.*

VIETNAMESE POS TAGGING USING CONDITIONAL RANDOM FIELDS (CRF)

Nguyen Van Tu*, Lu Quang Hung, Cao Van Huan*****
*** College of Economics, Ho Chi Minh City**
**** Hung Vuong University, Ho Chi Minh City**
***** University of Medicine and Pharmacy, Ho Chi Minh City**
Email: caovanhuan@ump.edu.vn

Abstract: *Natural Language Processing (NLP) is an area attracting significant research interest from experts both domestically and internationally. One of the critical tasks in NLP is Part-of-Speech (POS) tagging, which involves determining the grammatical category of words. Accurate POS tagging provides essential support for syntactic parsing, spell-checking, text classification, information extraction, and addressing lexical ambiguities. In this paper, the authors propose the use of Conditional Random Fields (CRF) for POS tagging in Vietnamese texts. Experimental results demonstrate an achieved F1 accuracy of 96.21%, highlighting the effectiveness and reliability of the proposed method.*

Keywords: *Conditional Random Fields, Natural Language Processing, Part-of-Speech tagging, Vietnamese.*

Nhận bài: 20/02/2025 *Phản biện: 23/03/2025* *Duyệt đăng: 26/03/2025*

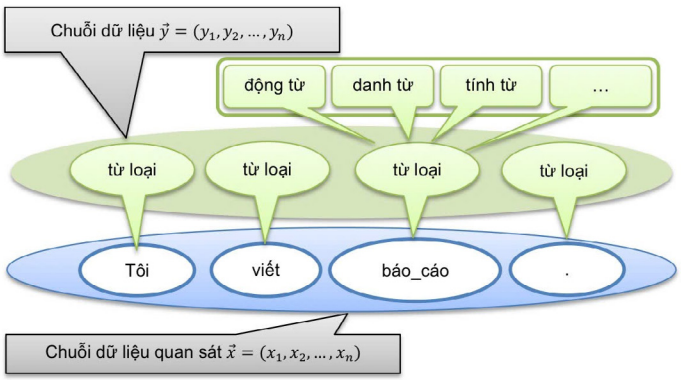
I. ĐẶT VẤN ĐỀ

Trong ngôn ngữ tự nhiên, từ vựng được chia thành các lớp (danh từ, động từ, tính từ, …) theo chức năng ngữ pháp. Bài toán gán nhãn từ loại chính là việc xác định chính xác chức năng ngữ pháp của các từ trong văn bản tiếng Việt. Việc xác định này giải quyết tính đa nghĩa của từ (Châu et al., 2006)

II. NỘI DUNG NGHIÊN CỨU

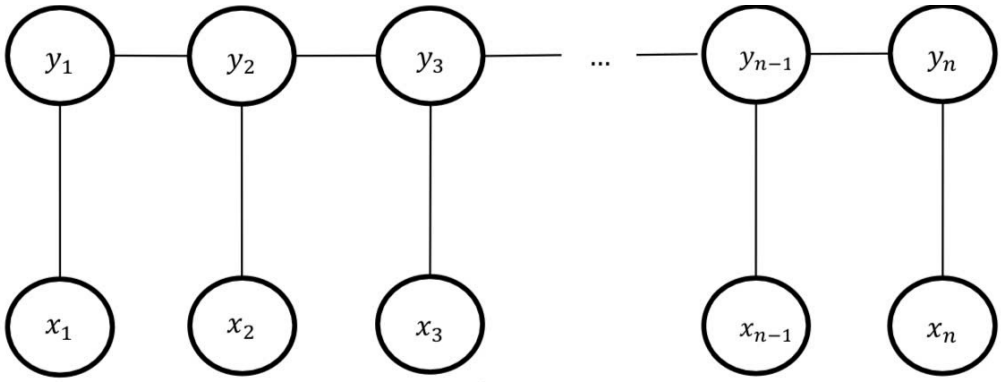
2.1. Mô hình CRF

Mô hình trường ngẫu nhiên có điều kiện CRF được giới thiệu lần đầu vào năm 2001 bởi Lafferty và các đồng nghiệp (Lafferty et al., 2001; Klinger & Tomanek, 2007). CRF là mô hình dựa trên xác suất điều kiện và có thể tích hợp được các thuộc tính đa dạng của chuỗi dữ liệu quan sát nhằm hỗ trợ cho quá trình phân lớp chính xác hơn. Dựa trên những tính chất đó, CRF giải quyết bài toán gán nhãn từ loại tiếng Việt như sau:



Hình 1. Mô hình gán nhãn từ loại

Gọi $\vec{x} = (x_1, x_2, \ldots, x_n)$ là biến ngẫu nhiên nhận giá trị chuỗi dữ liệu cần phải gán nhãn hay còn gọi dữ liệu đầu vào bao gồm n là thành phần dữ liệu và $\vec{y} = (y_1, y_2, \ldots, y_n)$ là biến ngẫu nhiên nhận giá trị của biến tương ứng (danh từ, tính từ, động từ, …). Vấn đề cần giải quyết ở đây là gán nhãn phù hợp cho x (Sutton & McCallum, 2004).



Hình 2. Đồ thị vô hướng mô tả CRF.

Một phân bố xác suất có thể được minh họa bởi một mô hình đồ thị vô hướng G(V,E) trong đó tập đỉnh V={$v_1, v_2, \ldots, v_T$} và tất cả các tập cạnh E mỗi đỉnh nhận giá trị từ một tập hữu hạn cho trước, phân bố cần quan tâm ở đây là p($v_1, v_2, \ldots, v_T$). Gọi C là tập tất cả các đồ thị con đầy đủ, sử dụng một tích các hàm không âm cực đại của đồ thị G. Các trọng số được xây dựng bằng cách mà các nút độc lập có điều kiện không xuất hiện chứa cùng hệ số, điều đó có nghĩa là chúng phụ thuộc vào các lớp khác nhau. Theo Hammmersley & Clifford (1971) định lý cơ bản về trường ngẫu nhiên (random fields) thì p($v_1, v_2, \ldots, v_T$) có thể được phân ra như sau:

$$p(v_1, v_2, \ldots, v_T) = \frac{1}{Z}\sum_{c\epsilon C} \Psi_c(\vec{v}_c) \qquad (2.1)$$

Các hệ số Ψ_c≥0 được gọi là hàm tiềm năng của biến ngẫu nhiên $\vec{v}_c$ chứa trong một lớp $c \in C$. Áp dụng kết quả từ định lý của Hemersley-Clifford cho trường ngẫu nhiên Markov, từ đó mô hình tổng quát của CRF là:

$$p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})}\prod_{c\epsilon C} \Psi_c(\vec{x}_c, \vec{y}_c) \qquad (2.2)$$

Hàm tiềm năng Ψc là các hệ số khác nhau đáp ứng với các cực đại trong đồ thị độc lập. Mỗi hệ số đáp ứng với một hàm tiềm năng mà kết hợp các đặc trưng f_i khác nhau của phần được xem xét của dãy quan sát và đầu ra. Chuẩn hóa theo với tham số $Z(\vec{x})$ bằng công thức sau:

$$Z(\vec{x}) = \sum_{\vec{y}'}\prod_{c\epsilon C} \Psi_c(\vec{x}_c, \vec{y}') \quad (2.3)$$

Thực tế, trong suốt cả quá trình huấn luyện và suy diễn, đối với mỗi thể hiện của đồ thị phần tách được xây dựng từ cái gọi là các mẫu lớp/nhóm. Các mẫu lớp tạo ra các giả thiết trên câu trúc của dữ liệu bằng việc định nghĩa cấu trúc của lớp.

2.2. Các dạng thuộc tính trong CRF

Thuộc tính chuyển f^e: biểu diễn sự phụ thuộc chuỗi bằng cách kết hợp nhãn (danh từ, tính từ, động từ, …) của trạng thái trước $y_{(j-1)}$ và nhãn (danh từ, tính từ, động từ, …) của trạng thái hiện tại y_j.

Thuộc tính quan sát f^o : mỗi đặc trưng trạng thái kết hợp nhãn (danh từ, tính từ, động từ, …) của trạng thái hiện tại y_j và một từ ngữ cảnh - một hàm nhị phân xác định các ngữ cảnh quan trọng của quan sát x tại vị trí đó.

Một số đặc trưng f_i quan trong cho bài toán gán nhãn từ loại tiếng Việt

Đặc trưng phát hiện tên thực thể:

- Kiểm tra từ hiện tại có được viết hoa ký tự đầu tiên hay không. Vì thực thể thường được viết hoa ký tự đấu tiên. Nhưng nếu nó đứng đầu câu thì thông thường không có ý nghĩa.

- Từ hiện tại có được viết hoa toàn bộ hay không. Thường những từ này có khả năng là tên tổ chức, tên người hay tên địa điểm.

Đặc trưng phát hiện từ láy:

Kiểm từ hiện tại có phải là từ láy hay không. Ta có từ láy 2 âm tiết, 3 âm tiết, 4 âm tiết được chia thành hai dạng là: (Ban, 2006; Cẩn, 1977)

- Từ láy bộ phận: là từ láy mà giữa các tiếng có sự giống nhau về phụ âm đầu hoặc phân vần.

- Từ láy toàn bộ: là từ láy có các tiếng lặp lại

nhau hoàn toàn (cũng có một số trường hợp tiếng đứng trước biến đổi thanh điệu hoặc phụ âm cuối).

Đặc trưng âm tiết tiếng Việt:

- Kiểm tra âm tiết có trong từ điển âm tiết tiếng Việt hay không (nếu có là tiếng Việt, không có thì nó là tiếng nước ngoài). Dùng để xác định từ nước ngoài.

- Các âm tiết tiếng Viêt có trong từ điển như sau: "á, à, ả, ã, ạ, ấ, ầ, ẩ, ẫ, ậ, ắ, ằ, ẳ, ẵ, ặ, é, è, ẻ, ẽ, ẹ, ế, ề, ể, ễ, ệ, í, ì, ỉ, ĩ, ị, ó, ò, ỏ, õ, ọ, ố, ồ, ổ, ỗ, ộ, ớ, ờ, ở, ỡ, ợ, ú, ù, ủ, ũ, ụ, ứ, ừ, ử, ữ, ự, ý, ỳ, ỷ, ỹ, ỵ"

Nhập nhằng trong từ loại tiếng Việt: Trong ngôn ngữ học, nhập nhằng là hiện tượng phổ biến trong giao tiếp hằng ngày, và con người thường dễ dàng xác định được nhờ ngữ cảnh. Tuy nhiên, với máy tính, việc xử lý nhập nhằng lại trở nên khó khăn hơn rất nhiều. Một số dạng nhập nhằng thường gặp bao gồm: nhập nhằng ranh giới từ, nhập nhằng từ đa nghĩa, nhập nhằng từ đồng âm, nhập nhằng từ loại và nhập nhằng thực thể. Do đó, việc lựa chọn các đặc trưng phù hợp đóng vai trò quan trọng trong quá trình khử nhập nhằng từ đa nghĩa. Các phương pháp phổ biến được sử dụng là:

- Khử nhập nhằng bằng đặt trưng N-gram.
- Khử nhập nhằng bằng luật chuyển.
- Khử nhập nhằng bằng đặc trưng tự điển.
- Khử nhập nhằng đặt trưng tiền tố và hậu tố.

Thử nghiệm và đánh giá

Chia tập ngữ liệu để thử nghiệm: Kho ngữ liệu ban đầu (corpus.pos.txt) được tổng hợp từ 73 file nhỏ hơn, bao gồm 5.475 câu với tổng cộng 115.425 từ đã được tách từ và gán nhãn từ loại (Trần, 2016). Sau đó, kho ngữ liệu này được chia thành các tập nhỏ hơn phục vụ mục đích thử nghiệm như sau:

- Tập huấn luyện (train.pos.txt): chiếm 90% dung lượng của tập corpus ban đầu, gồm 4.928 câu với tổng số 103.928 từ đã được tách từ và gán nhãn từ loại.

- Tập kiểm thử (test.pos.txt): chiếm 10% còn lại của corpus, gồm 547 câu với tổng số 11.497 từ đã được tách từ và gán nhãn từ loại.

- Tập thử nghiệm (test.raw.txt): bao gồm 547 câu tương ứng với 11.497 từ đã được tách từ nhưng đã loại bỏ nhãn từ loại từ tập test.pos.txt, nhằm phục vụ cho việc đánh giá độ chính xác của mô hình.

Quy trình thử nghiệm như sau:

- Bước 1: Đọc files dữ liệu huấn luyện đã được gán nhãn từ loại.

- Bước 2: Trích chọn đặc trưng, huấn luyện mô hình CRF.

- Bước 3: Gán nhãn từ loại cho văn bản sử dụng thuật toán quy hoạch động Viterbi.

- Bước 4: Đánh giá độ chính xác của chương trình.

*** Độ chính xác và cách tính cũng như phép đo độ chính xác:**

Để đánh giá hiệu quả của các phương pháp tác giả sử dụng phép đo độ chính xác như sau:

- Độ đo phủ (Recall): R = Nđ/Nm
- Độ đo chính xác (Precision): P = Nđ/Ng
- Độ đo cân đối: F1-score: F1 = 2RP/(R+P)

* Ký hiệu:

Nm: số đơn vị từ trong văn bản mẫu.

Ng: số đơn vị từ chương trình gán được.

Nđ: số đơn vị từ chương trình gán đúng.

*** Kết quả như sau**

Kết quả trình bày kết quả lần thử nghiệm thứ nhất về việc gán nhãn từ loại. Qua đó, nhận thấy các nhãn phổ biến như N, Np, CH, M, R, A, P, V, Nc, E, L, C, Ny, Nu, X đều có tỉ lệ gán sai không đáng kể, trong đó nhãn CH đạt độ chính xác cao nhất. Ngược lại, những nhãn ít phổ biến hơn như T, Z, Y, I có tỉ lệ gán sai khá cao; đáng chú ý nhất là nhãn T có tỉ lệ sai cao nhất, riêng hai nhãn Y và I không có kết quả gán đúng nào. Đối với các nhãn Nb, Vb, Ab là những từ viết tắt hoặc từ có nguồn gốc tiếng nước ngoài, chương trình còn hạn chế hoặc chưa nhận diện đúng. Do đó, việc xây dựng và cập nhật thường xuyên một từ điển các từ nước ngoài là điều rất cần thiết. Qua phân tích các kết quả thực nghiệm, có thể khẳng định rằng việc xây dựng kho ngữ liệu phong phú và từ điển có độ bao phủ cao là yếu tố quan trọng giúp giảm thiểu các lỗi gán sai và nâng cao chất lượng mô hình.

Kết quả thể hiện kết quả qua 10 lần thử nghiệm liên tiếp. Qua đó, nhận thấy rằng độ chính xác khi gán nhãn từ loại tiếng Việt khá ổn định: kết quả thấp nhất là 96,00% (ở lần thử nghiệm thứ 2), cao nhất đạt 96,52% (ở lần thứ 5), trung bình độ chính xác qua 10 lần thử nghiệm đạt 96,21%. Điều này

cho thấy mô hình hoạt động tương đối ổn định và có độ tin cậy cao trong việc gán nhãn từ loại cho tiếng Việt.

III.KẾT LUẬN

Kết quả của nghiên cứu này không chỉ hữu ích trong việc nâng cao chất lượng của các hệ thống xử lý ngôn ngữ tự nhiên, mà còn góp phần quan trọng vào việc phát triển các ứng dụng giáo dục, hỗ trợ dạy và học tiếng Việt, đặc biệt là trong lĩnh vực xây dựng phần mềm giáo dục và thiết kế chương trình đào tạo trực tuyến.

TÀI LIỆU THAM KHẢO

Bạn, D. Q. (2005). *Ngữ pháp Tiếng Việt (Tập 1 & 2*). Nhà xuất bản Giáo dục.

Cẩn, N. T. (1977). *Ngữ pháp tiếng Việt: Tiếng - từ ghép - đoản ngữ.* Nhà xuất bản ĐH & THCN.

Châu, N. Q., Tươi, P. T., & Trụ, C. H. (2006). *Gán nhãn từ loại cho tiếng Việt dựa trên văn phong và tính toán xác suất.* Tạp chí Phát triển KH&CN, 9(2).

Hammmersley, J. M., & Clifford, P. (1971). *Markov fields on finite graphs and lattices*. Retrieved from https://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf

Klinger, R., & Tomanek, K. (2007). *Classical probabilistic models and conditional random fields (Algorithm Engineering Report TR07-2-013)*. Department of Computer Science, Dortmund University of Technology. Retrieved from https://doi.org/10.17877/DE290R-7970

Lafferty, J., McCallum, A., & Pereira, F. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proceedings of the 18th International Conference on Machine Learning (ICML-2001) (pp. 282–289). Morgan Kaufmann. ISBN: 9781558607781.

Sutton, C., & McCallum, A. (2004). *Collective segmentation and labeling of distant entities in information extraction (Technical Report TR 04-49)*. University of Massachusetts. Presented at ICML 2004 Workshop on Statistical Relational Learning.

Trần, N. A. (2016). *Nghiên cứu phát triển một số kỹ thuật tách từ tiếng Việt (Luận án tiến sĩ toán học)*. Học viện Kỹ thuật Quân sự, Hà Nội.