Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00084

SỰ ẢNH HƯỞNG CỦA PHƯƠNG PHÁP TÁCH TỪ TRONG BÀI TOÁN PHÂN LỚP VĂN BẢN TIẾNG VIỆT

Phạm Nguyên Khang, Trần Nguyễn Minh Thư, Phạm Thế Phi, Đỗ Thanh Nghị
Khoa Công nghệ thông tin & Truyền thông, Trường Đại học Cần Thơ
{pnkhang, tnmthu, ptphi, dtnghi}@cit.ctu.edu.vn

TÓM TẮT— *Tách từ là một bước quan trọng không thể thiếu trong xử lý ngôn ngữ tự nhiên, nhằm xác định được ranh giới các từ có trong văn bản. Trong tiếng Việt, ngoài từ đơn (một âm tiết), còn có từ ghép (đa âm tiết). Điều này gây khó khăn cho việc tách từ tự động một cách chính xác, ảnh hưởng đến kết quả của các bài toán phân tích dữ liệu văn bản như: gom nhóm, phân lớp văn bản. Hai tiếp cận chính để tách từ là dựa trên từ điển và thống kê (hoặc kết hợp hai tiếp cận). Trong bài toán phân lớp văn bản, tách từ mới chỉ là bước tiền xử lý và biểu diễn dữ liệu. Bước kế tiếp là sử dụng một mô hình máy học để huấn luyện bộ phân lớp. Đối với một số mô hình máy học như máy học véc-tơ hỗ trợ (SVM), phân tích thành phần chính, phân tích tương ứng, các từ ghép có thể được phát hiện dựa vào sự đồng xuất hiện của các âm tiết mà không cần đến một bước tách từ chính xác. Trong bài báo này, chúng tôi nghiên cứu so sánh sự ảnh hưởng của các phương pháp tách từ lên hiệu quả phân lớp văn bản tiếng Việt, để từ đó chọn ra phương pháp hiệu quả nhất. Thực nghiệm trên tập dữ liệu 6,000 văn bản thuộc 10 chủ đề và tập dữ liệu 105,293 quyển sách thuộc 166 chủ đề với giải thuật máy học SVM cho thấy rằng kết quả phân lớp với các phương pháp tách từ khác nhau tuy có sự khác biệt nhưng không có ý nghĩa thống kê trong bài toán phân lớp văn bản tiếng Việt.*

Từ khóa— *Tách từ, phương pháp tách từ tiếng Việt, xử lý ngôn ngữ tự nhiên, phân lớp văn bản.*

I. GIỚI THIỆU

Với sự phát triển nhanh chóng của công nghệ thông tin, nguồn thông tin trực tuyến (online) dưới dạng văn bản xuất hiện càng ngày càng nhiều. Nguồn thông tin này đến từ các thư viện điện tử, thư điện tử, trang web, hệ thống tìm kiếm và tra cứu thông tin. Việc khám phá tri thức tiềm ẩn từ kho dữ liệu văn bản là cần thiết cho việc quản lý, khai thác hiệu quả nguồn thông tin văn bản khổng lồ này. *Phân lớp văn bản* (text categorization) là một trong những kỹ thuật chính để xử lý và tổ chức dữ liệu văn bản. Kỹ thuật phân lớp văn bản được dùng để gán nhãn tự động các bản tin, sắp xếp tổ chức email hay tập tin, nhận dạng thư rác. Có để định nghĩa ngắn ngọn bài toán phân lớp văn bản như sau: gán nhãn cho từng văn bản theo chủ đề đã được định nghĩa trước dựa vào nội dung của văn bản. Phân lớp văn bản thường được dựa trên mô hình ngữ nghĩa hoặc máy học. Tuy nhiên như bài phỏng vấn được thực hiện bởi M. Lucas (Tạp chí Mappa Mundi) năm 1999, M. Hearst cho rằng tiếp cận ngữ nghĩa là vấn đề rất khó, phức tạp. Vì vậy, tiếp cận dựa trên máy học tự động lại đơn giản và cho nhiều kết quả tốt trong thực tiễn. Hầu hết các phương pháp phân loại văn bản dựa trên mô hình thống kê từ và các giải thuật máy học phân lớp (Dumais et al., 1998), (Sebastiani, 1999), (Manning et al., 2008).

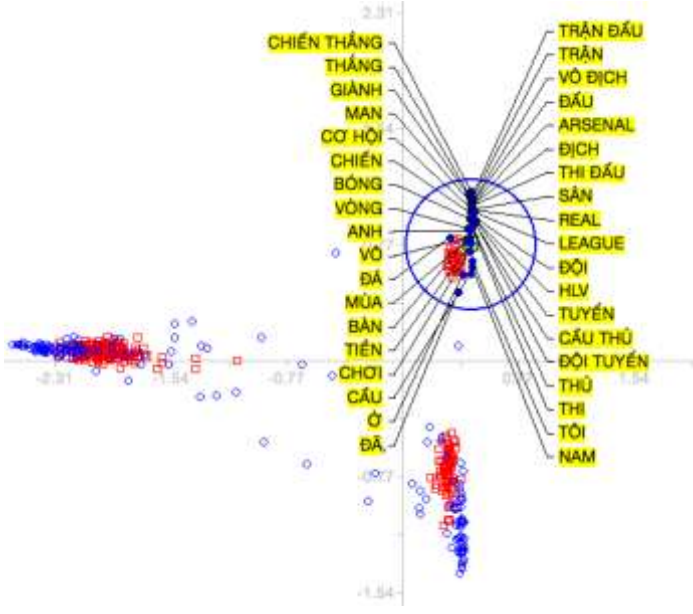
Bước đầu tiên trong phân lớp văn bản là biến đổi văn bản từ chuỗi ký tự về dạng phù hợp với các giải thuật học máy. Đặc điểm chung của nguồn dữ liệu văn bản là không có cấu trúc (độ dài khác nhau) trong khi đa số các giải thuật đòi hỏi dữ liệu huấn luyện phải có cấu trúc (chiều dài các véc-tơ đặc trưng phải giống nhau chẳng hạn). Các nghiên cứu trong lĩnh vực truy vấn thông tin đã chỉ ra rằng thứ tự của các từ trong văn bản đóng vai trò không quan trọng lắm đối với hầu hết các bài toán phân tích, xử lý dữ liệu văn bản (Joachims, 1999). Chính vì thế mô hình túi từ (Salton et al., 1975) là một mô hình phổ biến cho biểu diễn dữ liệu văn bản. Theo mô hình này, mỗi từ (khác nhau) trong văn bản sẽ là một *đặc trưng* (feature) và tần số xuất hiện của nó trong văn bản là giá trị của đặc trưng tương ứng. Quá trình trích đặc trưng bao gồm *tách từ* (word segmentation) và đếm số lần xuất hiện của các từ trong văn bản. Như thế, văn bản sẽ được biểu diễn dưới dạng véc-tơ tần số.

Bước tiếp theo là huấn luyện mô hình học tự động từ bảng dữ liệu này. Các mô hình máy học thường sử dụng như giải thuật *k*-NN (Fix & Hodges, 1952), naive Bayes (Good, 1965), cây quyết định (Quinlan, 1993), (Breiman et al., 1984), máy học véc-tơ hỗ trợ (Vapnik, 1995), giải thuật tập hợp mô hình bao gồm Boosting (Freund & Schapire, 1995), (Breiman, 1998) và rừng ngẫu nhiên (Breiman, 2001). Các nghiên cứu về máy học trước đây của (Phạm et al., 2006), (Phạm et al., 2008), (Đỗ, 2012), (Đỗ & Phạm, 2013) đề xuất các giải thuật máy học dựa trên tập hợp mô hình, máy học véc-tơ hỗ trợ, naive Bayes, cho phép phân lớp hiệu quả các tập dữ liệu có số chiều lớn như biểu diễn văn bản bằng mô hình túi từ.

Đối với các ngôn ngữ như tiếng Anh, tiếng Pháp, tiếng Đức việc tách từ được thực hiện khá đơn giản dựa vào các ký tự phân cách như: khoảng trắng, ký tự tab, các dấu câu, dấu ngoặc, v.v. Ngược lại, đối với tiếng Việt (và các ngôn ngữ châu Á khác như tiếng Trung Quốc, tiếng Nhật Bản, tiếng Hàn) khoảng trắng ngoài việc ngăn cách các từ với nhau, còn được dùng để ngăn cách các *âm tiết* (syllable) của một từ ghép, ví dụ: câu "Học sinh đi học" phải được tách thành "Học_sinh/đi_học". Khoảng trắng thứ nhất và thứ ba dùng để ngăn cách các âm tiết của một từ và khoảng trắng thứ hai dùng để ngăn cách hai từ với nhau. Điều này gây khó khăn cho quá trình tách từ. Các phương pháp tách từ tiếng Việt (và các ngôn ngữ châu Á khác) đều dựa trên thông tin về sự xuất hiện cạnh nhau của các âm tiết (colocation). Hai tiếp cận chính để tách từ tiếng Việt là (i) dựa trên từ điển và (ii) tiếp cận thống kê. Ngoài ra còn có một số phương pháp kết hợp cả hai tiếp cận trên. Trong tiếp cận dựa trên từ điển, một chuỗi các âm tiết sẽ được xem là một từ ghép nếu

chuỗi các âm tiết này có trong từ điển. Tiếp cận thống kê dựa trên sự xuất hiện cạnh nhau của các âm tiết, nếu sự xuất hiện cạnh nhau này xảy ra thường xuyên các âm tiết này rất có thể thuộc về một từ ghép nào đó. Cho dù sử dụng tiếp cận nào, nhập nhằng (ambiguous) trong việc tách từ cũng có thể xảy ra. Nhập nhằng xảy ra khi có nhiều hơn một cách tác các từ trong một câu. Để khử nhập nhằng, phương pháp thường dùng là cực đại hoá độ hợp lý (Maximum Likelihood Estimation) với giải thuật Viterbi-like. Điều này làm cho quá trình biểu diễn văn bản thường mất rất nhiều thời gian.

Trong khi nghiên cứu phân tích so sánh vai trò của các âm tiết và bản thân từ ghép trong việc hình thành các chủ đề văn bản cho bài toán phát hiện chủ đề văn bản, chúng tôi nhận thấy rằng các âm tiết của một từ ghép có vai trò tương đương với từ ghép được tạo nên từ các âm tiết này. Hình 1 hiển thị kết quả của việc áp dụng *Phân tích tương ứng* (Correspondence Analysis hay CA) (Benzécri, 1973) trên các văn bản của tập dữ liệu vnexpress (gồm 3 chủ đề: công nghệ thông tin, thể thao và nấu ăn[1]). Áp dụng CA trên dữ liệu văn bản cho phép (i) gom nhóm các văn bản có nội dung tương tự nhau (tạo nên chủ đề), (ii) gom nhóm các từ tạo nên chủ đề và (iii) hiển thị các nhóm văn bản và các nhóm từ tương ứng cạnh nhau trong không gian rút gọn của CA. Ta dễ dàng thấy rằng, mặt dù là một phương pháp không giám sát, CA vẫn cho phép phát hiện 3 nhóm văn bản tương ứng với 3 chủ đề có trong tập dữ liệu. Kết quả hiển thị trong hình 1 cũng chỉ ra rằng các từ ghép quan trọng trong chủ đề "**thể thao**" bao gồm: "**chiến thắng**", "**cơ hội**", "**trận đấu**", "**vô địch**", "**thi đấu**", "**cầu thủ**", "**đội tuyển**". Điều thú vị là các âm tiết tạo nên các từ ghép này "**chiến**", "**thắng**", "**vô**", "**địch**", "**cầu**", "**thủ**", … cũng xuất hiện nằm trong danh sách các từ tạo nên chủ đề "thể thao" và ở ngay bên cạnh các từ ghép tương ứng trong không gian rút gọn của CA.



Hình 1. Vai trò của các âm tiết trong việc tạo nên chủ đề của văn bản.

Kết quả phân tích trực quan với CA cho thấy rằng thông tin về sự đồng xuất hiện (không kể vị trí) của các âm tiết trong một văn bản cũng đủ để hình thành nên chủ đề của văn bản mà không cần đến quá trình tách từ (sử dụng thông tin về sự xuất hiện cạnh nhau). Nói cách khác bản thân âm tiết (chứ không phải từ ghép) cũng góp phần vào việc phân biệt các chủ đề/lớp văn bản. Kết quả này cho phép chúng ta đặt giả thiết: "*liệu chỉ với thông tin đồng xuất hiện của các âm tiết có đủ để huấn luyện một bộ phân lớp mạnh để phân lớp chính xác văn bản tiếng Việt*", cụ thể hơn:

- Biểu diễn văn bản dựa trên từ ghép (được tách từ đúng) so với biểu diễn văn bản chỉ đơn thuần dựa trên âm tiết có ảnh hưởng đến hiệu quả phân lớp không?
- Tách từ sai (ghép các âm tiết không cùng một từ ghép) có ảnh hưởng đến hiệu quả phân lớp không?

Trong bài báo này, chúng tôi thực hiện một nghiên cứu so sánh về sự ảnh hưởng của các phương pháp tách từ tiếng Việt đối với hiệu quả phân lớp trong bài toán phân loại văn bản tiếng Việt. Kết quả thực nghiệm trên tập dữ liệu văn bản gồm 6000 văn bản thuộc 10 chủ đề của trang báo điện tử vnexpress.net và tập dữ liệu thư viện gồm 166 chủ đề cho thấy rằng việc tách từ đa âm tiết (tạo ra các từ ghép) và tách từ đơn âm tiết (đơn thuần dựa trên khoảng trắng) có ảnh hưởng không đáng kể đối với hiệu quả phân lớp.

Phần tiếp theo của bài viết được trình bày như sau: phần II lượt khảo một số phương pháp tách từ tiếng Việt bao gồm: tiếp cận dựa trên từ điển, tiếp cận dựa trên thống kê và tiếp cận lại; phần III trình bày phân loại văn bản với mô hình túi từ và máy học véc-tơ hỗ trợ; phần IV trình bày các kết quả thực nghiệm trước khi kết luận và hướng phát triển.

[1]Chúng tôi đã xử lý tập dữ liệu này bằng phương pháp tách từ dựa trên từ điển và để chúng chứa cả các từ ghép lẫn các âm tiết tạo nên từ ghép.

II. TÁCH TỪ TIẾNG VIỆT

Từ trong tiếng Việt, ngoài từ đơn (một âm tiết), còn có từ ghép (đa âm tiết), chính vì vậy không thể dùng khoảng trắng để xác định ranh giới của các từ. Những âm tiết được kết hợp để tạo thành các từ khác nhau tùy thuộc vào ngữ cảnh của văn bản. Để nhận dạng đúng ranh giới của các từ (tách từ) phục vụ cho các bài toán phân tích dữ liệu văn bản như: gom nhóm, phân lớp văn bản, các nhà khoa học đã đề xuất nhiều phương pháp tách từ. Dựa trên đặc điểm của "từ" kết hợp với cách tiếp cận khác nhau, các phương pháp tách từ này có thể chia thành ba nhóm chính: dựa trên từ điển (dictionary-based), dựa trên thống kê (statistic-based) và phương pháp lai (hybrid).

A. Tiếp cận dựa trên từ điển

Ý tưởng chính của phương pháp tách từ dựa trên từ điển là từ một từ điển sẵn có, thực hiện so khớp từng âm tiết trong văn bản với các từ có trong từ điển. Tuỳ vào cách thức so khớp mà ta có các phương pháp khác nhau như: so khớp từ dài nhất (longest matching), so khớp từ ngắn nhất (short matching), so khớp chồng lắp (overlap matching) và so khớp cực đại (maximum matching) (Dinh et al., 2001), (Pham et al., 2009). Độ chính xác của phương pháp dựa trên từ điển phụ thuộc rất lớn vào kích thước từ điển được xây dựng. Với đặc điểm là không cần phải có bước huấn luyện nên thời gian xử lý của phương pháp này tương đối nhanh, đơn giản và dễ hiểu. Tuy nhiên, phương pháp này sẽ khó có thể xử lý được các tình huống nhập nhằng cũng như xử lý tình huống xuất hiện từ mới không tồn tại trong từ điển. Hai phương pháp thường được sử dụng của tiếp cận từ điển là phương pháp so khớp từ dài nhất và phương pháp so khớp cực đại:

- Phương pháp so khớp từ dài nhất (Surapant Meknavin et al., 1997): với mỗi câu, duyệt từ trái qua phải các âm tiết trong câu, kiểm tra xem có nhóm các âm tiết có tồn tại từ trong từ điển hay không. Chuỗi dài nhất các âm tiết được xác định là từ sẽ được chọn ra. Tiếp tục thực hiện việc so khớp cho đến hết câu. Ví dụ "Học sinh học sinh vật học", từ trái qua phải, âm tiết đầu tiên là "học", "học" cũng có thể là 1 từ đơn, nhưng "học" cũng có thể kết hợp với âm tiết "sinh" để tạo nên từ ghép "học sinh", ta được từ đầu tiên là "học sinh", xét tiếp các âm tiết còn lại cho đến khi hết câu ta có các từ sau: "học sinh", "học sinh", "vật", "học". Với ví dụ này, phương pháp so khớp từ dài nhất không đem lại kết quả như mong muốn.
- Phương pháp so khớp cực đại (Chih-Hao Tsai, 1996), (Surapant Meknavin et al., 1997): ứng với mỗi câu dữ liệu đầu vào, tìm tất cả các trường hợp mà các âm tiết có thể kết hợp lại để tạo nên các từ có nghĩa. Ứng với mỗi loại ngôn ngữ khác nhau thì sự lựa chọn các nhóm âm tiết này có thể khác nhau. Phương pháp này là so khớp toàn diện cho một câu thay vì so khớp cục bộ âm tiết đang được xét. Với ví dụ: "Học sinh học sinh vật học": các trường hợp kết hợp của các âm tiết có thể có "sinh vật học", "học sinh", "học", từ được tách trong câu sẽ chính xác hơn phương pháp so khớp từ dài nhất.

B. Tiếp cận dựa trên thống kê

1. Mô hình ngôn ngữ

Với cách tiếp cận dựa trên thống kê, các giải pháp cho việc tách từ thông thường dựa trên mô hình ngôn ngữ (language model – LM) (Jelinek et al., 1991). Một LM thường được xây dựng dựa trên việc thu thập thống kê số lần xuất hiện hoặc đồng xuất hiện của các từ trong một tập lớp các văn bản. Với một đoạn văn bản $w_1^n = w_1 w_2 ... w_n$, mô hình LM được dùng để tính xác suất $P(w_1^n)$ của đoạn văn bản này. Công thức tính xác suất tổng quát có thể được biểu diễn như sau:

$$P(w_1^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1}) = \prod_{k=1}^{n} P(w_k \mid w_1^{k-1}) \quad (1)$$

Ứng dụng giả thuyết của Markov rằng dự đoán kế tiếp chỉ phụ thuộc vào lịch sử gần đây thay vì toàn bộ lịch sử, chúng ta có thể biểu diễn công thức (1) bằng công thức sau:

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_{k-N+1}^{k-1}) \quad (2)$$

với *N* chỉ ra lịch sử gần nhất hay cụ thể hơn là số lượng từ gần nhất đứng trước từ thứ *k* hiện tại. Mô hình LM này thường được gọi là mô hình **n-grams.**

Việc ước lượng $P(w_k \mid w_{k-N+1}^{k-1})$ hay $P(w_k \mid w_{k-N+1},...,w_{k-1})$ được thực hiện như sau:

$$p(w_n \mid w_{n-N+1},...,w_{n-1}) = \frac{C(w_{n-N+1},...,w_n)}{C(w_{n-N-1},...,w_{n-1})} \quad (3)$$

với $C(w_{n-N+1},...,w_n)$ là số lần xuất hiện của dãy các từ $w_{n-N+1},...,w_n$ trong tập dữ liệu huấn luyện; và $C(w_{n-N+1},...,w_{n-1})$ là số lần xuất hiện của dãy các từ $w_{n-N+1},...,w_{n-1}$ trong tập dữ liệu huấn luyện.

2. Phương pháp tách từ sử dụng mô hình Markov ẩn

Phương pháp tách từ dựa theo thống kê (ở đây là mô hình **n-grams**) cơ bản nhất được đề xuất bởi Luo và đồng sự (Luo et al., 1996). Ở đó các tác giả đề xuất một mô hình Markov ẩn (Hidden Markov Model – HMM) để biểu diễn các khả năng tách các từ trong một câu tiếng Trung. Việc tách từ trong một câu tiếng Trung cũng tương đồng như việc tách các từ trong một câu tiếng Việt. Nghĩa là có một số từ nên đứng riêng, còn một số từ khác nên được ghép chung để thành từ ghép.

Chúng tôi sử dụng mô hình HMM này trong tách từ tiếng Việt như sau. Gọi *S* là một câu tiếng Việt bao gồm *n* từ $w_1 w_2 ... w_n$ với w_i là một từ trong câu. Bài toán đặt ra là tách câu này thành các cụm từ thích hợp:

$$\begin{aligned} S &= w_1 w_2 ... w_n \\ &= (w_1 ... w_{x_1})(w_{x_1+1} ... w_{x_2}) ... (w_{x_m+1} ... w_{x_m}) \\ &= C_1 C_2 ... C_m \end{aligned} \quad (4)$$

với x_i là vị trí của từ cuối cùng của cụm từ (từ ghép) thứ *i*: $C_i = w_{x_i-1} ... w_{x_i}$, với $i = 1,2,...,m$ và $x_0 = 0, x_m = n$.

Một cách phân tách các từ trong câu *S* bây giờ có thể được biểu diễn bởi một dãy các số nguyên $x_1,...,x_m$.

Gọi $G(S)$ là tập tất cả các cách để phân tách các từ trong câu *S*:

$$G(S) = \{(x_1 ... x_m) : 1 \le x_1 \le ... \le x_m, m \le n\} \quad (5)$$

Giả sử chúng ta sử dụng mô hình n-grams như trên thì với một cách phân đoạn $g(S) = (x_1 ... x_m) \in G(S)$, khả năng của cách phân đoạn này được ước lượng như sau:

$$\begin{aligned} L(g(S)) &= \log P_g(C_1 ... C_m) \\ &= \sum_{i=1}^{m} \log P_g(C_i \mid h_i) \end{aligned} \quad (6)$$

với h_i là lịch sử gần nhất của cụm từ C_i. Trong các thí nghiệm được trình bày trong phần sau, chúng tôi sử dụng mô hình ngôn ngữ **unigram**, nghĩa là một cụm từ được tính khả năng xuất hiện độc lập so với các cụm từ khác. Và chúng tôi cũng giới hạn mỗi cụm từ có tối đa 2 từ.

Trong tất cả các cách phân đoạn có thể có, chúng tôi sẽ chọn ra cách phân đoạn ***g**** là cách phân đoạn cuối cùng với khả năng được ước lượng cao nhất.

$$\begin{aligned} g^* &= \underset{g \in G(S)}{\operatorname{argmax}} L(g(S)) \\ &= \underset{g \in G(S)}{\operatorname{argmax}} \log P_g(C_1 ... C_m) \end{aligned} \quad (7)$$

Việc ước lượng ***g**** được thực hiện bởi phương pháp Viterbi như đề xuất của (Luo et al., 1996).

3. Phương pháp tách từ sử dụng mô hình trường xác suất có điều kiện và độ hỗn loạn cực đại

Phương pháp tách từ sử dụng mô hình trường xác suất có điều kiện (CRFs) và độ hỗn loạn cực đại (MaxEnt) được đề xuất bởi (Nguyen et al., 10). Bài toán tách từ được xem như là công việc gán nhãn cho một dãy các từ. Một từ đơn tiếng Việt mà đứng đầu một từ ghép được gắn nhãn B_W, một từ đơn nằm trong một từ ghép được gán nhãn I_W và những thứ khác ví dụ như dấu phẩy, dấu chấm được gán nhãn O (Outside of a word). Bài toán nhằm tìm ra ranh giới giữa các từ trong một câu trở thành bài toán gán nhãn các từ đơn trong câu với các loại nhãn như vừa nêu trên.

Nguyen và các cộng sự đề xuất sử dụng mô hình CRFs để mô hình hóa bài toán. CRFs được biểu diễn như là một chuỗi tuyến tính vô hướng các trạng thái của mô hình. Mỗi trạng thái ở đây được gán một trong các nhãn (như trình bày bên trên). Nhãn thích hợp nhất được xác định dựa trên quan sát của từ tương ứng với trạng thái đó cũng như các trạng thái đứng trước. Xác suất của một trạng thái biết trước một từ tương ứng được ước lượng dựa vào hàm đặc trưng (được xây dựng dựa trên ước lượng độ hỗn loạn MaxEnt). Nguyen và các cộng sự sử dụng hai loại hàm tính đặc trưng (feature function) trong các mô hình CRFs tuyến tính: đặc trưng dựa trên các cạnh của đồ thị và đặc trưng dựa trên từng trạng thái của đồ thị mà chúng được sinh ra bằng cách kết hợp thông tin xung quanh vị trí hiện hành trong dãy quan sát với nhãn hiện hành.

4. Phương pháp tách từ sử dụng mô hình Pointwise

Một phương pháp tách từ khác tên là Pointwise được đề xuất trong (Luu & Yamamoto, 2012) cho rằng những phương pháp tách từ như HMM, CRFs và MaxEnt có điểm chung là có tham khảo nhãn (hay kết quả) của những nhãn bên cạnh; các phương pháp này chỉ đạt kết quả tốt khi có một từ điển lớn. Với cách tiếp cận của Pointwise, các nhãn sẽ được đánh giá một cách độc lập, không có tham khảo đến kết quả của các nhãn trước đó. Các đặc trưng tại mỗi vị trí từ đơn đang xét nhãn có sử dụng thông tin văn bản (quan sát) xung quanh vị trí đó. Luu và đồng sự sử dụng 3 dạng đặc trưng trong phương pháp Pointwise: n-grams âm tiết (từ đơn), n-grams chủng loại của âm tiết (âm tiết viết hoa, viết thường, số và các loại khác) và đặc trưng từ điển (xét sự xuất hiện của các từ trong từ điển).

Bước sau cùng thực hiện huấn luyện mô hình máy học SVM để phân loại từng vị trí giữa các từ trong câu. Ở đây phương pháp thực hiện phân loại mỗi vị trí thành: vị trí tách từ hay vị trí liên kết từ (tạo thành từ ghép).

C. Tiếp cận lai

Như đã phân tích ở trên, phương pháp tiếp cận từ điển và phương pháp tiếp cận thống kê đều có những ưu và nhược điểm riêng. Để có thể tận dụng được những ưu điểm của mỗi loại tiếp cận, phương pháp tiếp cận lai được đề nghị. Một số phương pháp kết hợp giữa tiếp cận từ điển và tiếp cận thống kê có thể kể đến như: kết hợp giữa mô hình ngôn ngữ Weighted Finite State Transducer (WFST) và mạng Neural (Dinh et al., 2001), kết hợp giữa mô hình so khớp cực đại và máy học véc-tơ hỗ trợ (SVMs) (Dinh et al., 2006), kết hợp mô hình so khớp cực đại và ngôn ngữ mô hình n-grams (Le et al., 2008), hệ thống tách từ tiếng Việt WS4VN kết hợp giữa phương pháp so khớp cực đại và mô hình Markov ẩn (Pham et al., 2009). Le và cộng sự đã đề xuất phương pháp tách từ tiếng Việt dựa trên sự kết hợp giữa phương pháp tiếp cận dựa trên từ điển và phương pháp tiếp cận thống kê (Le et al., 2008).

III. MÔ HÌNH TÚI TỪ VÀ MÁY HỌC VÉC-TƠ HỖ TRỢ

Sau bước tách từ bằng các phương pháp trình bày ở trên, tập dữ liệu văn bản cần được biểu diễn về cấu trúc bảng để từ đó các giải thuật máy học có thể học để phân lớp tự động văn bản. Mô hình túi từ (Salton et al., 1975) là mô hình biểu diễn văn bản phổ biến (Lewis& Gale, 1994), (Dumais et al., 1998), (Sebastiani, 1999), (Manning et al., 2008). Một văn bản được biểu diễn dạng véc-tơ (có *n* thành phần, chiều) mà giá trị thành phần thứ *j* là tần số xuất hiện từ thứ *j* trong văn bản. Nếu xét tập *D* gồm *m* văn bản và từ điển có *n* từ vựng, thì *D* có thể được biểu diễn thành bảng *D* kích thước *m* x *n*, dòng thứ *i* của bảng là véc-tơ biểu diễn văn bản thứ *i* tương ứng.

Bảng 1. Ví dụ tập dữ liệu văn bản

STT	Nội dung	Chủ đề
1	Brazil - đối thủ khắc tinh của Italy	Thể thao
2	Mưa đá dữ dội, rất nhiều nhà dân bị thiệt hại	Xã hội
…	…	…
m	Đột nhập nhà đại gia trộm 2 kg vàng	Pháp luật

Xem ví dụ trong bảng 1 là tập dữ liệu văn bản sau bước tách từ đơn âm, tập dữ liệu văn bản được biểu diễn bằng mô hình túi từ như bảng 2.

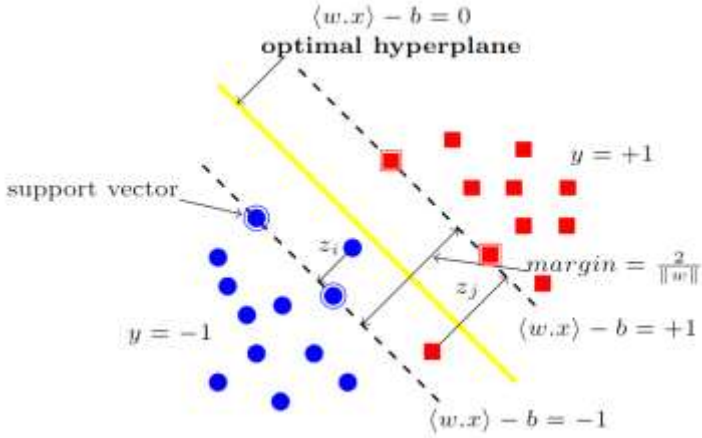
Bảng 2. Tập dữ liệu văn bản được biểu diễn bằng mô hình túi từ

STT	1 (bị)	2 (brazil)	…	n (tinh)	Chủ đề
1	0	1	…	1	Thể thao
2	1	0	…	0	Xã hội
…	…	…	…	…	…
m	0	0	…	0	Pháp luật

Bảng dữ liệu *D* có số chiều (cột) chính bằng số lượng từ vựng. Với tập dữ liệu khoảng vài trăm văn bản, tập từ vựng có thể lên đến hàng chục ngàn từ. Do đó bảng dữ liệu *D* có số cột *n* rất lớn đến vài chục ngàn.

Bước quan trọng tiếp theo là cần huấn luyện mô hình máy học để có thể phân lớp chính xác tập dữ liệu *D* có số chiều lớn. Trong các giải thuật phân lớp (Wu & Kumar, 2009), mô hình máy học véc-tơ hỗ trợ, SVM (Vapnik, 1995) là giải thuật cho độ chính xác cao nhất khi so sánh với các giải thuật máy học khác (Caruana et al., 2008).

Xét ví dụ phân lớp nhị phân tuyến tính đơn giản được mô tả như hình 2, giải thuật máy học SVM tìm siêu phẳng tối ưu để tách dữ liệu ra 2 lớp xa nhất có thể. Máy học SVM tìm siêu phẳng tối ưu dựa trên 2 siêu phẳng hỗ trợ song song của 2 lớp. Siêu phẳng hỗ trợ ($w.x - b = +1$) của lớp *+1* là siêu phẳng mà các phần tử x_p thuộc lớp $y_p = +1$ nằm về phía bên phải của nó. Tương tự, siêu phẳng hỗ trợ ($w.x - b = -1$) của lớp *-1* là siêu phẳng mà các phần tử x_n thuộc lớp $y_n = -1$ nằm về phía bên trái siêu phẳng hỗ trợ lớp *-1*. Những phần tử nằm ngược phía với siêu phẳng hỗ trợ được coi như lỗi, được biểu diễn bởi $z_i \geq 0$. Khoảng cách giữa 2 siêu phẳng hỗ trợ được gọi là lề. Siêu phẳng tối ưu (nằm giữa 2 siêu phẳng hỗ trợ) cần tìm phải thỏa 2 tiêu chí là cực đại hóa lề (lề càng lớn, mô hình phân lớp càng an toàn) và cực tiểu hóa lỗi.

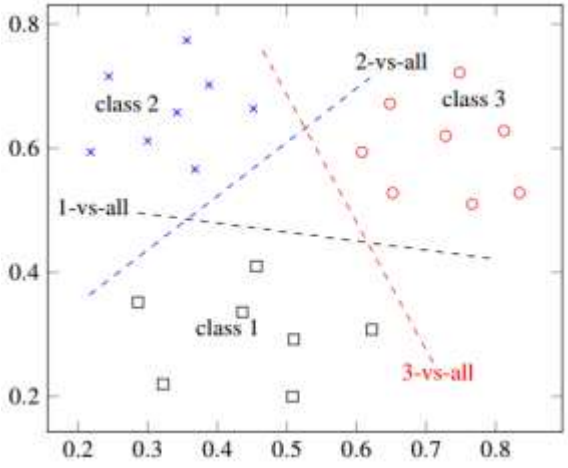


Hình 2. Phân lớp tuyến tính với máy học SVM

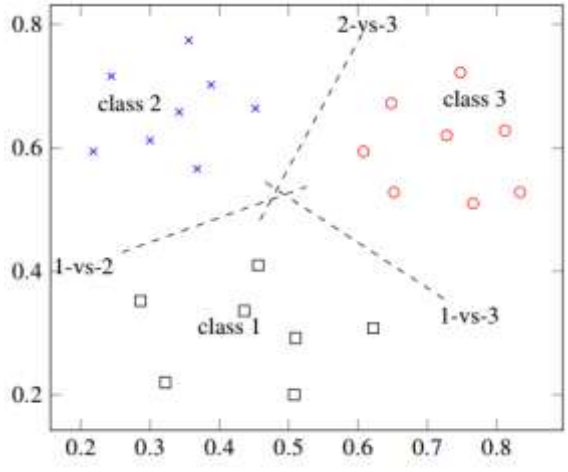
Máy học SVM có thể mở rộng để xử lý bài toán phân lớp *k* lớp (*k > 2* gọi là phân lớp đa lớp vì có số lớp lớn hơn 2). Phương pháp thường được sử dụng trong cài đặt LibSVM đa lớp (Chang & Lin, 2011):

- Phương pháp 1-tất cả, 1-vs-all (Vapnik, 1995): mỗi mô hình phân tách 1 lớp từ các lớp khác, xây dựng *k* mô hình cho *k* lớp (như hình 3),
- Phương pháp 1-1, 1-vs-1 (Kreßel, 1999): mỗi mô hình phân tách 2 lớp, xây dựng *k(k-1)/2* mô hình cho *k* lớp (như hình 4).

Phân lớp phần tử mới *x* dựa vào bình chọn khoảng cách từ *x* đến các siêu phẳng thu được từ các mô hình SVM nhị phân.



Hình 3. Phương pháp 1-tất cả của SVM đa lớp



Hình 4. Phương pháp 1-1 của SVM đa lớp

Mô hình máy học SVM cho kết quả cao, ổn định, chịu đựng nhiễu tốt và phù hợp với các bài toán phân lớp dữ liệu có số chiều lớn. Nghiên cứu của (Dumais et al., 1998) chỉ ra rằng máy học SVM cho hiệu quả cao nhất trong phân lớp tự động văn bản biểu diễn bằng mô hình túi từ. Chính vì lý do đó, chúng tôi sử dụng máy học SVM để phân lớp dữ liệu văn bản.

IV. KẾT QUẢ THỰC NGHIỆM

Chúng tôi tiến hành đánh giá hiệu quả của các phương pháp tách từ tiếng Việt được sử dụng trong phân lớp văn bản tiếng Việt được biểu diễn bằng mô hình túi từ, sử dụng máy học SVM. Chúng tôi tiến hành cài đặt bằng C/C++ các phương pháp:

- tách từ đơn Unigram, viết tắt là **Uni**
- tách từ theo phương pháp so khớp từ dài nhất trong từ điển của (Ho, 1997-2004), viết tắt là **Dic**
- phương pháp tách từ n-grams sử dụng thống kê từ.

Chúng tôi cũng sử dụng thư viện **JvnTextPro** của (Nguyen et al., 2010), thư viện cung cấp phương pháp tách từ tiếng Việt dựa trên trường xác suất có điều kiện (Conditional Random Fields - CRFs) và độ hỗn loạn cực đại (Maximum Entropy - MaxEnt), viết tắt là **Jvn.** Thư viện **vnTokenizer** của (Le et al., 2008) cung cấp phương pháp tách từ tiếng Việt dựa trên kỹ thuật lai (từ điển, automat hữu hạn trạng thái, biểu thức chính quy và so khớp từ dài nhất), viết

tắt là **vnTok**. Nhóm tác giả (Luu & Yamamoto, 2012) đề xuất phương pháp tách từ với n-grams, từ điển, máy học SVM trong thư viện **DongDu**. Thư viện LibSVM (Chang & Lin, 2011) cung cấp giải thuật máy học SVM đa lớp sử dụng phương pháp 1-1.

Tất cả các thí nghiệm được chạy trên máy tính cá nhân, cài hệ điều hành Linux Fedora 20, bộ vi xử lý Intel® Core i7-4790, 3.6 GHz, 4 nhân và bộ nhớ RAM 8 GB.

A. Chuẩn bị tập dữ liệu

Chúng tôi sử dụng 2 tập dữ liệu văn bản tiếng Việt để đánh giá sự ảnh hưởng của các phương pháp tách từ tiếng Việt trong phân lớp tự động văn bản tiếng Việt. Tập dữ liệu **vnexpress** là tập dữ liệu văn bản thu thập từ trang báo điện tử **vnexpress.net**, gồm có 10 chủ đề (10 lớp) bao gồm công nghệ thông tin, giải trí, giáo dục, kinh doanh, ẩm thực, pháp luật, y tế, thế giới, thể thao, tình yêu. Mỗi chủ đề chúng tôi thu thập khoảng 600 bản tin văn bản khác nhau tạo thành tập dữ liệu văn bản có 6000 bản tin. Vấn đề đặt ra là cần huấn luyện mô hình phân lớp từ tập dữ liệu **vnexpress**, để có thể phân lớp tự động một bản tin vào một trong 10 chủ đề. Sau bước tách từ, chúng tôi thu được các tập từ vựng tương ứng của mỗi phương pháp như trình bày trong bảng 3. Biểu diễn tập **vnexpress** bằng mô hình túi từ (Salton et al., 1975), chúng tôi thu được 6 bảng dữ liệu tương ứng với 6 phương pháp tách từ, mỗi bảng có 6000 dòng và số cột (chiều) là tổng số từ vựng thu được từ 6 phương pháp và 10 lớp.

Bảng 3. Tập dữ liệu văn bản vnexpress

Phương pháp tách từ	Tổng số từ vựng	Tổng số văn bản	Tổng số chủ đề
Unigram (từ đơn)	24214	6000	10
JVnTextPro (CRF, MaxEnt)	63827	6000	10
vnTokenizer (hybrid approach)	51018	6000	10
DongDu (Pointwise)	58811	6000	10
Dictionary (Longest matching)	34775	6000	10
n-grams (Statistical approach)	34746	6000	10

Tập dữ liệu **book collection** là tập dữ liệu văn bản thu được từ Trung tâm học liệu, Trường Đại học Cần Thơ. Tập dữ liệu có 105293 quyển sách, mỗi quyển sách được mô tả bởi tựa đề, từ khóa, tóm tắt và mã loại. Tập dữ liệu **book collection** rất phức tạp, có số lượng sách nhiều, mô tả mỗi quyển sách rất ít thông tin (khoảng 20 từ), tổng số lớp là 166. Vấn đề đặt ra là cần huấn luyện mô hình phân lớp từ tập dữ liệu **book collection**, để có thể phân lớp tự động một cuốn sách vào một trong 166 mã loại. Các phương pháp tách từ cho ra các tập từ vựng như trình bày trong bảng 4. Biểu diễn tập **book collection** bằng mô hình túi từ, chúng tôi thu được 6 bảng dữ liệu tương ứng với 6 phương pháp tách từ, mỗi bảng có 105293 dòng và số cột (chiều) là tổng số từ vựng thu được từ 6 phương pháp và 166 lớp.

Bảng 4. Tập dữ liệu văn bản book collection

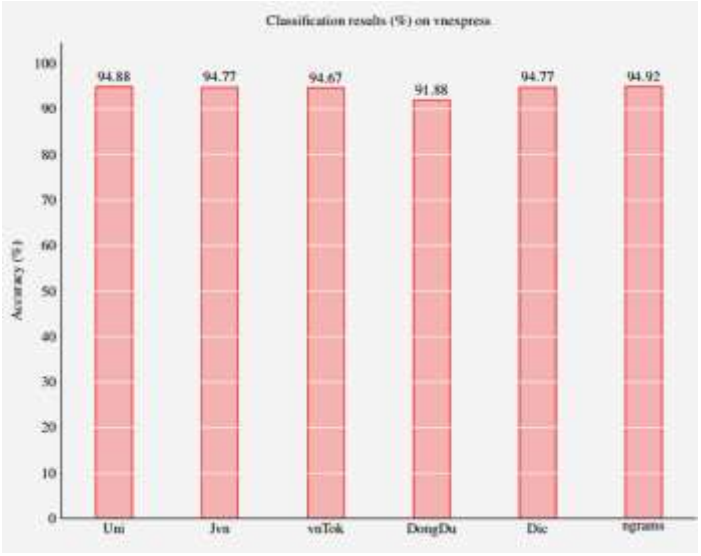
Phương pháp tách từ	Tổng số từ vựng	Tổng số sách	Tổng số loại
Unigram (từ đơn)	59263	105293	166
JVnTextPro (CRF, MaxEnt)	83061	105293	166
vnTokenizer (hybrid approach)	89595	105293	166
DongDu (Pointwise)	121589	105293	166
Dictionary (Longest matching)	68224	105293	166
n-grams (Statistical approach)	119864	105293	166

B. Kết quả thực nghiệm

Do các bảng dữ liệu thu được từ biểu diễn bằng mô hình túi từ có số cột (chiều) lên đến vài chục ngàn, chỉ cần huấn luyện mô hình máy học SVM sử dụng hàm nhân tuyến tính là có thể phân lớp chính xác các bảng dữ liệu có số chiều lớn (Dumais et al., 1998), (Sebastiani, 1999).

Chúng tôi sử dụng nghi thức kiểm tra chéo 3-fold để đánh giá kết quả phân lớp. Tập dữ liệu được xáo trộn ngẫu nhiên và chia thành 3 phần bằng nhau; ở mỗi lần thực nghiệm lấy 1 phần làm tập kiểm tra và 2 phần còn lại làm tập huấn luyện; dùng tập huấn luyện để xây dựng mô hình phân lớp SVM, tiếp đến là dùng mô hình SVM thu được để phân lớp tập kiểm tra thu được độ chính xác; ở lần thực nghiệm tiếp theo sử dụng 1 phần khác làm tập kiểm tra, 2 phần còn lại làm tập huấn luyện và thực hiện lặp lại các bước xây dựng mô hình, phân lớp tập kiểm tra; đến lần thực nghiệm thứ 3 thì kết thúc. Kết quả phân lớp là trung bình cộng của cả 3 lần thực nghiệm trên.

Hình 5 trình bày kết quả phân lớp trên tập dữ liệu **vnexpress**sử dụng 6 phương pháp tách từ **Uni**, **Jvn**, **vnTok**, **DongDu**, **Dic**, **n-grams** tương ứng. Kết quả cho thấy rằng phương pháp tách từ **n-grams** và **Uni** trên tập dữ liệu **vnexpress** được sử dụng trong phân lớp văn bản tiếng Việt cho độ chính xác cao nhất tương ứng là **94.92%** và **94.88%**. Trong khi đó mô hình phân lớp văn bản **vnexpress** sử dụng phương pháp tách từ **DongDu** cho độ chính xác thấp nhất tương ứng là **91.88%.**



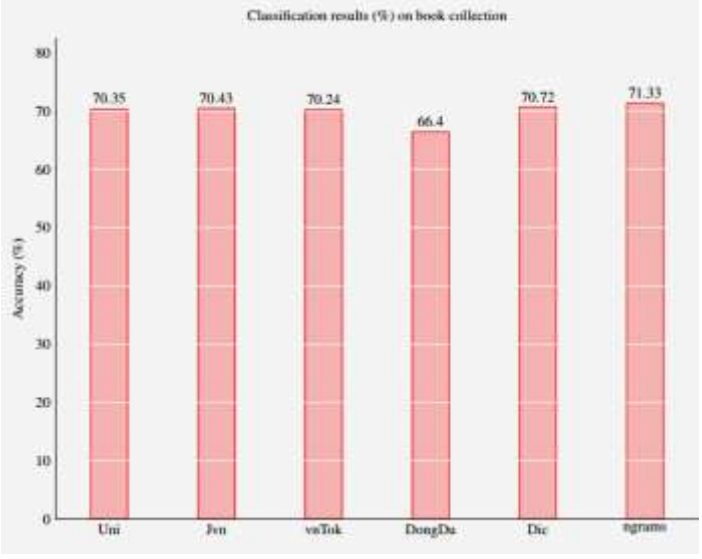
Hình 5. Kết quả phân lớp tập dữ liệu vnexpress

Bảng 5. Kiểm định Wilcoxon từng cặp phương pháp tách từ trên tập văn bản vnexpress

p-value	Jvn	vnTok	DongDu	Dic	n-grams
Uni	0.6442302	0.3342981	0.4817821	0.9580880	0.3460977
Jvn		0.1323545	0.2241462	0.6112262	0.1268114
vnTok			0.8983922	0.3936825	0.9552262
DongDu				0.4724780	0.9162778
Dic					0.3393654

Chúng tôi thực hiện kiểm định thống kê để kiểm chứng có sự khác biệt thật sự giữa các phương pháp tách từ tiếng Việt được sử dụng trong phân lớp văn bản tiếng Việt. Kết quả kiểm định Wilcoxon cho từng cặp phương pháp, thu được các giá trị **p** như trong bảng 5. Các giá trị **p** lớn hơn **0.05** cho thấy rằng sử dụng các phương pháp tách từ tiếng Việt trong phân lớp văn bản tiếng Việt thu được kết quả khác biệt không có ý nghĩa thống kê.

Tương tự với tập dữ liệu **book collection**, chúng tôi thu được kết quả phân lớp sử dụng 6 phương pháp tách từ tiếng Việt như trình bày trong hình 6. Kết quả cho thấy rằng sử dụng phương pháp tách từ **n-grams** và **Dic** cho tập dữ liệu **book collection** trong phân lớp sách tiếng Việt cho độ chính xác cao nhất tương ứng là **71.33%** và **70.72%**. Một lần nữa, mô hình phân lớp sách **book collection** sử dụng phương pháp tách từ **DongDu** cho độ chính xác thấp nhất tương ứng là **66.40%**.



Hình 6. Kết quả phân lớp tập dữ liệu book collection

Kết quả kiểm định Wilcoxon cho từng cặp phương pháp, thu được các giá trị **p** lớn hơn **0.05** như trong bảng 6, một lần nữa cho thấy rằng sử dụng các phương pháp tách từ tiếng Việt trong phân lớp văn bản tiếng Việt thu được kết quả khác biệt không có ý nghĩa thống kê.

Bảng 6. Kiểm định Wilcoxon từng cặp phương pháp tách từ trên tập book collection

p-value	Jvn	vnTok	DongDu	Dic	n-grams
Uni	0.3684242	0.3432044	0.3370766	0.3656872	0.13760121
Jvn		0.9618614	0.9860077	0.9955413	0.11892901
vnTok			0.9641109	0.9666230	0.13006150
DongDu				0.9790262	0.12772106
Dic					0.11832005

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày một nghiên cứu so sánh về sự ảnh hưởng của các phương pháp tách từ tiếng Việt đối với hiệu quả phân lớp trong bài toán phân loại văn bản tiếng Việt. Các kết quả thực nghiệm trên tập dữ liệu văn bản gồm 6000 văn bản thuộc 10 chủ đề của trang báo điện tử vnexpress.net và tập dữ liệu sách với 166 chủ đề cho thấy rằng việc tách từ đa âm tiết với các tiếp cận khác nhau và tách từ đơn âm tiết hoàn toàn không có ảnh hưởng hoặc ảnh hưởng không đáng kể đối với hiệu quả phân lớp. Một điều cần chú ý là đối với phương pháp tách từ dựa trên điểm tách (phương pháp DongDu) hoàn toàn dựa trên thông tin về sự xuất hiện cạnh nhau (collocation) của các từ có thể tạo ra các từ ghép mới (không phải là từ ghép). Điều này (i) làm cho số lượng từ vựng tăng lên và (ii) nghiêm trọng hơn là làm mất thông tin về sự xuất hiện của các từ có trong từ ghép mới này[2]. Lúc này, để đảm bảo giữ được hiệu quả phân lớp cần phải có số lượng lớn mẫu huấn luyện (hiện tượng được biết đến với tên gọi *curse of dimensionality*).Với kết quả như thế, ta hoàn toàn có thể sử dụng phương pháp tách từ đơn âm tiết (dựa trên khoảng trắng như tiếng Anh) hoặc tách từ theo phương pháp so khớp từ dài nhất trong từ điển vào bài toán phân loại văn bản tiếng Việt để tăng tốc độ xử lý trong khi vẫn giữ được hiệu quả phân lớp cao.

Chúng tôi tiếp tục thực hiện so sánh sự ảnh hưởng của tách từ với các giải thuật máy học khác như multinominal naive Bayes, cây quyết định, rừng ngẫu nhiên và với nhiều nhiều tập dữ liệu tiếng Việt khác nữa. Ngoài ra, tiếp cận này hoàn toàn có thể áp dụng lên các ngôn ngữ châu Á khác như tiếng Trung Quốc, tiếng Nhật hay tiếng Hàn. Chúng tôi dự định thực hiện điều này trong các nghiên cứu sắp tới.

TÀI LIỆU THAM KHẢO

[1] J-P. Benzécri, "*L'analyse des correspondances*", Paris:Dunod, 1973.

[2] L. Breiman, J.H. Friedman, R.A. Olshen and C. Stone, "*Classification and Regression Trees*", Wadsworth International, 1984.

[3] L. Breiman, "Arcing classifiers",*The annals of statistics* 26(3):801-849, 1998.

[4] L. Breiman, "Random forests",*Machine Learning* 45(1):5-32, 2001.

[5] C-C. Chang, and C-J. Lin, "LIBSVM: a library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27, pp.1-27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] R. Caruana, N. Karampatziakis, A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions", in proc. of the 25th intl conf. on Machine learning, pp. 96-103, 2008.

[7] Jan Daciuk, Stoyan Mihov, Bruce W. Watsonand Richard E.Watson, "Incremental Construction of Minimal Acyclic Finite-State Automata", *Computational Linguistics*, Vol. 26, No. 1, 2000.

[8] D. Dinh, K. Hoang, V-T. Nguyen, "Vietnamese Word Segmentation", The 6thNatural Language Processing Pacific Rim Symposium, pp.749-756, 2001.

[9] D. Dinh, D. Vu, N.L. Nguyen, "Application of Maximum matching and SVMs for Vietnamese word segmentation", ICT.rda'06, Đà Lạt, 2006.

[10] T-N. Đỗ,"Phân loại thư rác với giải thuật ARCX4-RMNB", Kỷ yếu hội nghị @CNTT, pp. 427-437, 2012.

[11] T-N. Đỗ, N-K. Phạm,"Phân loại văn bản: Mô hình túi từ và tập hợp mô hình máy học tự động",*Tạp chí khoa học ĐHCT*, Số 28: 9-16, 2013.

[12] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization",inproc. of ACM-CIKM98, pp. 148-155, 1998.

[13] E. Fix, and J. Hodges, "Discriminatoiry Analysis: Small Sample Performance",*Technical Report 21-49-004*, USAF School of Aviation Medicine, Randolph Field, 1952.

[14] Y. Freund, and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", In proc. of *Computational Learning Theory*, pp. 23-37, 1995.

[15] I. Good, "The Estimation of Probabilities: An Essay on Modern Bayesian Methods",*MIT Press*, 1965.

[16] N-D. Ho, "The Free Vietnamese Dictionary Project", 1997-2004. http://www.informatik.uni-leipzig.de/~duc/Dict

[17] F. Jelinek, R.L. Mercer and S. Roukos,"Principles of Lexical Language Modeling for Speech Recognition", *Advances in Speech Signal Processing*, S. Furui and J. Sondhi, Eds. M. Dekker Publishers, New York, pp.651-700, 1991.

[18] T. Joachims, "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features", in proc. of ECML '98, pp. 137-142, 1998.

[2]Khi ghép các âm tiết để tạo nên từ ghép, ta chỉ giữ lại từ ghép sau cùng và bỏ qua tất cả các âm tiết có trong từ ghép. Vì thế nếu ghép sai, ta mất đi thông tin về sự xuất hiện của từ/âm tiết có trong *từ ghép sai*.

[19] U. Kreßel, "Pairwise classification and support vector machines",*Advances in Kernel Methods: Support Vector Learning*, pp. 255-268, 1999.

[20] H-P. Le, T-M-H., Nguyen, A. Roussanaly, and T V. Ho, "A hybrid approach to word segmentation of Vietnamese texts", in proc. of the 2nd Intl Conf. on Language and Automata Theory and Applications, Spain, Springer, pp. 240-249, 2008.http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer.

[21] D. Lewis, andW. Gale, "A sequential algorithm for training text classifiers", in proc. of the 17th annual intl ACM SIGIR conf. on Research and development in information retrieval, pp.3-12, 1994.

[22] X. Luo and S. Roukos, "An iterative algorithm to build Chinese language models", In Proceedings of the 34th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 139-143, 1996.

[23] T-A. Luu and K. Yamamoto, "Ứng dụng phương pháp Pointwise vào bài toán tách từ cho tiếng Việt", NLP Lab., Dept. of Electrical Engineering, Nagaoka University of Technology, 2012. http://viet.jnlp.org/dongdu.

[24] C. Manning, P. Raghavan, andH. Schütze, "*Introduction to Information Retrieval*", Cambridge University Press, 2008.

[25] A. McCallum, and K. Nigam, "A comparison of event models for Naive Bayes text classification", In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48, 1998.

[26] Surapant Meknavin, Paisarn Charoenpornsawat, and Boonserm Kijsirikul, "Feature-based Thai Word Segmentation", in proc. of the Natural Language Processing Pacific Rim Symposium (NLPRS'97), Phuket, Thailand, 1997.

[27] C-T. Nguyen, X-H. Phan, and T-T. Nguyen, "JVnTextPro: A Java-based Vietnamese Text Processing Tool", 2010. http://jvntextpro.sourceforge.net

[28] D-D. Pham, G-B. Tran, S-B. Pham, "A hybrid approach to Vietnamese word segmentation using part of speech tags",in proc. of intl conf on Knowledge and Systems Engineering, pp. 154-161, 2009.

[29] N-K. Phạm, T-N. Đỗ, và C-Đ. Trần,"Phân Loại Dữ Liệu với Giải Thuật Arcx4-LSSVM", Kỷ yếu hội nghị ICTFIT, HCM, pp. 72-78, 2008.

[30] N-K. Phạm, T-N. Đỗ, và F. Poulet,"Phân loại văn bản với BPSVM", Kỷ yếu hội nghị @CNTT, pp. 269-278, 2006.

[31] J-R. Quinlan, "*C4.5: Programs for Machine Learning*", Morgan Kaufmann, San Mateo, 1993.

[32] G. Salton, A. Wong, and C-S. Yang, "A vector space model for automatic indexing",*Communications of the ACM*, vol.18(11):613-620,1975.

[33] F. Sebastiani, "Machine learning in automated text categorization",*ACM Computing Surveys*vol.34(1):1-47, 1999.

[34] Chih-Hao Tsai, "MMSEG: A Word Identification System for Mandarin Chinese TextBased on Two Variants of the Maximum Matching Algorithm.", 1996.http://technology.chtsai.org/MMSEG/.

[35] V. Vapnik, "*The Nature of Statistical Learning Theory*", Springer-Verlag, 1995.

[36] X. Wu, and V. Kumar,"*Top 10 Algorithms in Data Mining*", Chapman & Hall/CRC, 2009.

A COMPARISON OF WORD SEGMENTATION METHODS IN VIETNAMESE TEXT CATEGORIZATION

Pham Nguyen Khang, Tran Nguyen Minh Thu, Pham The Phi, Do Thanh Nghi

ABSTRACT— *Word segmentation, which determines the boundaries of words in a text document, is an important step in natural language processing. In Vietnamese, besides one-syllable words, there are also words with multiple syllables. Hence, the approach of separating words simply using the white space is believed to be not effective. Many approaches to segmenting words in written Vietnamese (dictionary-based, statistical-based or combination of both) are proposed competing for accuracy. It is common sense that good word segmentation results will contribute to better language processing and understanding works, e.g. text clustering, text classification, part-of-speech tagging, semantic role labeling, machine translation, and so on. But is that really so for the task of Vietnamese text classification? In this paper, we present a comparative study of the effect of various word segmentation methods to Vietnamese text classification. The experiments are conducted on two datasets: (i) 6000 texts of 10 topics and (ii) 105293 book abstracts of 166 topics with the SVM classification model. We discover that the classification accuracies with different word segmentation methods are not statistically different.*