

# SO SÁNH CÁC PHẦN MỀM TÁCH TỪ TIẾNG VIỆT

## ON COMPARING VIETNAMESE WORD SEGMENTATION SOFTWARES

*Phan Tấn Phát<sup>1</sup>, Quách Luyt Đa<sup>2</sup>, Dương Trung Nghĩa<sup>3</sup>*

<sup>1</sup>*Đại học FPT Cần Thơ; [phatptce130319@fpt.edu.vn](mailto:phatptce130319@fpt.edu.vn)*

<sup>2</sup>*Đại học FPT Cần Thơ; [luyldaquach@gmail.com](mailto:luyldaquach@gmail.com)*

<sup>3</sup>*Đại học FPT Cần Thơ; trường đại học Kỹ thuật – Công nghệ Cần Thơ; [duong-trung@ismll.de](mailto:duong-trung@ismll.de)*

### Tóm tắt

Bài viết này nghiên cứu các phần mềm xử lý tiếng Việt trong việc thực hiện tác vụ tách từ. Việc xác định đầu là một từ giúp máy tính rút trích các đơn vị đặc trưng có ý nghĩa đối với văn bản. Đây là tác vụ tiền xử lý quan trọng, đảm bảo thông tin đặc trưng được rút trích, tránh nhiễu, giảm chiều không gian dữ liệu, làm tăng độ chính xác và giảm thời gian thực thi của các thuật toán xử lý ngôn ngữ tự nhiên. Từ đó hiểu các bộ công cụ để áp dụng cho việc phân tích, xử lý dữ liệu phục vụ cho các ứng dụng như nhận diện giọng nói, tự động sửa lỗi chính tả và ngữ pháp, tóm tắt văn bản. Nghiên cứu được thực hiện bằng cách áp dụng các phần mềm phân tích từ tiếng Việt lên các tập dữ liệu có kích thước khác nhau. Sau đó, ghi nhận tốc độ thực thi và độ chính xác. Kết quả của nghiên cứu cho thấy thuật toán kết hợp giữa từ điển và n-gram mang lại độ chính xác cao với thời gian thực thi ngắn.

**Từ khóa:** Xử lý ngôn ngữ tự nhiên; cấu trúc; tách từ; tiếng Việt; phần mềm tách từ tiếng Việt.

### 1. Đặt vấn đề

Xử lý ngôn ngữ tự nhiên là một lĩnh vực quan trọng trong nghiên cứu máy học, trí tuệ nhân tạo, kỹ thuật thông tin liên quan đến sự tương tác giữa máy tính và ngôn ngữ con người[1]. Bài toán tách từ là một phần trong xử lý ngôn ngữ tự nhiên.

Tách từ là tác vụ xác định ranh giới các từ trong văn bản, đảm bảo tính nguyên tử trong ý nghĩa. Việc xác định được đầu là một từ giúp máy tính rút trích được các đơn vị đặc trưng có ý nghĩa đối với văn bản. Đây là tác vụ tiền xử lý quan trọng, đảm bảo thông tin đặc trưng được rút trích, tránh nhiễu, giảm chiều không gian dữ liệu, từ đó làm tăng độ chính xác và giảm thời gian thực thi của các thuật toán xử lý ngôn ngữ tự nhiên.

Đối với các ngôn ngữ đa âm đơn nghĩa như tiếng Anh, Pháp việc phân tách từ một câu khá là đơn giản vì mỗi từ được cách nhau tự nhiên bởi ký tự khoảng trắng. Tuy nhiên, đối với các ngôn ngữ tượng âm và tượng hình thuộc vùng Á châu như tiếng Việt, tiếng Trung thì tác vụ tách từ lại không đơn giản. Khoảng trắng trong các ngôn ngữ này vừa dùng để tách từ, vừa dùng để tách tiếng. Trong các loại ngôn ngữ đơn âm tiết như tiếng Việt, một từ có thể được cấu tạo bởi nhiều tiếng. Từ xác định theo tiếng có một ý nghĩa và từ xác định theo tổ hợp các tiếng lại có ý nghĩa hoàn toàn khác. Vì vậy điểm khó khăn trong việc tách từ tiếng Việt là xem một tổ hợp các tiếng là một từ ghép hay nhiều từ đơn lẻ. Nếu tìm hiểu ý nghĩa của một câu tiếng Việt bằng cách dịch nghĩa từng từ trong câu thì sẽ không chính xác. Ví dụ, đối với từ “xanh ngắt” nếu ta dịch riêng

### Abstract

This research focuses on several Vietnamese text processing softwares regarding of the task of word segmentation. Identifying word boundaries helps machines extract crucial characteristics given a chunk of continuous text. Segmentation is a significant task within text pre-processing procedure ensuring information extraction, noise reduction, dimensionality reduction, increase in accuracy and decrease in computational time. By comprehending how these softwares work, one can apply them to natural language processing tasks, text analytics, voice recognition, automatic grammar correction, text summarization. The research is conducted by applying the softwares upon numerous Vietnamese text sources with different sizes; observing time of execution and accuracy. The results show that algorithm combining dictionary and n-gram gains the most accuracy and time efficiency.

**Key words:** Natural language processing; syntax; word segmentation; Vietnamese; Vietnamese word segmentation softwares.

từ “xanh” và từ “ngắt” thì ý nghĩa của từ sẽ hoàn toàn khác so với từ gốc.

Để xử lý vấn đề trên thì việc sử dụng các công cụ xử lý ngôn ngữ tự nhiên nói chung hoặc sử dụng các công cụ tách từ trong tiếng Việt nói riêng một cách hiệu quả là rất cần thiết. Các công cụ tách từ mã nguồn mở được sử dụng trong nghiên cứu là VnTokenizer với thuật toán sử dụng từ điển kết hợp với n-gram, DongDu với phương pháp Pointwise, JvnSegmenter với việc sử dụng kết hợp trường điều kiện ngẫu nhiên (Conditional Random Fields - CRFs) và máy vector hỗ trợ (Support Vector Machines - SVMs) và UETSegmenter sử dụng thuật toán so khớp từ dài nhất kết hợp với hồi quy logistic.

Trong quá trình thực hiện tách từ, kho ngữ liệu là quan trọng nhất. Trên thế giới hiện nay có nhiều kho ngữ liệu khác nhau, tồn tại dưới nhiều dạng khác nhau, cấu trúc và định dạng của kho ngữ liệu rất đa dạng, phải kể đến các kho ngữ liệu của Anh là ICE[2], kho ngữ liệu của Hoa Kỳ là ANC[3],... Ở Việt Nam, kho ngữ liệu cũng được quan tâm và xây dựng như: kho ngữ liệu Sketch của tác giả Phan Thị Hà (94 triệu từ) [4]; VietTreebank của Nguyễn Phương Thái và các cộng sự [5], Kho ngữ liệu dự án VLSP của Lê Thanh Hương[6]. Cả hai kho ngữ liệu VietTreebank và kho ngữ liệu dự án VLSP đều thuộc dự án KC01.01/2006-2010 của tác giả Hồ Tú Bảo[7].

Để đảm bảo tính độc lập của dữ liệu, nghiên cứu thực hiện xây dựng bộ dữ liệu khoảng 10000 câu và được chia làm 5 tập dữ liệu với kích thước câu khác nhau. Nghiên cứu thực hiện kiểm tra độ chính xác và tốc độ thực thi, từ đó

đưa ra những đánh giá về kết quả thực hiện tách từ trong xử lý ngôn ngữ tự nhiên tiếng Việt đối với các phần mềm nghiên cứu được đề cập.

Đối chiếu với các nghiên cứu đã được thực hiện trước đây, với sự hiểu biết tốt nhất, bài viết này có một số đóng góp như sau:

1) Nghiên cứu thực nghiệm các phần mềm tách từ trên bộ dữ liệu độc lập do nhóm nghiên cứu tự xây dựng. Bộ dữ liệu sẽ được công bố cùng với bài viết, phục vụ cho các nghiên cứu về sau.

2) So sánh cùng lúc độ chính xác và thời gian thực hiện của các phần mềm trên bộ dữ liệu độc lập mà chưa có nghiên cứu nào thực hiện.

3) Đưa ra được những lựa chọn trong việc sử dụng phần mềm hướng thời gian hay hướng độ chính xác khi thực hiện tác vụ tách từ.

## 2. Giải quyết vấn đề

Để thực hiện giải quyết vấn đề, nghiên cứu thực hiện nghiên cứu các phần mềm tách từ tiếng Việt đã được giới thiệu với những phần mềm như sau:

### 2.1. Phần mềm VnTokenizer

Phần mềm VnTokenizer[8] được training với kho giữ liệu Vietlex với 40,181 từ được tạo nên từ 7,729 âm tiết khác nhau.

VnTokenizer sử dụng bộ quy tắc phân đoạn được quy định tại bộ chuẩn ISO/TC37/SC4[9] để xác định tính chính xác của công cụ tách từ. Các tiêu chí đó là: từ ghép, từ láy, từ ngữ gồm nhiều từ, tên riêng, các mẫu tự thường gặp.

Dữ liệu khi đưa vào chương trình, sẽ được phân tích bằng cách tìm các danh từ riêng, thời gian, ngày tháng, tên viết tắt, thư điện tử. Sau đó, sẽ được phân tích bằng cách tìm các từ ghép, từ láy hoặc từ ghép gồm nhiều từ đơn thông qua thuật toán tham ăn, nếu một nhóm từ thỏa mãn một mô hình có trong mẫu huấn luyện thì nhóm từ đó sẽ được gom lại thành một cụm. Để tăng độ chính xác, thì khi tìm ra được một cụm từ phù hợp với mẫu đã huấn luyện, thì những từ này sẽ được lưu lại và sau đó chọn lại từ phù hợp nhất.

Thuật toán biểu diễn các chuỗi đầu vào như một đồ thị có hướng tuyến tính  $G = (V, E)$  với  $V = \{v_0, v_1, \dots, v_n, v_{n+1}\}$  biểu diễn vị trí của từng âm tiết trong câu. Sẽ có một cặp  $(v_i, v_j)$  sao cho các âm tiết liên tiếp được đánh dấu từ  $s_{i+1}, s_{i+2}, \dots, s_j$  tạo nên một từ, khi  $i < j$ . Ta gọi hàm  $\text{accept}(A, s)$  – hàm này có chức năng lấy ra những cụm từ thoãn mãn điều kiện. Tuy vậy vẫn có khả năng thuật toán chọn những cụm từ không rõ nghĩa, ví dụ khi trong câu tồn tại các cụm từ đều có khả năng kết hợp với nhau.

Để giải quyết, các cụm từ không rõ nghĩa. Thuật toán n-gram, dùng để tính xác suất xuất hiện của một cụm từ trong ngôn ngữ, được áp dụng trong công cụ để xác định xem từ đang được xét có phù hợp hay không, dựa vào tần suất xuất hiện của nó.

$$P(s) = \prod_{i=1}^m P(w_i | w_i^{i-1}) \approx \prod_{i=1}^m P(w_i | w_{i-n+1}^{i-1}) \quad (1)$$

Trong công thức (1) ta có  $s$  là chuỗi đầu vào với  $s = w_1, w_2, \dots, w_m$ . Các  $w_i^j$  biểu thị các từ  $w_i \dots w_j$ . Giá trị  $n$  thường là 2 hoặc 3 tương đương với mô hình bi-gram hoặc tri-gram [10]. Sau khi xác suất xuất hiện của một từ được tính, thì công cụ sẽ dựa vào xác suất đó để chọn từ phù hợp để nhóm lại.

Thuật toán tạo một đồ thị được trình bày như sau:

---

```

V ← ∅;
for i = 0 to n + 1 do
    V ← V ∪ {vi};
end for
for i = 0 to n do
    for j = i to n do
        if (accept(A, si ... sj)) then
            E ← E ∪ {(vi, vj+1)};
        end if
    end for
end for
return G = (V, E);

```

---

### 2.2. Phần mềm JvnSegmenter

Phần mềm JvnSegmenter sử dụng dữ liệu huấn luyện có 2000 tên người, 707 địa danh cũng với 7800 câu thuộc các chủ đề xã hội khác nhau[11]. Khác với các công cụ trước, phần mềm này thực hiện gán nhãn các âm tiết thành các nhóm BW (Beginning of word), IW (Inside of word) và O (Others).

Phần mềm JvnSegmenter sử dụng mô hình chuỗi tuyến tính Conditional Random Fields (CRFs)[12] được dùng để dự đoán nhãn dựa trên chuỗi đầu vào. Sau khi được huấn luyện với kho ngữ liệu gồm các cụm từ tiếng Việt đã được tách. Thuật toán sẽ hoạt động bằng cách làm tăng tối đa sự tương đồng của dữ liệu kiểm tra hoặc đầu vào so khớp với mô hình đã được huấn luyện từ trước và đặt nhãn các từ trong từng chuỗi trong dữ liệu.

Với  $o = (o_1, o_2, \dots, o_T)$  là chuỗi dữ liệu quan sát được và  $S$  là tập hợp các máy trạng thái hữu hạn, mỗi trạng thái liên kết với nhãn  $l \in L$ . Gọi  $s(s_1, s_2, \dots, s_T)$  là một chuỗi trạng thái, để xác định mối liên hệ của chuỗi dữ liệu quan sát với chuỗi trạng thái cho một chuỗi quan sát sử dụng thuật toán CRFs[13] để xác định dựa trên xác suất.

$$p_\theta(s|o) = \frac{1}{Z(o)} \exp \left[ \sum_{t=1}^T \sum_k \lambda_k f_k(s_{t-1}, s_t, o, t) \right] \quad (2)$$

Trong công thức (2) ta có:

$$Z(o) = \sum_{s'} \exp \left( \sum_{t=1}^T \sum_k \lambda_k f_k(s'_{t-1}, s'_t, o, t) \right) \quad (3)$$

Công thức (3) là hệ số chuẩn hoá trên tất cả các chuỗi có thể nhận diện. Tương ứng với mỗi đặc trưng  $f_k$  có trọng số là  $\lambda_k$ , đây là mục đích chính của CRFs. Đặc trưng  $f_k$  được xem xét ở đặc trưng trạng thái (4) và đặc trưng chuyển tiếp (5).

$$f_k^{(per-state)}(s_t, o, t) = \delta(s_t, l)x_k(o, t) \quad (4)$$

$$f_k^{(transition)}(s_{t-1}, s_t, t) = \delta(s_{t-1}, l')\delta(s_t, l) \quad 5$$

Bên cạnh đó, thuật toán máy vector hỗ trợ (SVMs) được dùng để phân loại các âm tiết và phân loại nhãn vào một trong ba nhóm BW, IW hoặc O. Sau khi phân loại các từ thì công cụ sẽ nhóm các âm tiết thuộc BW và IW kế tiếp nhau thành một từ hoàn chỉnh.

Công cụ bước đầu sẽ duyệt qua các từ xem có thuộc vào tập tên người, địa danh, chữ viết tắt để xác định đó là một cụm từ. Sau đó áp dụng thuật toán mô hình CRFs cùng với SVMs để chọn các cụm từ có khả năng và đem so sánh với dữ liệu có trong từ điển.

### 2.3. Phần mềm UETSegmenter

Phần mềm UETSegmenter sử dụng hướng tách từ bằng cách đánh giá khoảng trắng ở giữa các từ là dùng để tách âm tiết hay dùng để tách từ. Với cách tiếp cận này, ta có thể thay đổi loại (tách âm tiết hay tách từ) của khoảng trắng mà không làm ảnh hưởng đến những khoảng trắng kế nó [14].

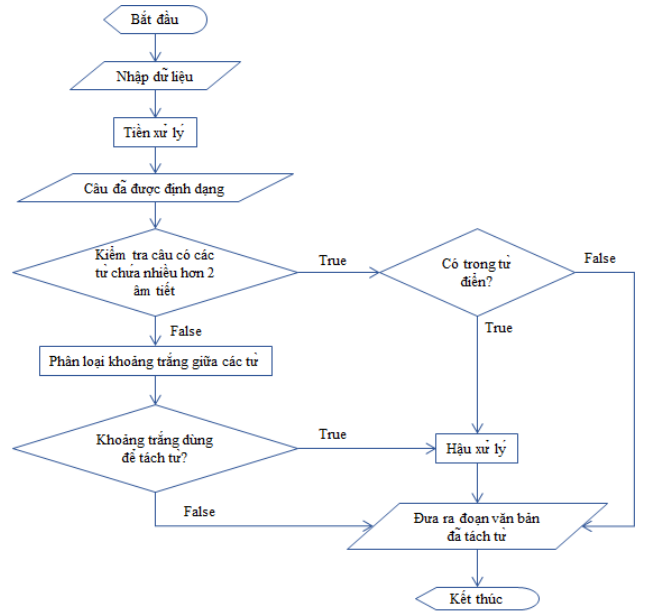
Vì hầu hết các từ trong tiếng Việt đều gồm một hoặc hai âm tiết (71%), nên đối với các từ nhiều hơn hai âm tiết ta có thể lưu chúng vào trong từ điển. Ta có thể áp dụng thuật toán so khớp từ dài nhất để tìm ra các cụm từ có nhiều âm tiết và so sánh chúng với từ có trong từ điển. Vì vậy ta chỉ cần xử lý các từ có hai âm tiết và các danh từ riêng.

Để giải quyết các từ có hai âm tiết hoặc các danh từ riêng, ta sử dụng thuật toán hồi quy logistic để phân loại các khoảng trắng. Sau khi phân loại được đâu là khoảng trắng để tách từ và đâu là khoảng trắng tách âm tiết, công cụ thay thế các khoảng trắng đó bằng dấu gạch dưới ‘\_’. Để tăng thêm độ chính xác, ta lại so sánh những từ ghép gồm ba âm tiết do thuật toán hồi quy logistic đưa ra, và xét trong từ điển. Nếu không có trong từ điển, ta lại tiếp tục tách ra và gọi thuật toán hồi quy để đưa ra các kết quả có khả năng khác. Tóm tắt hoạt động của phần mềm UETSegmenter dựa trên thuật toán hồi quy logistic được mô tả trong (Hình 1).

Hồi quy logistic dùng để phân loại khoảng trắng trong dữ liệu văn bản. Ta có tập huấn luyện  $D = \{(X, Y)\}$  với  $X$  là vector đặc trưng,  $Y$  là các nhãn tương ứng với cái loại khoảng trắng. Ta gọi nhãn  $1$  biểu thị cho khoảng trắng tách âm tiết của một từ, và nhãn  $0$  biểu thị cho khoảng trắng tách các từ. Xác suất nhãn  $1, 0$  được tính bằng công thức (3,4).

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_i X_i)} \quad (3)$$

$$P(Y = 0|X) = 1 - P(Y = 1|X) \quad (4)$$



Hình 1: Nguyên lý hoạt động của UETSegmenter.

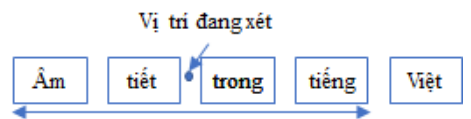
### 2.4. Phần mềm DongDu

Kết quả tách từ có sử dụng các phương pháp như mô hình Markov ẩn[15], CRFs đều cho kết quả thực nghiệm với độ chính xác không cao vì cần lượng từ điển lớn và đối với từ mới thì không nhận diện được, do đó một từ sai sẽ kéo theo các từ khác sai khi tách từ.

Để khắc phục sự hạn chế, phần mềm DongDu được xây dựng dựa trên phương pháp Pointwise[16]. Một lượng từ vựng vừa phải và có khả năng xác định từ mới không có trong từ điển bằng sự hỗ trợ của công cụ máy véc tơ hỗ trợ (SVMs), cách này giúp tập trung vào từng nhãn độc lập và giảm kích thước dữ liệu của mô hình.

Ba đặc trưng cơ bản của phương pháp Pointwise được sử dụng trong phần mềm DongDu là: n-gram âm tiết, n-gram chủng loại âm tiết, đặc trưng từ điển. Cụ thể gồm:

- N-gram âm tiết: Theo thực tế thì từ ghép trong tiếng Việt có 70% từ gồm hai âm tiết và 14% từ gồm ba âm tiết. Vì vậy khi đánh giá từ đang được chọn có phải là từ ghép ta có thể kiểm tra ba hoặc hai từ xung quanh của từ đang xét (Hình 2).
- N-gram chủng loại âm tiết, bao gồm:
  - Âm tiết viết hoa: Những âm tiết bắt đầu bằng chữ hoa.
  - Âm tiết viết thường: Những âm tiết chỉ gồm những chữ cái thường.
  - Số: Là các chữ số.
  - Loại khác: Kí hiệu, tiếng nước ngoài, và không thuộc vào ba loại trên.



Hình 2: Ví dụ về n-gram âm tiết.

- Đặc trưng từ điển: Là những từ tồn tại trong từ điển. Đối với hai phương pháp trên sẽ thực hiện gom các từ trong phạm vi một hoặc hai từ xung quanh vị trí đang xét, và xem

cụm từ vừa ghép đó có phải là từ ghép có trong từ điển hay không. Do đó, trong đặc trưng từ điển chỉ cần xét từ đó có trong từ điển hay không.

### 3. Xây dựng tập dữ liệu

Tập dữ liệu nghiên cứu được xây dựng dựa trên việc sưu tầm các mẫu tin tức trên Báo VnExpress[17] trong chuyên mục thời sự. Sau quá trình thực hiện xử lý: Loại bỏ hình, đường link, quảng cáo, các ký tự đặc biệt,...nghiên cứu thu được bộ dữ liệu gồm 10.000 câu. Sau đó nghiên cứu thực hiện tách từ thủ công và chia ra làm năm tập dữ liệu với số lượng câu được mô tả trong *Bảng 1*.

**Bảng 1:** Thống kê tập dữ liệu kiểm tra.

Tập dữ liệu	1	2	3	4	5
Số câu	2000	4000	6000	8000	10000

### 4. Kết quả nghiên cứu và bình luận

Nghiên cứu sử dụng những phần mềm đã được giới thiệu trong phần II để kiểm tra tập dữ liệu đã được xây dựng ở phần III, kết quả được thể hiện ở Bảng 2 như sau:

**Bảng 2:** Kết quả thực nghiệm trên tập dữ liệu được xây dựng.

Phần mềm	Tổng số từ tách được	Độ chính xác	Thời gian xử lý (giây)
<b>Tập dữ liệu 1</b>			
VnTokenizer	24592	98.30%	<b>0.49</b>
JvnSegmenter	23110	93.73%	15.41
Dongdu	21714	90.80%	1.46
UTESegmenter	24230	<b>99.91%</b>	28.05
<b>Tập dữ liệu 2</b>			
VnTokenizer	54517	98.70%	<b>1.11</b>
JvnSegmenter	50776	93.57%	62.04
Dongdu	48123	89.70%	3.27
UTESegmenter	54012	<b>99.91%</b>	29.00
<b>Tập dữ liệu 3</b>			
VnTokenizer	84775	98.70%	<b>1.72</b>
JvnSegmenter	79204	93.80%	141.00
Dongdu	75959	88.80%	5.09
UTESegmenter	84038	<b>99.93%</b>	31.11
<b>Tập dữ liệu 4</b>			
VnTokenizer	110169	97.90%	<b>2.22</b>
JvnSegmenter	103021	94.51%	248.11
Dongdu	99450	89.23%	7.38
UTESegmenter	109306	<b>99.50%</b>	35.06
<b>Tập dữ liệu 5</b>			
VnTokenizer	143933	97.95%	<b>2.85</b>

JvnSegmenter	134886	94.52%	458.10
Dongdu	128946	89.19%	9.53
UTESegmenter	142737	<b>99.50%</b>	37.46

Qua quá trình thực nghiệm, nghiên cứu thấy rằng VnTokenizer có tốc độ xử lý cao nhất khi thực hiện lần lượt tăng số lượng câu kiểm tra. Với việc lọc ra những danh từ riêng, thư điện tử, ngày tháng làm cho đoạn văn bản xử lý ngắn đi hơn rất nhiều. Cùng với đó, việc kết hợp kho từ điển và thuật toán xây dựng đồ thị để chọn ra các từ ghép có khả năng tách làm giảm đi thời gian xử lý so với việc tìm các từ xung quanh và so sánh với từ điển như công cụ DongDu hoặc phải tăng độ dài của từ đang xét một cách lần lượt rồi so với dữ liệu trong từ điển với thuật toán CRFs được sử dụng trong công cụ JvnSegmenter.

Xét về độ chính xác, công cụ UTESegmenter cho ra kết quả với độ chính xác cao nhất nhưng tốc độ xử lý dữ liệu thì chưa thật sự hiệu quả. Công cụ đã liệt kê ra hết tất cả cả cụm từ từ ba âm tiết trở lên, nên việc nhận diện các từ ghép từ ba âm trở lên là rất chính xác. Ngoài ra với việc phân loại khoảng trắng sẽ giúp cho việc xác định từ ghép trở nên đơn giản hơn thay vì việc xác định lượng lớn các từ khác nhau và phân loại từ đó thuộc từ ghép. Tuy vậy việc lưu lượng lớn các từ có lớn hơn hai âm tiết và việc dùng hồi quy logistic dẫn đến kho từ điển và dữ liệu huấn luyện cho model lớn gây ra việc tốn nhiều thời gian cho việc tải chương trình lên bộ nhớ máy tính.

### 5. Kết Luận

Sau khi thực nghiệm các phần mềm trên với các tập dữ liệu văn bản được thu thập, dễ dàng nhận thấy công cụ VnTokenizer đạt được tốc độ xử lý nhanh, do đó sẽ phù hợp với việc xây dựng các ứng dụng trên website bị giới hạn về thời gian đáp ứng. Tuy nhiên, để cần độ chính xác, và có những yêu cầu không quá khắc khe về thời gian có thể sử dụng phần mềm UTESegmenter. Bên cạnh đó, nghiên cứu cũng góp phần tạo ra dữ liệu chuẩn để phục vụ cho các nghiên cứu về sau cũng như áp dụng vào các công trình trí tuệ nhân tạo có sử dụng tiếng Việt.




### TÀI LIỆU THAM KHẢO

- [1] Aravind K. Joshi, "Natural Language Processing", *Science* 13, Vol. 253, Issue 5025, 1991, pp. 1242-1249, DOI: 10.1126/science.253.5025.1242.
- [2] Kho Ngữ liệu quốc tế Anh Quốc, <http://ice-corpora.net/ice>
- [3] Kho Ngữ liệu quốc gia Hoa Kỳ, <http://www.anc.org>
- [4] Phan Thị Hà, Nguyễn Thị Minh Huyền, Lê Hồng Phương, Adam Kilgarriff, Siva Reddy, "Nghiên cứu từ vựng tiếng Việt với hệ thống Sketch Engine", *Tạp chí Tin học Và Điều khiển học*, Số 3(27), 2011, tr. 206 - 217.
- [5] Nguyễn Phương Thái và các cộng sự, *Báo cáo kết quả sản phẩm SP 7.3- Kho ngữ liệu tiếng Việt có chú giải*, KC01/01, Dự án VLSP, 2009.
- [6] Lê Thanh Hương, *Báo cáo kết quả sản phẩm SP 8.5 - Nghiên cứu xây dựng công cụ phân tích câu Việt*, KC01/01, Dự án VLSP, 2009.
- [7] Hồ Tú Bào, *Về xử lý tiếng Việt trong công nghệ thông tin*, Viện Khoa học và Công nghệ Tiên tiến Nhật Bản. VLSP - KC01/06-10, 2012.
- [8] H. P. Lê, T. M. H. Nguyen, A. Roussanaly and T. V. Ho, "A Hybrid Approach to Word Segmentation of Vietnamese Texts", *International Conference on Language and Automata Theory and Applications, LATA 2008: Language and Automata Theory and Applications*, Tarragona, Spain, 2008, pp. 240-249, ISBN: 978-3540882817.
- [9] ISO/TC 37/SC 4 AWI N309, *Language Resource Management* -

- Word Segmentation of Written Texts for Mono-lingual and Multi-lingual Information Processing - Part I: General Principles and Methods*, Technical Report, ISO, 2006.
- [10] Sven Martin, Jör Liermann, Hermann Ney, “Algorithms for bigram and trigram word clustering”, *Speech Communication*, Volume 24, Issue 1, 1998, pp.19-37, ISSN: 0167-6393.
- [11] Nguyen, C.-T., Nguyen, T.-K., Phan, X.-H., Nguyen, L.-M., and Ha, Q.-T., “Vietnamese Word Segmentation with CRFs and SVMs: An Investigation”, *In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, Publisher: Tsinghua University Press, China, 2006, pp. 215–222.
- [12] John Lafferty, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, *ICML 2001 In Proceedings of International Conference on Machine Learning*, Publisher: ACM New York, 2001, pp. 282-289.
- [13] Lafferty, J., McCallum, A., Pereira, F., “Conditional Random Fields: probabilistic models for segmenting and labeling sequence data”, *The 18th International Conference on Machine Learning*, Massachusetts, USA, 2001, pp.282—290, ISBN:1-55860-778-1
- [14] Nguyen, T.-P. and Le, A.-C., A Hybrid Approach to Vietnamese Word Segmentation, In Proceedings of the 2016 IEEE RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for the Future, 2016, pp.114–119.
- [15] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, Maximum mutual information estimation of hidden markov model parameters for speech recognition, in Proc. International Conference on Acoustics, Speech and Signal Processing, Tokyo, April 1986, pp. 49–52.
- [16] Tuan Anh Luu, Kazuhide Yamamoto, A Pointwise Approach for Vietnamese Diacritics Restoration, IALP 2012 International Conference on Asian Language Processing, Vietnam, Publisher: IEEE Computer Society Washington, 2012, ISBN: 978-0-7695-4886-9.
- [17] Chuyên mục Thời sự của VnExpress: <https://vnexpress.net/tin-tuc/thoi-su>

(BBT nhận bài: .../.../2018, phản biện xong: .../.../2018)

## Thông tin về tác giả

	<p>Phan Tấn Phát:</p> <ul style="list-style-type: none"> <li>- Tóm tắt quá trình đào tạo, nghiên cứu : 2017-2021: Học kỹ thuật phần mềm, Đại học FPT Cần Thơ</li> <li>- Tóm tắt công việc hiện tại (chức vụ, cơ quan); Sinh viên Đại học FPT Cần Thơ</li> <li>- Lĩnh vực quan tâm: Xử lý ngôn ngữ tự nhiên, Machine Learning, Artificial Intelligence.</li> <li>- Điện thoại: 0899.467.737</li> </ul>
	<p>ThS. Quách Luyt Đa:</p> <ul style="list-style-type: none"> <li>- Tóm tắt quá trình đào tạo, nghiên cứu: 2011: Kỹ sư Tin học, Trường ĐH Tây Đô. 2016: Thạc sỹ Hệ thống thông tin, ĐH Cần Thơ.</li> <li>- Tóm tắt công việc hiện tại (chức vụ, cơ quan): Khoa Kỹ thuật – Công nghệ, trường Đại học Tây Đô. Giảng viên thỉnh giảng Đại học FPT Cần Thơ.</li> <li>- Lĩnh vực quan tâm: Xử lý ngôn ngữ tự nhiên, Xử lý ảnh, Machine learning;</li> <li>- Điện thoại: 0976 703 075</li> </ul>
	<p>TS. Dương Trung Nghĩa</p> <ul style="list-style-type: none"> <li>- Tóm tắt quá trình đào tạo, nghiên cứu: 2008: Kỹ sư Tin học, trường Đại học Cần Thơ. 2011: Thạc sỹ Công nghệ phần mềm và quản lý, trường Đại học Heilbronn, cộng hòa liên bang Đức. 2017: Tiến sĩ Khoa học máy tính, trường Đại học Hildesheim, cộng hòa liên bang Đức.</li> <li>- Tóm tắt công việc hiện tại: Trưởng Bộ môn Khoa học máy tính, khoa Công nghệ thông tin, trường Đại học Kỹ thuật – Công nghệ Cần Thơ.</li> <li>- Giảng viên thỉnh giảng Đại học FPT Cần Thơ.</li> <li>- Lĩnh vực nghiên cứu: Machine Learning, Data Mining, Artificial Intelligence.</li> <li>- Điện thoại: 0939.657.063</li> </ul>