Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.000111

ỨNG DỤNG ĐỒ THỊ CÓ HƯỚNG KHÔNG TRỌNG SỐ NHẬN DIỆN TỪ TRONG VĂN BẢN TIẾNG KHMER

Trần Văn Nam[1,2], Sơn Phú Quý[2], Nguyễn Thị Huệ[1], Phan Huy Khánh[2]

[1] Trường Đại học Trà Vinh, Việt Nam

[2] Trường Đại học Bách khoa, Đại học Đà Nẵng, Việt Nam

namtv@tvu.edu.vn, sonqphu@gmail.com, *huetvu@tvu.edu.vn*, phkhanh@dut.udn.vn

TÓM TẮT: *Chúng tôi tiếp tục thực hiện bài toán tách từ còn tồn động ở bài báo trước [5]. Bài báo đề xuất giải pháp tách từ dựa trên so khớp âm tiết, sử dụng kho ngữ liệu cấu trúc âm tiết tiếng Khmer đã xây dựng, sử dụng kết quả so khớp được thực hiện xây dựng đồ thị có hướng không có trọng số đón nhận từ vựng. Từ đó có được tất cả các cách phân tách câu. Hướng tiếp cận hoàn toàn khả thi và có kết quả thử nghiệm đạt độ chính xác cao, nhận diện từ mới ở mức âm tiết, nhận diện câu có nhập nhằng, góp phần giải quyết bài toán tách từ ứng dụng hiệu quả trong xử lý tiếng Khmer.*

Từ khóa: *đồ thị, tách từ, kho ngữ liệu cấu trúc âm tiết, nhập nhằng, tiếng Khmer.*

I. GIỚI THIỆU

Bài toán tách từ cho ngôn ngữ đơn lập đã đặt ra từ lâu, chủ yếu để giải quyết cho tiếng Trung Quốc, tiếng Nhật [11] [12] [13] [15]. Cho đến hiện nay thì có rất nhiều ngôn ngữ được áp dụng. Các thuật toán tách từ có thể gồm hai loại. Loại thứ nhất là dựa trên luật như: Longest Matching, Mô hình so khớp tối đa [4] [6] [11] [14] (đối với mô hình này gồm: so khớp tối đa tiến và so khớp tối đa lùi), phương pháp này hoàn toàn phụ thuộc vào từ điển. Loại thứ hai là dựa trên thống kê: Phương pháp này dựa vào ngữ cảnh xung quanh để đưa ra quyết định thích hợp [7] [8] [9], giải pháp này gồm hai vấn đề chính. Vấn đề thứ nhất là độ rộng ngữ cảnh, nếu độ rộng ngữ cảnh lớn thì thuật toán rất phức tạp. Vấn đề thứ hai là cách áp dụng thống kê, giải pháp này phụ thuộc nhiều vào ngữ liệu huấn luyện. Các cách tách từ khác thường dựa trên sự lai tạo giữa các mô hình trên.

Tuy nhiên trong xử lý tiếng Khmer, bài toán tách từ luôn đặt ra thách thức, cho đến nay vẫn chưa được giải quyết một cách thỏa đáng [4] [6] [10]. Về mặt ngôn ngữ, chính tả tiếng Khmer rất phức tạp, trong câu gồm nhiều từ cũng như trong từ gồm nhiều âm tiết viết liền nhau mà không có dấu phân cách.

Bài báo sử dụng cách tiếp cận tách từ tiếng Việt để xây dựng phương pháp tách từ tương tự cho tiếng Khmer.

Phương pháp tách từ tiếng Việt được mô tả chi tiết trong bài báo [16], khá giống như trường hợp của tiếng Khmer và đạt được kết quả khá cao. Chúng tôi sử dụng phương pháp này áp dụng cho tiếng Khmer.

Bài báo đề xuất giải pháp tách từ tiếng Khmer dựa trên kho ngữ liệu cấu trúc âm tiết đã được xây dựng. Xây dựng đồ thị có hướng không trọng số để phân tích các phương án tách từ. Sau phần mở đầu, bài báo gồm các nội dung như sau: Mã hóa ngữ liệu cấu trúc âm tiết, mô hình và phương pháp tách từ, trình bày giai đoạn thử nghiệm và đánh giá kết quả. Phần cuối là kết luận và hướng phát triển tiếp theo.

II. MÃ HÓA DỮ LIỆU KHO NGỮ LIỆU ÂM TIẾT TỪ VỰNG TIẾNG KHMER

A. Kho ngữ liệu cấu trúc âm tiết từ vựng tiếng Khmer

Chúng tôi đã xây dựng kho ngữ liệu cấu trúc âm tiết tiếng Khmer nhằm phục vụ cho việc tách từ và kiểm tra lỗi chính tả ở pha sau, phương pháp xây dựng kho ngữ liệu cấu trúc âm tiết là dựa vào các mô hình đặc trưng âm tiết của ngôn ngữ Khmer [5]. Cấu trúc từ điển mã hóa theo cấu trúc âm tiết là việc sắp xếp các đơn vị mục từ theo trật tự được xác định. Mục từ gồm các từ đơn, từ ghép, cụm từ. Mỗi đơn vị mục từ có dấu phân cách (space) giữa các âm tiết.

Ví dụ: Kho ngữ liệu cấu trúc âm tiết cho các từ như sau:

Bảng 1. Kho ngữ liệu cấu trúc âm tiết tiếng Khmer

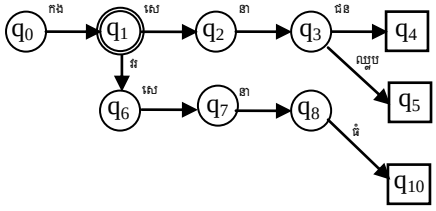
Kho từ vựng	Kho ngữ liệu cấu trúc âm tiết	Nghĩa tiếng Việt
កង	កង	Vòng đeo cổ
កងសេនាជន	កង សេ នា ជន	Dân quân
កងសេនាឈ្លប	កង សេ នា ឈ្លប	Du kích
កងវរសេនាធំ	កង វ រ សេ នា ធំ	Trung đoàn
...	...	...

B. Mã hóa kho ngữ liệu theo cấu trúc âm tiết

Mục đích mã hóa kho ngữ liệu cấu trúc âm tiết nhằm phục vụ cho việc so khớp theo từng âm tiết của từ vựng, kho ngữ liệu cấu trúc âm tiết được sắp xếp giảm dần theo số lượng âm tiết của từ vựng trước khi mã hóa.

Dựa vào cấu trúc dữ liệu, chúng tôi xây dựng Automat bằng thuật toán để mã hóa lưu trữ kho ngữ liệu cấu trúc âm tiết cho việc so khớp âm tiết. Kho ngữ liệu cấu trúc âm tiết gồm có khoảng 48947 từ vựng đã xây dựng, mỗi âm tiết trong từ vựng có dấu phân cách bởi khoảng trắng. Mỗi cung của Automat có nhãn là một âm tiết, các cung đi từ một trạng thái xuống các trạng thái con của nó phải mang các nhãn khác nhau.

Ví dụ, với bốn từ vựng theo cấu trúc âm tiết như: កង (Vòng đeo cổ), កង សេ នា ជន (Dân quân),កង សេ នា ឈ្លបDu kích), កង វ សេ នា ធំ (Trung đoàn).



Hình 1. Xây dựng Automat mã hóa kho ngữ liệu cấu trúc âm tiết tiếng Khmer

Trong Hình 1:

○ Ký hiệu trạng thái chưa kết thúc.

◯ Ký hiệu trạng thái kết thúc của một từ.

□ Ký hiệu trạng thái vừa là kết thúc, vừa là nút lá của từ.

Thuật toán xây dựng Automat kho ngữ liệu cấu trúc âm tiết

Input: kho ngữ liệu cấu trúc âm tiết.
Output: Automat kho ngữ liệu cấu trúc âm tiết.
1. Lập trạng thái khởi đầu q_0;
2. Vòng lặp đọc cho tới khi hết tệp dữ liệu, lấy ra từng mục từ *word*. Gọi các âm tiết của *word* là $a_0, a_1, \dots, a_{n-1}$;
   a. p := q_0; i := 0;
   b. Vòng lặp trong khi (i ≤ n - 1)
      Lấy ra âm tiết a_i;
      Tìm trong các cung chuyển từ trạng thái p cung trên đó ghi âm tiết a_i.
      Nếu có cung (p, q) thì i:=i+1; p:=q;
      Ngược lại thì thoát khỏi vòng lặp b.
3. Vòng lặp với j từ i đến n - 1
   Tạo mới trạng thái q , ghi nhận q là trạng thái không kết thúc;
   Thêm cung chuyển (p, q) trên đó ghi âm tiết a_j;
   p := q;
4. Nếu q có tồn tại cung chuyển (p,q), ghi nhận q là trạng thái kết thúc;
   Ngược lại, ghi nhận q là trạng thái vừa là kết thúc, vừa là nút lá của từ.

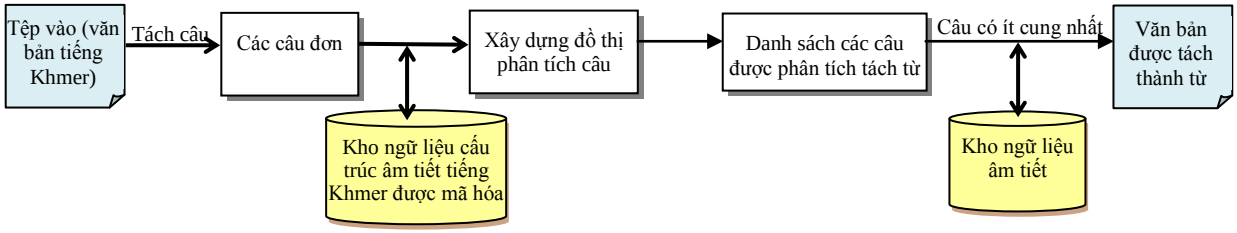
Sau khi đã xây dựng xong Automat, ta lưu vào tệp để dùng trong bước phân tách từ vựng và bắt lỗi chính tả ở pha sau.

III. MÔ HÌNH VÀ GIẢI PHÁP TÁCH TỪ

A. Đề xuất mô hình tách từ tiếng Khmer

Giải quyết bài toán đặt ra là đầu vào là một văn bản tiếng Khmer, hãy tách văn bản đó thành các câu, trong mỗi câu tách thành những đơn vị từ vựng (từ) hoặc chỉ ra những âm tiết nào không có trong từ điển (phát hiện đơn vị từ vựng mới).

Trên cơ sở tìm hiểu chính tả, ngữ pháp tiếng Khmer [1] [2] [3], chúng tôi đề xuất mô hình tách từ dựa trên so khớp âm tiết và kho ngữ liệu cấu trúc âm tiết.



Hình 2. Mô hình tách từ tiếng Khmer

B. Giải pháp tách từ tiếng Khmer

Chúng tôi lần lượt xử lý theo mô hình tách từ tiếng Khmer các bước sau:

1) Tách câu

Hệ thống đọc văn bản đầu vào cần tách câu, dựa trên dấu kết thúc câu để ngắt câu ra thành từng câu để xử lý. Mỗi câu sẽ được xử lý độc lập với nhau. Các câu được phân cách bởi các dấu kết thúc câu gồm: ?, ! ,។, ៕.

Phân tích văn bản U gồm m câu đơn $U=w_1w_2...w_m,(m\geq 1)$. Chúng tôi sử dụng lại mô hình cấu trúc âm tiết, tách lần lượt từng câu đơn thành từng âm tiết riêng biệt.

Mỗi câu đơn chứa n âm tiết $w_m= a_0, a_1, \ldots, a_{n-1}$.

2) Xây dựng đồ thị phân tích câu

Ý tưởng phân tách từ vựng là quy việc phân tách câu về việc tìm đường đi trên một đồ thị có hướng, không có trọng số. Giả sử câu đơn ban đầu là một dãy gồm n âm tiết $a_0, a_1, ..., a_{n-1}$. Xây dựng một đồ thị có n+1 đỉnh $q_0, q_1, ..., q_n$. Sắp theo thứ tự trên một đường thẳng từ trái sang phải; Với các cung là các từ trong câu, đỉnh là đường nối giữa hai từ kề nhau, hướng thể hiện của các từ trong câu.

Để kiểm tra có bao nhiêu từ vựng được thành lập do thực hiện so khớp lần lượt các âm tiết trong câu theo thứ tự từ trái sang phải, ta xét từ nút gốc của Automat và xét lần lượt các âm tiết trong câu, mỗi khi xét qua cung chuyển rẽ sang nhánh con theo cung có nhãn tương ứng với a_i và nếu quá trình di chuyển gặp trạng thái kết thúc hoặc trạng thái nút lá ở một trạng thái nào đó trên Automat thì đó là từ của câu, lưu vào danh sách từ vựng, cứ tiếp tục như vậy cho đến hết câu. Nếu tại một bước nào đó việc chuyển xuống nhánh con thất bại do không tìm được cung có nhãn tương ứng thì không là từ của câu, thoát khỏi vòng lặp.

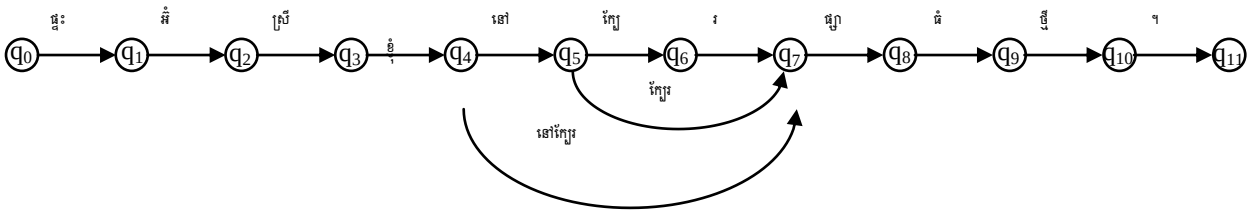
Dựa vào danh sách từ vựng vừa thu được, thực hiện vẽ các cung trên đồ thị bắt đầu tại đỉnh đang xét. Sau đó tìm tiếp tục như trên cho các từ còn lại, bắt đầu từ âm tiết kế sau đỉnh đang xét, thực hiện cho đến khi hết các từ trong câu.

Sau khi xây dựng đồ thị xong, bước tiếp theo, liệt kê tất cả các cách phân tích câu dựa trên đồ thị bằng phương pháp duyệt theo chiều sâu. Khi đó mỗi cách phân tách câu khác nhau tương ứng với một đường đi trên đồ thị từ đỉnh đầu q_0 đến đỉnh cuối q_{n+1}. Dựa trên đồ thị phân tích câu, chúng tôi thống kê kết quả bằng thủ công, tập dữ liệu thử nghiệm gồm 500 câu về lĩnh vực xã hội, do trường Đại học Trà Vinh cung cấp. Kết quả được chia ra làm hai nhóm như sau: Nhóm 1 là tập câu có duy nhất một đường đi ngắn nhất, đúng cú pháp lớn hơn 95% trên tổng số câu, nhóm 2 là các câu còn lại, đúng cú pháp khoảng 5%. Nhận thấy các câu có đường đi trên đồ thị nhiều cung nhất đều sai cú pháp. Chúng tôi quyết định tạm thời chọn kết quả phân tích câu đúng đắn nhất ứng với đường đi qua ít cung nhất trên đồ thị.

Trường hợp, câu có nhiều hơn một đường đi ngắn nhất trên đồ thị, xác định câu xảy ra nhập nhằng. Liệt kê tất cả các cách phân tích câu nhập nhằng.

Trường hợp trong câu có âm tiết không có mặt trong từ điển âm tiết, xác định một đơn vị âm tiết (từ vựng) mới. Âm tiết mới gồm âm tiết bị sai lỗi chính tả hoặc âm tiết mới chưa tồn tại trong từ điển.

Ví dụ: Câu văn bản đầu vào của tiếng Khmer là: "ផ្ទះអ៊ំស្រីខ្ញុំនៅក្បែរផ្សារធំថ្មី។" (*Nhà bác gái của tôi ở bên cạnh chợ lớn mới*)



Hình 3. Xây dựng đồ thị phân tích câu

Kết quả các phương án được phân tích từ đồ thị:

Phương án 1: ផ្ទះ | ឆ្នាំ | ស្រី | ខ្ញុំ | នៅក្បែរ | ផ្សារ | ធំ | ថ្មី | ។ |

Phương án 2:ផ្ទះ | ឆ្នាំ | ស្រី | ខ្ញុំ | នៅ | ក្បែរ | ផ្សារ | ធំ | ថ្មី | ។ |

Phương án 3:ផ្ទះ | ឆ្នាំ | ស្រី | ខ្ញុំ | នៅ | ក្បែ | រ | ផ្សារ | ធំ | ថ្មី | ។ |

Kết quả được chọn là phương án 1, có ít cung nhất.

IV. CHẠY THỬ NGHIỆM VÀ ĐÁNH GIÁ GIẢI PHÁP

Để tiến hành thử nghiệm, chúng tôi đã sử dụng kho từ vựng âm tiết tiếng Khmer có độ lớn như sau:

Số lượng các từ đơn: 7278 từ

Số lượng các từ ghép: 17095 từ

Số lượng các cụm từ: 24574 từ

Chúng tôi đã sử dụng máy tính cá nhân cho các văn bản thuộc lĩnh vực thông tin xã hội, do trường Đại học Trà Vinh cung cấp, với năm trường hợp khác nhau về độ lớn văn bản.

Kết quả thử nghiệm tách từ được đánh giá dựa trên sự kết hợp giữa hai độ đo**Error! Reference source not found.**:

Độ chính xác (Precision) là tỷ lệ giữa các âm tiết tách đúng trên tổng số âm tiết tách được, Precision bằng 100% có nghĩa là tất cả các âm tiết đều phù hợp.

$$\text{Precision} = \frac{\text{Số âm tiết tách đúng}}{\text{Tổng số âm tiết tách được}}$$

Độ bao phủ (Recall) là tỷ lệ giữa các từ tách đúng trên tổng số từ cần tách.

$$Recall = \frac{\text{Số từ tách đúng}}{\text{Tổng số từ cần tách}}$$

Chỉ số F-score được sử dụng để đánh giá hiệu quả tổng thể của hệ thống bằng cách kết hợp hai chỉ số Precision và Recall.

$$F - score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

Bảng 2. Kết quả thử nghiệm phương pháp tách từ

Độ lớn văn bản	Số lượng từ đơn	Số từ ghép	Số lượng cụm từ	F-score
<30 KB	300	180	391	98%
>30 KB và <100 KB	620	257	598	97%
>100 KB và <150 KB	735	486	702	94%
>150 KB và <250 KB	852	591	758	92%
>250 KB và <500 KB	970	622	799	89%
Trung bình F-score				**94%**

Phương pháp tách từ tiếng Khmer dùng kho ngữ liệu cấu trúc âm tiết, xây dựng đồ thị phân tích câu của ngôn ngữ Khmer đã đạt được độ chính xác trung bình F-score khoảng 94%. Hạn chế của phương pháp là chưa nhận diện được những từ ghép sai lỗi chính tả. Mặt khác, kho từ vựng chưa đủ lớn nên gây ra nhiều từ mới hay từ sai lỗi chính tả. Chưa giải quyết triệt để câu có nhập nhằng.

Để đạt kết quả hợp lý, chúng tôi nghiên cứu tiếp theo xử lý từ mới hay từ sai lỗi chính tả, để tăng độ chính xác cho việc tách từ đồng thời giải quyết nhập nhằng triệt để.

V. KẾT LUẬN

Giải pháp đề xuất mô hình phân tách từ sử dụng kho ngữ liệu âm tiết tiếng Khmer, kết hợp kho ngữ vựng từ điển. Xây dựng đồ thị phân tích câu, hoàn toàn khả thi và có kết quả thử nghiệm đạt độ chính xác cao so với một số phương pháp tách từ trước đây. Phương pháp này nhận diện được nhận dạng từ mới: âm tiết mới (hay âm tiết sai lỗi chính tả), chưa giải quyết được từ ghép bị sai chính tả.

Hướng nghiên cứu tiếp theo của chúng tôi là nhận dạng các từ có chứa từ ghép bị sai lỗi chính tả, có thể tạo thêm các cung mờ (lưới từ) trên đồ thị, gợi ý cho người dùng các từ đúng chính tả và nghiên cứu phương pháp xử lý nhập nhằng trong câu triệt để. Để chọn phương án đúng nhất giữa nhiều phương án, có thể dùng các qui tắc cú pháp do chuyên gia ngôn ngữ Khmer xây dựng. Tiến hành phân tích cú pháp của câu với những phương án tách từ vựng có thể, từ đó loại ra những phương án sai cú pháp.

LỜI CẢM ƠN

Để hoàn thành bài báo này, các tác giả chân thành cảm ơn Sư Thạch Qui và Sư Keo Samane, chuyên gia ngôn ngữ Khmer đang theo học đại học ngành Công nghệ thông tin tại trường Đại học Trà Vinh. Cảm ơn trường Đại học Trà Vinh, đặc biệt là ông Tăng Văn Thòn, giảng viên giảng dạy ngôn ngữ Khmer, khoa Ngôn ngữ-Văn hóa-Nghệ thuật Khmer Nam Bộ, đã tạo điều kiện thuận lợi cho việc cập nhật dữ liệu văn bản phục vụ cho việc chạy thử nghiệm.

TÀI LIỆU THAM KHẢO

[1] Hắc Sóc Hi, "Ngữ pháp tiếng Khmer", Học viện Giáo dục Dân tộc, 2012.

[2] K. Sok, "Khmer Language Grammar", First Edition of Royal Academic of Cambodia, 2004.

[3] Chan Som Nop, "Từ và Các phương thức cấu tạo từ trong tiếng Khmer", Nxb. Campuchia, 2010.

[4] Ly Vattana. "Các tiếp cận tách từ tiếng Khmer dùng trong cơ sở dữ liệu văn bản". Tạp chí Khoa học ĐHQGHN, Khoa học Tự nhiên và Công nghệ 27 pp251-258, 2011.

[5] Tran Van Nam, Nguyen Thi Hue, Phan Huy Khanh, "Building a Syllable Database to Solve the Problem of Khmer Word Segmentation". International Journal on Natural Language Computing (IJNLC) Vol. 6, No.1. 2017.

[6] Narin Bi, Nguonly Taing. "Khmer word segmentation based on bi-directional maximal matching for plaintext and microsoft word document". Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014, Chiang Mai, Thailand, IEEE, December 9-12, pages 1-9. 2014.

[7] Villavon Souksan, Phan Huy Khánh. "Tách từ tiếng Lào sử dụng kho ngữ vựng kết hợp với các đặc trưng ngữ pháp tiếng Lào". Hội thảo quốc gia lần thứ XVI, 14-15/11/2013.

[8] Villavon Souksan, Phan Huy Khánh. "Khử bỏ nhập nhằng trong bài toán tách từ tiếng Lào". Tạp chí Khoa học & Công nghệ ĐH Đà Nẵng no1 (62), pp 113-119. 2013.

[9] Vichet Chea, Ye Kyaw Thu, Chenchen Ding, Masao Utiyama, Andrew Finch, Eiichiro Sumita. "Khmer Word Segmentation Using Conditional Random Fields". In Khmer Natural Language Processing, December 4, 2015.

[10] Chea Sok Huor, Top Rithy, Ros Pich Hemy, Vann Navy. "Word Bigram Vs Orthographic Syllable Bigram in Khmer Word Segmentation". PAN Localization Team, Cambodia, pp 249-253, 2007.

[11] C. Chan P. Wong, "Chinese Word Segmentation Based on Maximum Matching and Word Binding Force". Proceedings of Coling 96, p200-203, 1996.

[12] K. J. Chen and S. H. Liu. "Word identification for madarin chinese sentences". Proceedings of the Fifthteenth International Conference for Computational Linguistics, 1992.

[13] Richard W Sproat, Chilin Shih, William Gale, and Nancy Chang. "A stochastic finite-state word-segmentation algorithm for Chinese". CL, 22(3):377-404, 1996.

[14] Dinh Dien, Hoang Kiem, and Nguyen Van Toan. "Vietnamese word segmentation", NLPRS, 11 2001.

[15] David Palmer. "A trainable rule-based algorithm for word segmentation", Proceedings of the 35th Annual Meeting of the Association for Computational Linguistisc, 1997.

[16] Phuong Le-Hong, Huyen Nguyen-Thi-Minh, Azim Roussanaly, Vinh Ho-Tuong. "A hybrid approach to word segmentation of Vietnamese texts", Language and Automata Theory and Applications pp 240-249. LATA 2008.

APPLICATIONS FOR THE GRAPH HAS DIRECTION FOR THE PROBLEM OF KHMER WORD SEGMENTATION

Tran Van Nam, Son Phu Quy, Nguyen Thi Hue, Phan Huy Khanh

ABSTRACT: *We continue to perform the word separation problem in the previous article. The paper proposes a solution for word separation based on syllable matching, using a Khmer syllabus constructed using the result of a non-weighted directional mapping. Take vocabulary From there all the sentences are separated. The approach is feasible and has high accuracy test results, new vocabulary recognition at syllabic level, ambiguous sentence identification, contributes to solve the separation problem from the problem of Khmer word segmentation.*