Tạp chí Khoa học ĐHQGHN, Khoa học Tự nhiên và Công nghệ 27 (2011) 251-258

Các tiếp cận tách từ tiếng Khmer dùng trong cơ sở dữ liệu văn bản

Ly Vattana*

Trung tâm nghiên cứu đa phương tiện MICA, Trường Đại học Bách khoa Hà Nội, Số 1 Đại Cồ Việt, Hà Nội, Việt Nam

Nhận ngày 11 tháng 8 năm 2010

Tóm tắt. Bài báo đề cập bài toán tách từ, sử dụng trong việc tổ chức dữ liệu văn bản bằng tiếng Khmer. Bài toán này quan trọng trong xử lí ngôn ngữ tiếng Khmer. Cũng như tiếng Trung Quốc, tiếng Thái, tiếng Khmer không có các dấu hiệu phân tách để phân biệt các từ trong câu. Bài báo sẽ phân tích và so sánh hai phương pháp tiếp cận khác nhau trong bài toán tách từ tiếng Khmer : Tiếp cận dựa trên ký tự (*Character-based approaches*) và Tiếp cận dựa trên từ (*Word-based approachs*). Hai cách tiếp cận này được thử nghiệm trong các ngôn ngữ độc lập như tiếng Trung Quốc, và tiếng Thái. Đây là một trong những giải pháp cho bài toán tách từ tiếng Khmer.

Từ khóa: Tách từ, tiếng Khmer, ngôn ngữ.

1. Giới thiệu

Tách từ là một bài toán quan trọng trong các hệ thống đánh chỉ mục và tìm kiếm văn bản tiếng Khmer [1]. Mục đích của bài toán nhằm xác định ranh giới của các từ ở trong câu. Không giống như tiếng Anh và một số tiếng khác, tách từ của ngôn ngữ tiếng Khmer (cũng như một số ngôn ngữ châu Á) rất phức tạp bởi vì trong ngôn ngữ này, các từ được viết liền nhau, không có ranh giới giữa các từ ví dụ như các khoảng trắng,... Nhiều nghiên cứu nhằm đề xuất các phương pháp tách từ cho các tiếng Trung Quốc, Thái Lan đã được đề xuất [2, 3]. Do tiếng Thái và tiếng Khmer có nhiều điểm tương tự nhau về mặt hình thái và cú pháp. Ta có thể áp dụng và cải tiến một số phương pháp tách từ trong tiếng Thái vào bài toán tách từ tiếng Khmer.

Trước tiên, các đặc trưng của tiếng Khmer được phân tích, rồi đưa ra các phương pháp thích hợp cho bài toán tách từ tiếng Khmer. Cấu trúc của bài báo gồm các phần như sau: phần (2) giới thiệu phương pháp tách từ đã được đề xuất cho các ngôn ngữ khác như tiếng Thái. Phần (3) sẽ tập trung vào phân tích các đặc trưng của tiếng Khmer. Trong phần (4), hai hướng tiếp cận chính được trình bày cho bài toán tách từ trong tiếng Khmer. Phần (5) trình bày các kết quả đánh giá thử nghiệm cũng như các phân tích về kết quả thử nghiệm. Phần (6) đưa ra các kết luận và các hướng phát triển tiếp theo.

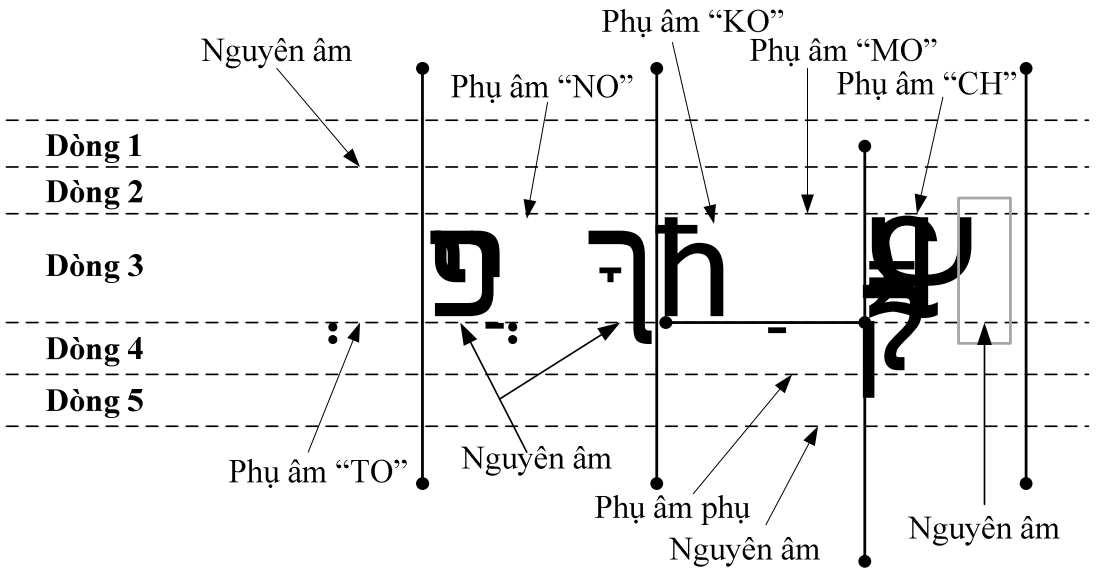
* E-mail: Vattana.ly@mica.edu.vn

2. Giới thiệu phương pháp tách từ

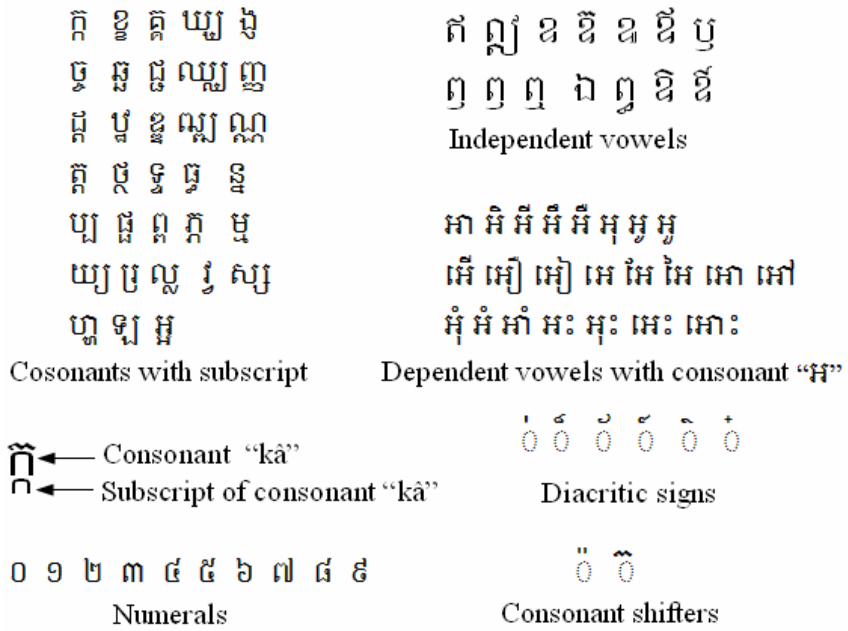
Trên thực tế, đã có nhiều phương pháp đề xuất cho bài toán tách từ tiếng Thái [4]. Có thể phân loại các phương pháp đó thành: tách từ dựa trên quy tắc, tách từ dựa trên từ điển. Thairatananond và Chamyapompong [3] đã phát triển hệ thống tách từ dựa trên quy tắc cho tiếng Thái. Tuy nhiên, hệ thống chỉ mới tách được các âm tiết mà chưa tách được các từ. Các hướng giải quyết từ trước đến nay đều dựa trên các từ điển được xây dựng bằng tay để lấy thông tin về các từ. Việc phân tách các từ được thực hiện bằng cách áp dụng các chiến lược khác nhau như đối sánh xâu dài nhất (Poowarawan 1986), đối sánh cực đại (Sornlertlamvanich 1993). Bên cạnh đó, các phương pháp mô hình tri-gram (Kawtrakul 1997), và tách từ dựa trên các đặc trưng (Meknavin 1997) cũng được đề xuất. Một số phương pháp tách từ sử dụng thống kê các từ trên Internet và không sử dụng từ điển để tránh phải giải quyết bài toán xử lí các từ không tìm có trong từ điển (Theeramunkong 2000).

3. Đặc điểm của tiếng Khmer

Hệ thống chữ viết Khmer, được gọi là *aksaa khmae* "ký tự Khmer" là hệ thống chữ viết chính thức của Campuchia. Tiếng Khmer không sử dụng các kí hiệu phân tách từ trong câu một cách rõ ràng. Trong tiếng Khmer, các ký tự phụ âm kết hợp với các ký hiệu đặt thêm phía trước, phía trên, phía dưới, và/hoặc phía sau để tạo thành một âm tiết và tiếng Khmer viết từ trái sang phải.



Hình 1. Phân tích hệ thống chữ viết tiếng Khmer "Đây là Campuchia".
Bộ chữ tiếng Khmer bao gồm: 33 phụ âm, 14 nguyên âm độc lập, 23 nguyên âm, các dấu kết thúc, xuống dòng, dấu nhắc và các chữ số.



Hình 2. Bộ chữ tiếng Khmer [5] .

Tất nhiên, đối với con người khi sử dụng những ngôn ngữ này sẽ không gặp bất kì một trở ngại nào. Họ vẫn có thể liên lạc với nhau hiệu quả và chính xác dưới cả hai hình thức nói hoặc viết mà không cần xác định rõ các từ trong câu. Tuy nhiên, máy tính không biết được cách phân tách các từ trong câu. Việc phân tách nhầm lẫn các từ sẽ làm ảnh hưởng nghiêm trọng tới kết quả của các chương trình ở mức cao hơn.

Ví dụ câu

ពណ៌សម្តេចថាខ្មៅ

tách ra

ពណ៌ | ស | ម្តេច | ថា | ខ្មៅ

Tức màu | trắng | sao | bảo | đen" hoặc

ពណ៌ | សម្តេច | ថា | ខ្មៅ

Tức "màu | vua | bảo | đen".

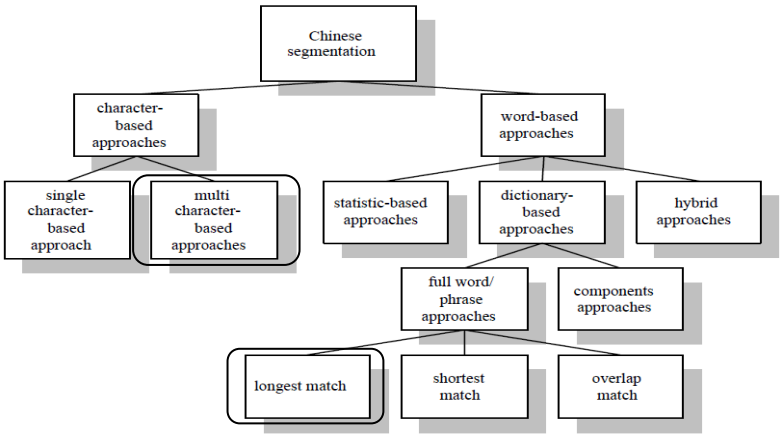
Cấu trúc chữ viết của tiếng Khmer rất phức tạp, không chỉ các từ viết liền nhau có thể tách ra, mà sau khi tách các ký tự ra cũng có thể mang lại các ý nghĩa khác nhau. Việc tách từ không chính xác có thể làm thay đổi hoàn toàn nghĩa của câu.

4. Các phương pháp và hướng tiếp cận

Để có một kết quả tách được từ tốt, người ta sử dụng hai hướng tiếp cận để giải quyết vấn đề này [6, 7] :

- Hướng tiếp cận dựa trên ký tự : Hướng tiếp cận này đơn thuần rút trích một số lượn nhất định các tiếng trong câu như rút trích *1 ký tự* (uni-gram) hay *nhiều ký tự* (n-gram); ví dụ câu ABCDEF, sẽ chia câu thành AB CD EF. Một biến thể của phương pháp này là chia chồng lên nhau, cho phép ABCDEFG chia thành AB BC CD DE EF FG.

- Tiếp cận dựa trên từ : Hướng tiếp cận dựa trên từ với mục tiêu tách được các từ hoàn chỉnh trong câu. Hướng tiếp cận này có thể chia ra 3 hướng: *dựa trên thông kế* (statistic-based) *dựa trên từ điển* (dictionary-based) và *hydrid.*



Hình 3. Dựa trên hướng tiếp cận tách từ của Tiếng Trung [2].

4.1. Phương pháp tiếp cận dựa trên ký tự

Trật tự chuẩn của các thành phần trong một âm tiết chính tả được trình bày như sau:

CLUSTER := B {R|C} {S{R}}*{{Z}V} {O} {S} [8]

Trong đó => **B** là ký tự cơ bản (ký tự phụ âm, ký tự nguyên âm độc lập)

R là một robat

C là phụ âm

S là một phụ âm phụ hoặc dấu nguyên âm độc lập

V là một dấu nguyên âm độc lập

Z là khoảng trống có độ rộng bằng 0

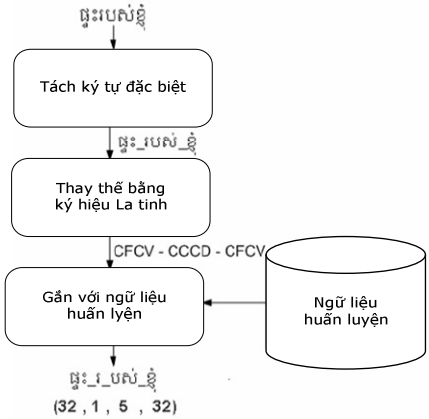
O là bất kỳ dấu nào khác

Với tiếp cận này, ví dụ có ផ្ទះរបស់ខ្ញុំ "nhà của tôi" và hệ thống sẽ thực hiện các bước như sau:

a) Tìm những ký tự đặc biệt: đây là công việc đầu tiên hệ thống cần làm là tìm những ký tự đặc biệt và sau đó tách nó ra những ký tự đằng sau nó, các ký tự đặc biệt ở đây gồm ◌ះ, ◌ំ

b) Thay thế bằng ký hiệu la tinh: sau khi tách ra những ký tự đặc biệt ផ្ទះ_របស់_ខ្ញុំ công việc tiếp theo là thay thế bằng ký hiệu La tinh.

c) Gắn với ngữ liệu huấn luyện: những ký hiệu La tinh được tập hợp lại và sau đó so sánh với ngữ liệu huấn luyện.



Hình 4. Công đoạn tách từ. **C**: consonant (phụ âm), **FC**: foot consonant (phụ âm phụ), **D**: diacritic (dấu), **V**: vowel (nguyên âm), **I**: independent vowel (nguyên âm độc lập).

Trong chuẩn Unicode, coeng nyo và om thường được tách xa hơn, và một từ hoàn chỉnh được trình bày bởi năm ký tự được mật mã. Mỗi câu đưa vào, hệ thống tạm thời tách ký tự đặc biệt và kiểm tra âm tiết theo dữ liệu huấn luyện sau đó tổ hợp lại âm tiết đó thành từ và so sánh trong từ điển Khmer (Chhoun Nat dictionary, official Khmer dictionary) cho đến khi thu được cụm từ hoàn chỉnh.

Một thí dụ như trong đoạn sau:

Dữ liệu đưa vào:
កម្ពុជាបានអំពាវនាវសុំអោយបណ្តាក្រុមហ៊ុនបរទេសទៅធ្វើវិនិយោគថែមទៀតនៅក្នុងវិស័យទូរស័ព្ទចល័តនៅប្រទេសកម្ពុជា

Kết quả tách thủ công:
កម្ពុជា - បាន - អំពាវនាវ - សុំអោយ - បណ្តា - ក្រុមហ៊ុន - បរទេស - ទៅ - ធ្វើ - វិនិយោគ - ថែមទៀត - នៅ - ក្នុង - វិស័យ - ទូរស័ព្ទ - ចល័ត - នៅ - ប្រទេស - កម្ពុជា

Kết quả tách tự động:
កម្ពុ | ជា | បាន | អំ | ពា | វ | នាវ | សុំ | អោយ | បណ្តា | ក្រុម | ហ៊ុន | ប | រ | ទេស | ទៅ | ធ្វើ | វិ | និ | យោគ | ថែម | ទៀត | នៅ | ក្នុង | វិ | ស័យ | ទូ | រ | ស័ព្ទ | ច | ល័ត | នៅ | ប្រ | ទេស | កម្ពុ | ជា

Dịch:
Campuchia kêu gọi các công ty nước ngoài đầu tư thêm nữa trong lĩnh vực viễn thông di động ở Campuchia.

4.2. Phương pháp tiếp cận dựa trên từ

Phương pháp khớp tối đa (Longest-Matching). Phương pháp này, người ta sẽ duyệt một ngữ hoặc câu từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ từ điển Khmer (Chhoun Nat dictionary, official Khmer dictionary) và cứ thực hiện lặp lại như vậy cho đến hết câu. Dạng đơn giản của phương pháp dùng để giải quyết nhập nhằng từ đơn. Giả sử có một chuỗi ký tự $C_1, C_2, \ldots, C_n$. Người ta sẽ áp dụng phương pháp từ đầu chuỗi. Đầu tiên kiểm tra xem C_1 có phải là từ hay không, sau đó kiểm tra xem C_1C_2 có phải là từ hay không. Tiếp tục thực hiện như thế cho đến khi tìm được từ dài nhất. Dạng phức tạp dạng này là phân đoạn từ. Thông thường người ta chọn phân đoạn ba từ có chiều dài tối đa. Thuật toán bắt đầu từ dạng đơn giản, cụ thể là nếu phát hiện ra những cách tách từ gây nhập nhằng, như ở ví dụ trên, giả sử C_1 là từ và C_1C_2 cũng là một từ, khi đó chúng ta kiểm tra ký tự kế tiếp trong chuỗi $C_1, C_2, \ldots, C_n$ để tìm tất cả các đoạn ba từ có bắt đầu với C_1 hoặc C_1C_2.

Thí dụ minh họa như trong đoạn sau:

Dữ liệu đưa vào:
កម្ពុជាបានអំពាវនាវសុំអោយបណ្តាក្រុមហ៊ុនបរទេសទៅធ្វើវិនិយោគថែមទៀតនៅក្នុងវិស័យទូរស័ព្ទចល័តនៅប្រទេសកម្ពុជា

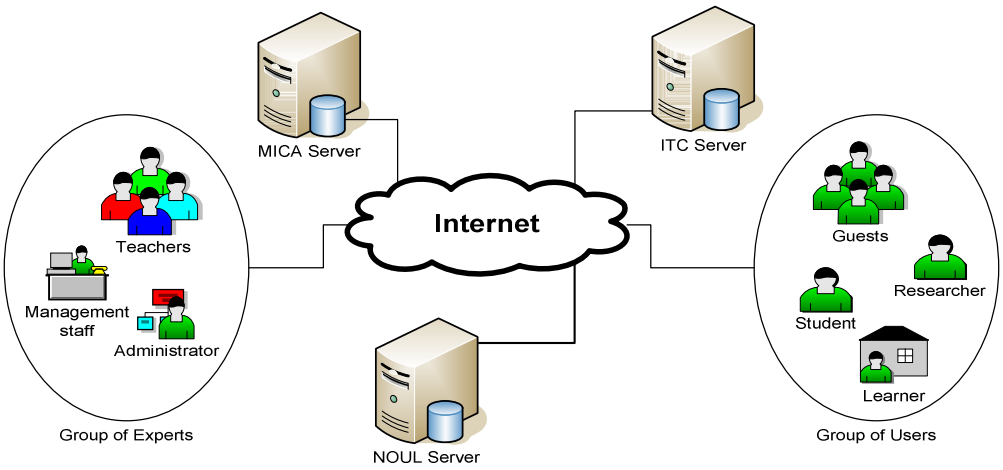
Kết quả tách thủ công:
កម្ពុជា - បាន - អំពាវនាវ - សុំអោយ - បណ្តា - ក្រុមហ៊ុន - បរទេស - ទៅ - ធ្វើ - វិនិយោគ - ថែមទៀត - នៅក្នុង - វិស័យ - ទូរស័ព្ទ - ចល័ត - នៅ - ប្រទេស - កម្ពុជា

Kết quả tách tự động:
កម្ពុជា | បាន | អំពាវនាវ | សុំ | អោយ | បណ្តា | ក្រុមហ៊ុន | បរទេស | ទៅធ្វើ | វិនិយោគ | ថែមទៀត | នៅក្នុង | វិស័យ | ទូរស័ព្ទចល័ត | នៅ | ប្រទេស | កម្ពុជា

Dịch:
Campuchia kêu gọi các công ty nước ngoài đầu tư thêm nữa trong lĩnh vực viễn thông di động ở Campuchia.

5. Đánh giá kết quả

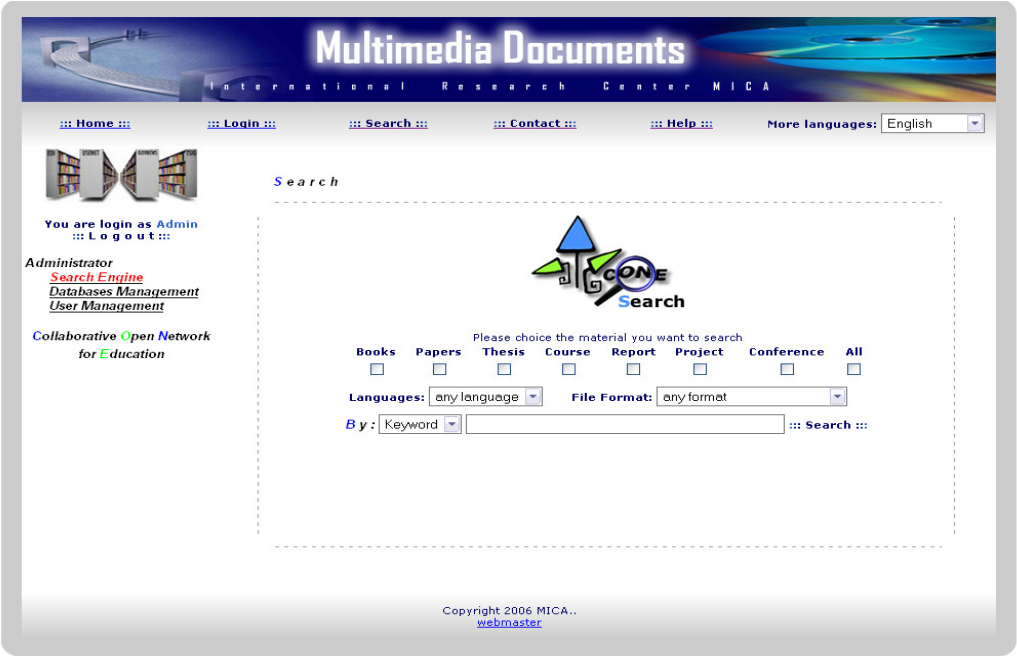
Hệ thống CONE được triển khai tại Trung tâm MICA, Trường Đại học Bách khoa Hà Nội, nhằm cung cấp học liệu điện tử cho các Việt Nam, Lào và Cămpuchia. Người dùng truy cập CONE từ trang tin Internet. Người ta cần đăng nhập hệ thống theo tài khoản đã được cấp phát. Việc sử dụng giao thức HTTP để truyền kiến thức phục vụ đào tạo là phù hợp trong điều kiện và môi trường kinh tế xã hội của Việt Nam, Lào và Cămpuchia.



Hình 5. Mô hình của hệ thống.

Hình 6. Trang chủ của CONE trên Internet.

Trung tâm MICA có thể đảm bảo nguồn tư liệu học tập bằng tiếng Việt. Một phần do người Việt đã đầu tư nghiên cứu về xử lí và tổng hợp tiếng Việt, xử lí văn bản tiếng Việt. Với văn bản tiếng Khmer, các kết quả nghiên cứu đã có không mấy khả quan. Cần thiết có mô hình tổ chức dữ liệu văn bản tiếng Khmer với các đặc thù ngôn ngữ, ngữ cảnh sử dụng ngôn ngữ và nguồn từ vựng tiếng Khmer.



Hình 7. Chức năng tìm kiếm.

Trong thử nghiệm này, tập văn bản thử nghiệm được xây dựng, với trên 50 văn bản tiếng Khmer và kết quả thử nghiệm tách từ đánh giá giữa trên sự kết hợp của hai độ đo: độ bảo phủ (Recall), độ chính xác (Precision). Từ kết quả trả về, chúng ta có thể biết được khả năng tách từ của hai cách tiếp cận trên. Độ bao phủ là tỉ lệ giữa các từ tách đúng trả về trên tổng số các từ được trong cơ sở dữ liệu. Trong khi đó, độ chính xác là tỉ lệ giữa các từ tách được đúng trên từ tách được.

Tập văn bản thử nghiệm trên 50 văn bản tiếng Khmer và kết quả của cách tiếp cận dựa trên từ điển đạt được: 95% từ đúng và tiếp cận dựa trên nguyên tắc là 85%.

6. Kết luận

Sau khi xem xét hai hướng tiếp cận trong tách từ tiếng Khmer, kết quả chỉ ra rằng phương pháp tách từ dữa trên từ mang lại kết quả có độ chính xác cao hớn điều này điều này có được nhờ vào tập từ điển lớn, được đánh dấu ranh giới giữa các từ chính xác giúp cho việc so sánh để tách từ cho các văn bản tiếng Khmer được tốt đẹp, tuy nhiên dễ nhận thấy hiệu suất của phương pháp hoàn toàn phụ thuộc vào tập từ điển. Hướng tiếp cận dựa trên ký tự có ưu điểm là dễ thực hiện, thời gian thực hiện tương đối nhanh, tuy nhiên lại cho kết qủa không chính xác bằng hướng tiếp cận dựa trên từ. Hướng tiếp cận này nói chung phù hợp cho các ứng dụng không cần độ chính xác tuyệt đối. Mỗi

phương pháp đều có những ưu điểm và nhược điểm riêng vì trong tiếng Khmer ngữ nghĩa của từ có thể thay đổi theo ngữ cảnh cho nên hướng phát triển tiếp theo là ghiên cứu sâu về ngữ nghĩa của từ và ngữ cảnh của câu trong việc tách từ nói chung và tìm kiếm thông tin trong hệ thống học tiếng Khmer nói riêng dựa vào thống kê, dựa vào từ điển và dựa vào ngữ pháp.

Tài liệu tham khảo

[1] F.E. Huffman, *Cambodian systems for writing and begining reader*, 1970

[2] Li, S.F.a.H., *Chinese Word Segmentation and Its Effect on Information Retrieval,* 2004.

[3] Aroonmanakun, W., *Collocation and Thai Word Segmentation,* 2002.

[4] T. Theeramunkong, S. Usanavasin, *Non-dictionary-based Thai word segmentation using decision trees, in Proceedings of the first international conference on Human language technology research.* 2001, Association for Computational Linguistics: San Diego.///4

[5] O.P. Ye Kyaw Thu, Yoshiyori URANO and Mitsuji MATSUMOTO, *A Word-based Predictive Text Entry Method for Khmer Language*, 2008.///5

[6] P. Hok, *Development of a Khmer Spell Checker Based on a Hidden Markov Model,* 2005.

[7] D.D. Palmer, *A Trainable Rule-based Algorithm for Word Segmentation.* 1996.

[8] J. Solá, *Issues in Khmer Unicode* 4.0. 2004.

Approaches for segmenting words in Khmer language in text database application

Ly Vattana

MICA Center, Polytechnic Institute of Hanoi, 1 DaiCoViet Street, Hanoi, Vietnam

Word segmentation is an important problem in processing Khmer language. Like Chinese and Thailand, Khmer language has no space sign to distinguish words in a sentence. In this research, we will analyse and compare two different approaches in Khmer word segmentation problem: Character-based approaches and word-based approaches. This two appraches were tested in independent language as Chinese and Thailand. This is one of solutions to Khmer word segmentation problem.