PHÁT TRIỂN CÔNG CỤ TÁCH TỪ VÀ GIẢI NGHĨA TỪ HÁN - VIỆT TRONG VĂN BẢN

DEVELOPING WORD SEGMENTATION TOOL AND SINO - VIETNAMESE WORD INTERPRETATION IN DOCUMENT

Võ Ngọc Đạt

Trường Cao đẳng CNTT Hữu nghị Việt - Hàn, Khoa Thương mại điện tử và Truyền thông;
Email: vongocdatit@gmail.com

Tóm tắt

Hiện nay đã có rất nhiều công trình nghiên cứu về xử lý ngôn ngữ như tiếng Anh, tiếng Nhật, tiếng Trung Quốc trên văn bản tương đối tốt. Và cũng có rất nhiều phần mềm giải nghĩa từ Hán - Việt như Song Kiều, HVDic... nhưng chỉ cho phép giải nghĩa từng từ. Thực tế chưa có những công cụ trợ giúp để giải nghĩa từ Hán - Việt trên văn bản tiếng Việt. Trên văn bản, để xác định được từ thuần Việt, từ Hán - Việt, từ khác, điều bắt buộc đối với tiếng Việt đó là việc tách từ. Trong bài viết này, chúng tôi giới thiệu về một số phương pháp tách từ trên văn bản tiếng Việt, áp dụng tách từ vào việc giải nghĩa từ Hán - Việt trên văn bản, đặc biệt giúp cho người dùng tra cứu từ Hán - Việt trên văn bản được nhanh và tiện lợi.

Từ khóa: *Tách từ, xử lý ngôn ngữ tự nhiên, từ tiếng Việt, cơ cấu ngữ âm tiếng Việt, ngôn ngữ tiếng Việt, từ Hán - Việt.*

Abstract

At present there are many studies on language processing, including English, Japanese, Chinese writing as well as a lot of Sino-Vietnamese word intepretation softwares as Song Qiao, HVDic... but with interpretation in words. In fact, there are no tools that help to interpret Sino-Vietnamese words in Vietnamese documents. In documents, to determine whether it is a Vietnamese word, a Sino-Vietnamese word, or another word, word segmentation is compulsory. In this article, we introduce a number of word segmentation methods in Vietnamese documents, applying word segmentation to interpret Sino-Vietnamese words in a document and specially providing users with a fast and convenient search for Sino-Vietnamese words in documents.

Keywords: *Word segmentation, natural language processing, Vietnamese words, Structure of Vietnamese phonetics, Vietnamese language, Sino-Vietnamese words.*

1. Giới thiệu

Ngôn ngữ chính là một trong những nét thể hiện tâm hồn của con người, bản sắc văn hóa của một dân tộc. Nhưng trong quá trình hình thành tiếng Việt thì nó đã vay mượn khá nhiều từ của tiếng

nước ngoài, đặc biệt tiếng Hán, tiếng Pháp. Vì thế việc hiểu đúng nghĩa của từ đặc biệt là từ Hán - Việt sẽ giúp mỗi chúng ta trao dồi thêm ngôn ngữ cho chính mình, tìm lại những nét văn hóa cổ của dân tộc và góp phần vào quá trình giao lưu quốc tế khi Việt Nam đã là thành viên của WTO [1].

Đối với tiếng Anh, từ là một nhóm các ký tự có nghĩa được tách biệt với nhau bởi khoảng trắng trong câu, do vậy việc tách từ trở nên rất đơn giản. Trong khi đối với tiếng Việt, ranh giới từ không được xác định mặc định là khoảng trắng mà tùy thuộc vào ngữ cảnh dùng câu tiếng Việt. Vấn đề trên thực sự đưa ra một thách thức đối với những người làm tin học.

Trong bài viết này, chúng tôi trình bày giải pháp và kết quả thử nghiệm việc phát triển công cụ tách từ và áp dụng tách từ vào việc giải nghĩa từ Hán - Việt trên văn bản. Điều này giúp người sử dụng hiểu từ Hán - Việt, biết cách dùng từ Hán - Việt.

2. Tình hình nghiên cứu

2.1. Các vấn đề trong bài toán tách từ

2.1.1. Xử lý nhập nhằng

Nhập nhằng trong tách từ được phân thành hai loại:

- Nhập nhằng chồng (Overlapping Ambiguity);

- Nhập nhằng hợp (Combination Ambiguity).

Ta gọi D là tập hợp các từ tiếng Việt (từ điển tiếng Việt). Các trường hợp nhập nhằng trên được mô tả hình thức như sau:

- Chuỗi $\alpha\beta\gamma$ được gọi là nhập nhằng chồng nếu $\{\}$, $D\ \alpha\beta\ \beta\gamma$.

- Chuỗi $\alpha\beta$ được gọi là nhập nhằng hợp nếu $\{\}$, $D\ \alpha\ \beta\ \alpha\beta$.

Trong thực tế loại nhập nhằng chồng thường xảy ra hơn loại nhập nhằng hợp bởi hầu hết các tiếng của tiếng Việt đều có thể đóng vai trò là một từ đơn độc lập. Do đó, hầu hết các từ ghép đều có thể bị nhập nhằng hợp [2].

2.1.2. Nhận diện từ chưa biết

Trong văn bản không chỉ có sự tồn tại của từ thuần tuý có trong từ điển, mà còn có các đơn vị thông tin khác nữa. Từ chưa biết bao gồm các tên riêng tiếng Việt hoặc từ tiếng nước ngoài và các factoids.

2.2. Phương pháp Maximum Matching [4]

Maximum Matching (MM) được xem như là phương pháp tách từ dựa trên từ điển đơn giản nhất. MM cố gắng so khớp với từ dài nhất có thể có trong từ điển. Đó là một thuật toán ăn tham (Greedy Algorithms) nhưng bằng thực nghiệm đã chứng minh được rằng thuật toán này đạt được độ chính xác >90% nếu từ điển đủ lớn. Tuy nhiên, nó không thể giải quyết vấn đề nhập nhằng và không thể nhận diện được các từ chưa biết bởi vì chỉ những từ tồn tại trong từ điển mới được phân đoạn đúng.

Giải quyết MM gồm 2 giải thuật con: FMM (Forward Maximum Matching: so khớp cực đại theo chiều tiến) và BMM (Backward Maximum Matching: so khớp cực đại theo chiều lùi). Nếu chúng ta nhìn vào kết quả của FMM và BMM thì sự khác biệt này cho chúng ta biết nơi nào nhập nhằng xảy ra.

Ngoài ra, MM là phương pháp tách từ hoàn toàn phụ thuộc vào từ điển, từ điển phải đủ lớn, đủ chính xác và độ tin cậy phải cao thì mới cho kết quả tách từ chấp nhận được. Đây cũng là nhược điểm rất lớn của phương pháp này.

Ví dụ: Người nông dân ra sức cải tiến bộ công cụ lao động của mình.

Đầu ra FMM:

Người# nông dân #ra sức# cải tiến# bộ# công cụ# lao động# của# mình.

Đầu ra BMM:

Người# nông dân #ra sức# cải# tiến bộ# công cụ# lao động# của# mình.

2.3. Phương pháp tách từ bằng WFST (Weighted Finite State Transducer) [2]

Gồm các bước:

Xây dựng từ điển trọng số: theo mô hình WFST, việc phân đoạn từ được xem như là một sự chuyển dịch trạng thái có xác suất. Chúng ta miêu tả từ điển D là một đồ thị biến đổi trạng thái hữu hạn có trọng số. Giả sử:

H: là tập các từ chính tả tiếng Việt (còn gọi là "tiếng");

P: là từ loại của từ (Pos: Part - Of - Speech).

Mỗi cung của D có thể là:

Từ một phần tử của H tới một phần tử của H;

Hoặc từ ε (ký hiệu kết thúc từ) tới một phần tử của P.

Các nhãn trong D biểu thị một chi phí ước lượng (estimated cost) bằng công thức:

$$cost = -\log(f/N)$$

Với f: tần số của từ, N: kích thước tập mẫu.

Đối với các trường hợp từ mới chưa gặp, tác giả áp dụng xác suất có điều kiện Goog-Turning (Baayen) để tính toán trọng số.

Xây dựng các khả năng phân đoạn từ: Để giảm sự bùng nổ tổ hợp khi sinh ra các dãy các từ có thể từ một dãy các tiếng trong câu, tác giả đề xuất một phương pháp mới là kết hợp dùng từ điển để hạn chế sinh ra các bùng nổ tổ hợp. Khi phát hiện thấy một cách phân đoạn từ nào đó không phù hợp (không có trong từ điển, không phải là từ láy, không phải là danh từ riêng...) thì tác giả loại bỏ các nhánh xuất phát từ cách phân đoạn từ đó.

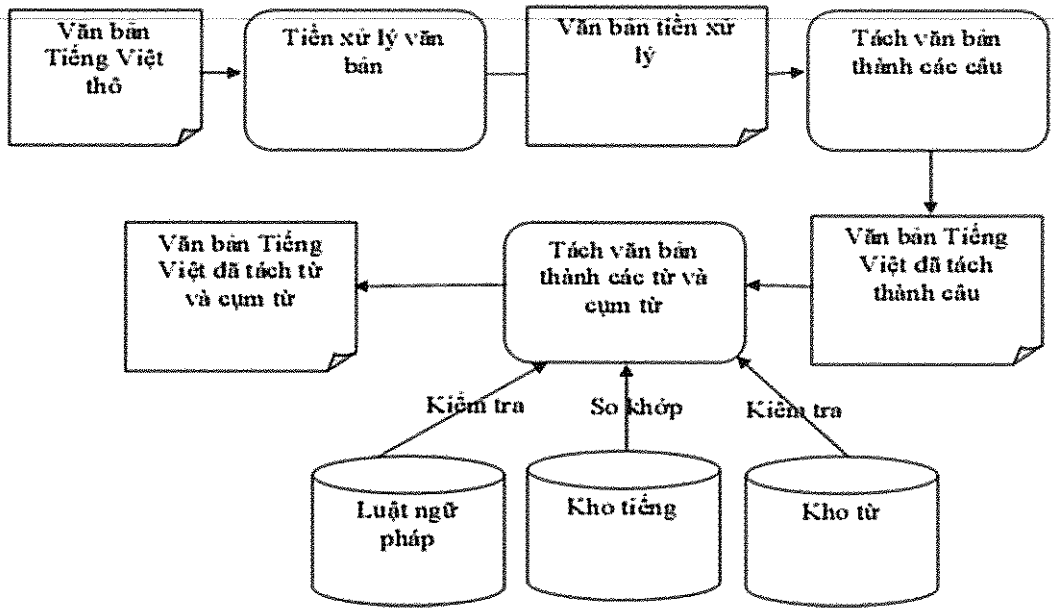
Lựa chọn khả năng phân đoạn từ tối ưu: Sau khi được một danh sách các cách phân đoạn từ có thể có của câu, tác giả chọn trường hợp phân đoạn từ có trọng số bé nhất.

Tầng mạng neural:

Sau khi cho câu được tách qua mô hình WFST để xác định kết quả tách từ trên có thực sự hợp lệ hay không, tác giả đã định nghĩa một ngưỡng giá trị t_0 với ý nghĩa như sau: Nếu sự chênh lệch về trọng số trong các câu được tách so với câu có trọng số bé nhất lớn hơn t_0 thì đó là kết quả tách từ trên thực sự được chấp nhận. Còn nếu có một vài câu mà sự chênh lệch về trọng số của chúng so với câu có trọng số bé nhất nhỏ hơn t_0 thì mô hình WFST chưa thể xác định được ranh giới từ trong câu, lúc này đưa những câu này qua mô hình mạng Neural để xử lý.

3. Giải pháp đề xuất

3.1. Xây dựng mô hình cho bài toán tách từ



Hình 1. Quá trình tách từ tiếng Việt thành các từ và cụm từ

3.2. Áp dụng phương pháp tách từ tiếng Việt

Qua các phương pháp tách từ đã nêu trên, tôi đề xuất chọn phương pháp so khớp tối đa (Maximum Matching). Vì với phương pháp này, ta dễ dàng tách được chính xác các ngữ/câu như "hợp tác xã // mua bán","thành lập // nước // Việt Nam // dân chủ // cộng hòa", cách tách từ đơn giản, nhanh chỉ dựa vào từ điển. Để xử lý nhập nhằng ta sử dụng các luật sau:

Luật 1: Sử dụng Simple Maximum Matching lấy từ với chiều dài dài nhất, Complex maximum matching lấy từ đầu tiên từ dãy với chiều dài dài nhất. Nếu có nhiều dãy có chiều dài dài nhất, áp dụng luật kế tiếp.

Luật 2: hai từ 2 tiếng giống nhau không đi liền nhau. Điều này hoàn toàn đúng trong tiếng Việt, chúng ta xem ví dụ sau đây:

Học sinh học sinh học

Có một số cách tách từ sau đây:

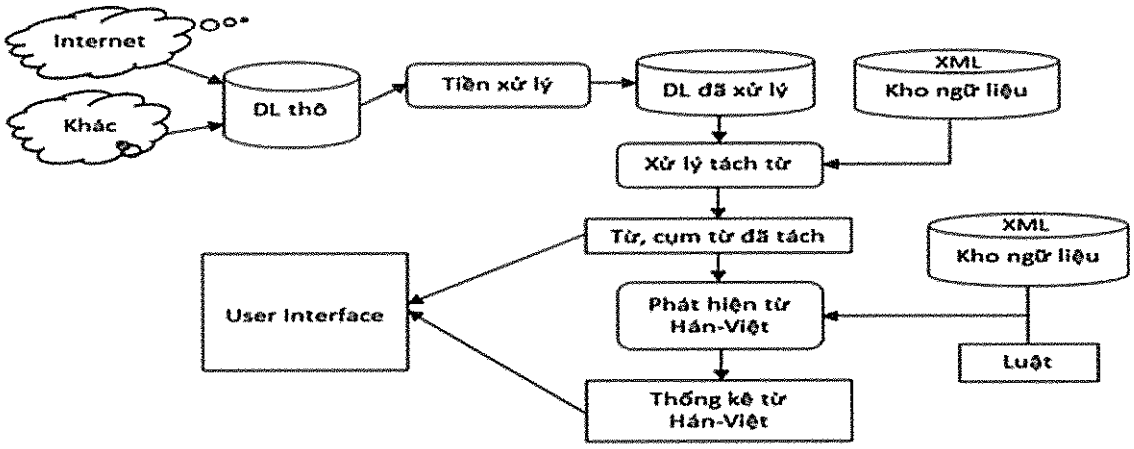
Học sinh# học sinh# học

Học# sinh học# sinh học

Hai từ "Học sinh" và "học sinh" không bao giờ đi liền nhau, cũng như "sinh học" và "sinh học" không bao giờ đi liền nhau.

Luật 3: dựa vào phương pháp WFST để lựa chọn khả năng phân đoạn từ tối ưu.

3.3. Thuật toán phát hiện từ Hán - Việt trong tệp văn bản dựa vào luật



Hình 2. Mô hình xử lý tổng quát

Luật 1: Nếu một chữ có nghĩa nhưng không hoạt động làm thành một từ được mà chỉ làm thành một bộ phận của từ thôi thì đó là một chữ Hán - Việt [1], [3].

Luật 2: Nếu một chữ mà ta không hiểu nghĩa của nó, và nó không làm thành từ, lại xuất hiện trong hai từ có một sự giống nhau nào đó về nghĩa thì đó là một chữ Hán - Việt [1], [3].

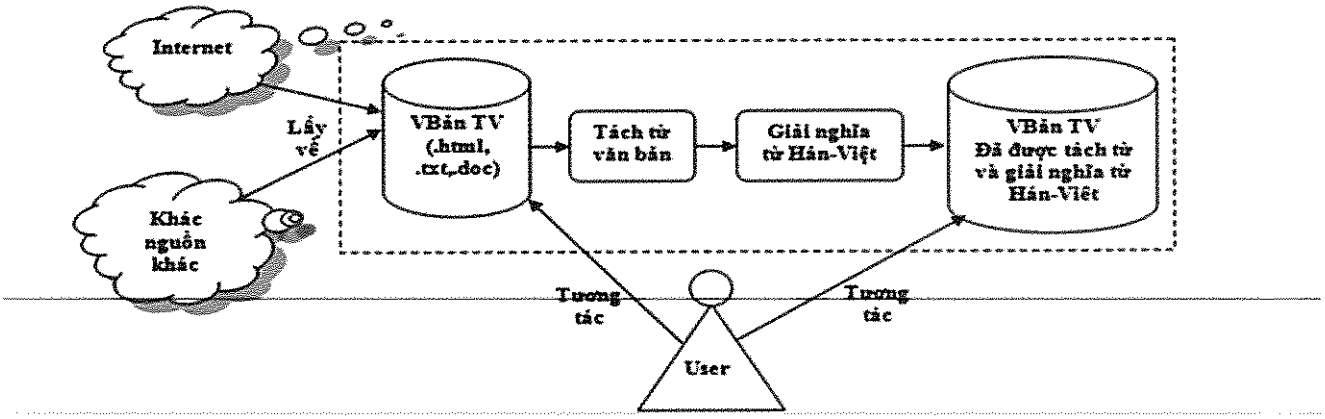
Luật 3: Trong một từ Hán - Việt thì các yếu tố và các chữ của nó đều là từ Hán - Việt. Vậy nếu ta tạo một từ có yếu tố Hán - Việt như ở luật 1 và luật 2 thì các chữ trong từ Hán - Việt đều là từ Hán - Việt [1], [3].

Luật 4: Nếu ta có thể sắp vào trước hoặc sau chữ mà chúng ta đang xét một trong các chữ nhất, hữu, vô, dân, nhân, bất thì đó là chữ Hán - Việt [1], [3].

4. Thiết kế chương trình

4.1. Kiến trúc tổng quát của hệ thống

Hệ thống của chúng tôi tuy có nhiều chức năng khác nhau như: tách từ, cập nhật kho từ tiếng Việt, thống kê… Tuy nhiên, chức năng chính của chương trình vẫn là giải nghĩa từ Hán - Việt trong văn bản.



Hình 3. Mô hình tổng quan của hệ thống

4.2. Thống kê kết quả đạt được

Sau khi đã xây dựng ứng dụng hoàn chỉnh chúng tôi thống kê được thông tin về kho từ vựng sau đây:

4.2.1. Thống kê các từ trong kho tiếng Việt

Bảng 1. Thống kê số lượng từ trong kho tiếng Việt

Kho Từ Vựng	Số Từ	Dung lượng
Kho tiếng	5.465	346 Kb
Kho từ	39.443	2.429 Mb
Kho tính trọng số của từ	9.076	883 KB

4.2.2. Thống kê các từ đơn Hán - Việt theo thứ tự chữ cái

Có tổng số: 1854 từ đơn Hán - Việt trong kho dữ liệu.

Bảng 2. Thống kê các từ đơn Hán - Việt

Vần	Số từ đơn	Dung lượng	Tỉ lệ %
A, Ă, Â	35	87 Kb	1,88%
B	98	314 Kb	5,28%
C	140	400 Kb	7,55%
D, Đ	178	483 KB	9,60%
E, Ê	2	3 Kb	0,10%
G	24	111 Kb	1,29%
H	152	355 Kb	8,19%
I	1	1 Kb	0,05%

K	138	293 Kb	7,44%
L	88	192 Kb	4,74%
M	70	152 Kb	3,77%
N	186	303 Kb	10,03%
O, Ô, Ơ	19	24 Kb	1,02%
P	85	173 Kb	4,58%
Q	42	109 Kb	2,26%
R	0	1 Kb	0,00%
S	83	128 Kb	4,47%
T	391	890 Kb	21,08%
U, Ư	25	42 Kb	1,34%
V	38	99 Kb	2,04%
X	45	70 Kb	2,42%
Y	14	37 Kb	0,75%

4.2.3. Thống kê các từ ghép Hán - Việt theo thứ tự chữ cái

Có tổng số: 5434 từ ghép Hán - Việt trong kho dữ liệu.

Bảng 3. Thống kê các từ ghép Hán - Việt

Vần	**Số từ ghép**	**Dung lượng**	**Tỉ lệ %**
A, Ă, Â	231	87 Kb	4,25%
B	816	314 Kb	15,01%
C	641	400 Kb	11,79%
D, Đ	768	483 KB	14,13%
E, Ê	1	3 Kb	0,01%
G	169	111 Kb	3,11%
H	515	355 Kb	9,47%
I	1	1 Kb	0,01%
K	426	293 Kb	7,83%
L	284	192 Kb	5,22%
M	209	152 Kb	3,84%
N	449	303 Kb	8,26%
O, Ô, Ơ	29	24 Kb	0,53%

P	2	173 Kb	0,03%
Q	146	109 Kb	2,68%
R	0	1 Kb	0,00%
S	32	128 Kb	0,58%
T	371	890 Kb	6,82%
U, Ư	55	42 Kb	1,01%
V	147	99 Kb	2,70%
X	81	70 Kb	1,49%
Y	61	37 Kb	1,12%

Nhìn vào số lượng từ Hán - Việt đã được lưu trữ trong kho dữ liệu trên ta thấy đó mới chỉ là một phần nhỏ trong kho từ vựng đồ sộ. Các từ Hán - Việt sẽ được cập nhật vào kho một cách dần dần để kho dữ liệu này càng đầy đủ để có thể phát triển ứng dụng này một cách toàn diện.

5. Kết luận

Các kết quả nghiên cứu của chúng tôi đã mở ra nhiều hướng tiếp cận mới để giải quyết cho các bài toán liên quan đến xử lý ngôn ngữ tự nhiên mà kết quả hoàn toàn chấp nhận được. Tuy nhiên chúng tôi cũng cần nhấn mạnh rằng, để có thể giải quyết tốt hơn nữa các bài toán liên quan đến xử lý ngôn ngữ tự nhiên trên tiếng Việt, chúng ta cần phải giải quyết tốt các bài toán cơ bản ví dụ như tách từ tiếng Việt. Và điều quan trọng không kém khi sử dụng các phương pháp máy học là chúng ta phải xây dựng được một bộ ngữ liệu hoàn chỉnh và phải được công nhận.

Những mặt còn hạn chế trong quá trình xây dựng kho ngữ liệu tiếng Việt (kho tiếng, kho từ và kho trọng số) và kho ngữ liệu từ Hán - Việt. Chúng tôi sẽ tiếp tục công việc nghiên cứu để hoàn thiện, mở rộng đưa vào phục vụ và khai thác chương trình.

TÀI LIỆU THAM KHẢO

[1]. **Phan Huy Khánh, Huỳnh Ngọc Chiến**, *"Xây dựng công cụ chuyển đổi nhanh giữa văn bản Hán Việt và văn bản chữ Hán"*, Tạp chí Bưu Chính Viễn Thông & Công nghệ thông tin, Tổng biên tập: TS. Chu Văn Vệ, 56/GP-BC.

[2]. **Đinh Điền, Hoàng Kiếm, Nguyễn Văn Toán**, *Vietnamese Word Segmentation. In Processding of NLPPRS*, Tokyo, Japan, 2001, pp 749 - 756.

[3]. **Đoàn Ngọc Diễm My, Bùi Như Diệu**, *Xây dựng các luật xử lý từ Hán - Việt và ứng dụng vào việc phát triển từ Hán - Việt trong văn bản*, Báo cáo tốt nghiệp, Đại học Bách Khoa, Hướng dẫn PGS. TS. Phan Huy Khánh, Đà Nẵng, 2007.

[4]. **J. Berker,** *Multilingual Word Processing Microsystems*, February, 1984, Tr. 96-t106.