1. GIỚI THIỆU

Xử lý ngôn ngữ tự nhiên là một lĩnh vực nghiên cứu của trí tuệ nhân tạo nhằm xây dựng một hệ thống xử lý cho máy tính, làm cho máy tính có thể hiểu được ngôn ngữ của con người gồm cả ngôn ngữ nói và viết. Không chỉ với một ngôn ngữ của một dân tộc, của một quốc gia mà máy tính có thể hiểu được ngôn ngữ của tất cả các dân tộc, các quốc gia trên thế giới. Nhờ đó, mọi người trên thế giới dựa vào máy tính cũng có thể hiểu và giao tiếp được với nhau mà không cần học, hiểu ngôn ngữ của nhau,... Và hơn thế nữa, máy tính có thể phân tích, tổng hợp ngôn ngữ để đưa ra tri thức cho con người một cách nhanh chóng và chính xác. Nhất là khi các dữ liệu liên quan đến ngôn ngữ đang dần trở nên là kiểu dữ liệu chính của con người.

Xử lý ngôn ngữ tự nhiên nhằm mục đích: Phân tích, nhận biết, tổng hợp ngôn ngữ tự nhiên. Là cơ sở chính để hiểu ngôn ngữ, dịch ngôn ngữ, xử lý tiếng nói, xử lý văn bản,... Để xử lý ngôn ngữ tự nhiên bằng máy tính, trên thế giới người ta đã cho ra đời một ngành học mới được kết hợp giữa hai ngành máy tính và ngôn ngữ học, được gọi là ngôn ngữ học máy tính. Trong tương lai máy tính sử dụng ngôn ngữ tự nhiên để giao tiếp giữa người và máy, máy có khả năng hiểu được ngôn ngữ tự nhiên của con người và trả lời các câu hỏi của con người. Thậm chí máy sẽ dịch được các ngôn ngữ tự nhiên từ một ngôn ngữ này sang một một ngôn ngữ khác một cách nhanh chóng và chính xác.

Với một hệ thống xử lý ngôn ngữ tự nhiên, đầu vào của một hệ thống có thể là một hoặc nhiều câu dưới dạng tiếng nói hay văn bản. Các dữ liệu liên quan đến ngôn ngữ viết (văn bản) và nói (tiếng nói) đang dần trở nên kiểu dữ liệu chính con người có và lưu trữ dưới dạng điện tử. Đặc điểm chính của các kiểu dữ liệu này là không có cấu trúc hoặc nửa cấu trúc và chúng không thể lưu trữ trong các khuôn dạng cố định như các bảng biểu. Theo đánh giá của công ty Oracle, hiện có đến 80% dữ liệu không cấu trúc trong lượng dữ liệu của loài người đang có [Oracle Text]. Với sự ra đời và phổ biến của Internet, của sách báo điện tử, của máy tính cá nhân, của viễn thông, của thiết bị âm thanh, … người người ai cũng có thể tạo ra dữ liệu văn bản hay tiếng nói. Vấn đề là làm sao ta có thể xử lý chúng, tức chuyển chúng từ các dạng ta chưa hiểu được thành các dạng ta có thể hiểu và giải thích được, tức là ta có thể tìm ra thông tin, tri thức hữu ích cho mình [1].

2. CÁC BƯỚC XỬ LÝ VĂN BẢN

Quá trình xử lý văn bản hay quá trình phân tích và kiểm tra tính chính xác của một văn bản là một vấn đề khá phức tạp, trải qua nhiều bước khác nhau. Ở mỗi bước xử lý đòi hỏi người nghiên cứu phải có một nền tảng kiến thức vững vàng về ngôn ngữ cũng như nhiều kiến thức bổ trợ khác mới có thể xử lý tốt được. Quá trình này có thể được chia thành các bước sau.

- Tiền xử lý văn bản: Sẽ xử lý sơ bộ văn bản đầu vào (làm sạch văn bản) bằng cách xóa bỏ những ký tự, những mã điều khiển, những vùng không cần thiết cho việc xử lý và phân rã nó ra thành các câu là đơn vị cơ sở của một văn bản.
- Phân tích hình thái: phân tích câu thành một bảng các từ (hay cụm từ) riêng biệt, đồng thời kèm theo tất cả các thông tin về từ đó, như là: Từ loại, phạm trù ngữ pháp, các biến cách của từ, tiền tố, hậu tố của từ (nếu có). Trong trường hợp gặp từ mới, hệ thống sẽ để nguyên và đánh dấu một từ loại đặc biệt để chuyển sang phần xử lý tên riêng hay từ mới.
- Phân tích cứu pháp: Phân tích một câu thành những thành phần văn phạm có liên quan với nhau và được thể hiện thành cây cú pháp. Khi nhập câu, ta phải phân thành các thành phần như chủ ngữ, vị ngữ; gán vai trò chủ từ, đối từ của động từ chính, bổ nghĩa,..
- Phân tích ngữ nghĩa: là kiểm tra ý nghĩa của câu có mâu thuẫn với ý nghĩa của đoạn hay không. Dựa trên mối liên hệ logic về nghĩa giữa các cụm từ trong câu và mối liên hệ giữa các câu trong đoạn, hệ thống sẽ xác định được một phần ý nghĩa của câu trong ngữ cảnh của đoạn.
- Tích hợp văn bản: Ngữ nghĩa của một câu riêng biệt có thể phụ thuộc vào những câu đứng trước, đồng thời nó cũng có thể ảnh hưởng đến các câu phía sau.
- Phân tích thực nghĩa: phân tích nhằm xác định ý nghĩa câu dựa trên mối liên hệ của câu với hiện thực. Ý nghĩa thực tế của câu phụ thuộc rất nhiều vào ý tứ của người nói và ngữ cảnh diễn ra lời nói.

3. HƯỚNG TIẾP CẬN VỚI BÀI TOÁN TÁCH TỪ

Các nhà nghiên cứu đã đề xuất một số hướng tiếp cận để giải quyết bài toán tách từ. Nhìn chung, các hướng tiếp cận đó được chia thành 2 hướng: tiếp cận dựa trên từ và tiếp cận dựa trên ký tự từ[2].

Hướng tiếp cận dựa trên từ với mục tiêu tách được các từ hoàn chỉnh trong câu. Hướng tiếp cận này được chia thành 3 nhóm: dựa trên thống kê (statistics-based), dựa trên từ điển (dictionary-based) và kết hợp nhiều phương pháp (hydrid-based).

Hướng tiếp cận dựa vào thống kê cần phải dựa vào thông tin thống kê như từ hay tần số ký tự, hay xác suất cùng xuất hiện trong một tệp dữ liệu cơ sở. Do đó, tính hiệu quả của các giải pháp này chủ yếu dựa vào dữ liệu huấn luyện cụ thể được sử dụng. Tác giả Đinh Điền [7] đã xây dựng ngữ liệu huấn luyện riêng (khoảng 10Mb) dựa vào các tài nguyên, tin tức và sách điện tử trên Internet, bộ dữ liệu này khá nhỏ và không toàn diện tức là không bao quát nhiều lĩnh vực, nhiều chủ đề.

Hướng tiếp cận dựa trên từ điển: Ý tưởng của hướng tiếp cận này là những cụm từ được tách ra từ văn bản phải được so khớp với các từ trong từ điển. Từ điển sử dụng để so khớp thì lại có 2 loại: từ điển hoàn chỉnh (full word/pharse) và từ điển thành phần (component). Trong từ điển hoàn chỉnh thì chia thành 3 loại: so khớp dài nhất (longest match), so khớp ngắn nhất (shortest match) và so khớp kết hợp (overlap). Hướng tiếp cận này có đặc điểm là đơn giản, dễ hiểu tuy nhiên hiệu quả mang lại chưa được cao. Lý do là bởi nó chưa xử lý được nhiều trường hợp nhập nhằng cũng như khả năng phát hiện từ mới trong văn bản chưa cao. Hiện nay, hướng tiếp cận so khớp cực đại được xem là phương pháp quan trọng và có hiệu quả nhất trong hướng tiếp cận từ điển.

Hướng tiếp cận nhiều phương pháp với mục đích kết hợp các phương pháp tiếp cận khác nhau để thừa hưởng các ưu điểm của nhiều kỹ thuật và hướng tiếp cận khác nhau nhằm nâng cao hiệu quả. Hướng tiếp cận này thường kết hợp giữa hướng tiếp cận thống kê và dựa trên từ điển nhằm tận dụng những mặt mạnh của các phương pháp này. Tuy nhiên, hướng tiếp cận này lại mất nhiều thời gian xử lý, không gian đĩa và chi phí cao.

Hướng tiếp cận dựa trên ký tự từ: Hướng tiếp cận này đơn thuần là rút trích ra một số lượng nhất định các tiếng trong văn bản như rút trích 1 ký tự (unigram) hay nhiều ký tự (n-gram). Phương pháp này tuy đơn giản nhưng mang lại kết quả quan trọng được chứng minh qua một số công trình nghiên cứu đã được công bố, như của tác giả Lê An Hà [3].

Trong bài báo gần đây của H.Nguyễn et al, đề xuất năm 2005. Đây là phương pháp tách từ dựa trên thống kê từ Internet và giải thuật di truyền thay vì sử dụng dữ liệu thô, để tìm ra những cách phân cách đoạn văn bản tối ưu nhất cùng một văn bản. Khi so sánh kết quả của tác giả Lê An Hà và H.Nguyễn thì thấy công trình nghiên cứu của H.Nguyễn cho được kết quả tốt hơn khi tiến hành tách từ, tuy nhiên thời gian xử lý lâu hơn. Ưu điểm của hướng tiếp cận dựa trên nhiều ký tự là tính đơn giản, dễ ứng dụng, chi phí thấp. Qua nhiều công trình nghiên cứu của các tác giả đã được công bố, hướng tiếp cận dựa trên ký tự từ được cho là sự lựa chọn thích hợp.

4. MỘT SỐ PHƯƠNG PHÁP TÁCH TỪ

4.1 Phương pháp so khớp cực đại (Maximum Matching)

Phương pháp này đã được ChihHao Tsai [4] giới thiệu năm 1996. Ý tưởng chính của phương pháp này là duyệt một câu từ trái qua phải và chọn từ có nhiều tiếng nhất có mặt trong từ điển tiếng Việt, rồi cứ thế tiếp tục cho từ kế tiếp cho đến hết câu.

Phương pháp so khớp cực đại dạng đơn giản: Giả sử chúng ta có một câu S={c1, c2, c3,...,cn} với c1, c2, c3,... cn là các tiếng được tách bởi khoảng trắng trong câu. Chúng ta bắt đầu duyệt từ đầu chuỗi, xác định đâu là từ. Trước tiên, chúng ta sẽ kiểm tra xem c1 có phải là từ có trong từ điển hay không, sau đó kiểm tra c1c2 có trong từ điển hay không. Tiếp tục như vậy c1c2c3, c1c2c3c4,..., c1c2c3... cn, với n là số tiếng lớn nhất của một từ có thể có nghĩa (nghĩa là có trong từ điển tiếng Việt). Sau đó, chúng ta chọn từ có nhiều tiếng nhất có mặt trong từ điển và đánh dấu từ đó. Tiếp tục quá trình trên với tất cả các từ còn lại trong câu và trong toàn bộ văn bản.

Phương pháp so khớp cực đại dạng phức tạp: Phương pháp này về cơ bản cũng giống như so khop cực đại dạng đơn giản. Tuy nhiên, dạng này có thể tránh được một số nhập nhằng gặp phải trong dạng đơn giản. Độ chính xác cao lên đến 99.69% và 93.21 nhập nhằng được giải quyết. Đầu tiên, chúng ta sẽ kiểm tra xem c1 có phải từ có trong từ điển hay không, sau đó kiểm tra tiếp c1c2 có nằm trong từ điển hay không. Giả sử có 1 trường hợp xảy ra như sau: ta có c1 và c1c2 đều có trong từ điển thì thuật toán thực hiện chiến thực 3 từ tốt nhất được Chen & Liu (1992) đưa ra như sau:

Độ dài trung bình của từ lớn nhất: ở cuối mỗi chuỗi thường gặp những bộ chỉ có một hoặc hai từ. Luật này chỉ có lợi khi thiếu một hoặc một vài vị trí trong bộ. Khi bộ là bộ ba thì luật này không được hiệu quả lắm. Vì bộ ba từ

có cùng tổng độ dài, đương nhiên nó sẽ có cùng độ dài chung bình. Nên giải pháp này không đạt hiệu quả cao vì thế chúng ta cần một giải pháp khác.

Sự chênh lệch độ dài của 3 từ là ít nhất: là độ biến đổi nhỏ nhất chiều dài từ. Luật này cho phép lấy bộ đầu tiên với độ biến đổi chiều dài từ nhỏ nhất. Trong ví dụ trên, ta lấy từ C1C2 từ bộ đầu tiên. Giả thiết của luật này là những từ có chiều dài đều bằng nhau.

Đánh giá phương pháp: Phương pháp so khớp cực đại là cách tách từ đơn giản, dễ hiểu và chạy nhanh. Hơn chúng ta chỉ cần một tập từ điển đầy đủ là có thể tiến hành tách các văn bản. Tuy nhiên, phương pháp này không giải quyết 2 vấn đề quan trọng của bài toán tách từ tiếng Việt là thuật toán gặp phải nhiều nhập nhằng; độ chính xác của phương pháp này phụ thuộc vào tính đầy đủ và tính chính xác của từ điển.

4.2 Phương pháp chuyển dịch trạng thái hữu hạng có trọng số

Chuyển dịch trạng thái hữu hạn có trọng số (Weighted Finite-State Transducer – WFST) [5]. Mô hình chuyển dịch trạng thái hữu hạn có trọng số WFST đã được đề xuất năm 1996. Ý tưởng chính của phương pháp này áp dụng cho tách từ tiếng Việt là các từ sẽ được gán trọng số bằng xác suất xuất hiện của từ đó trong dữ liệu. Sau đó duyệt qua các câu, cách duyệt có trọng số lớn nhất sẽ là cách dùng để tách từ. Trong phương pháp này, tầng tiền xử lý có nhiệm vụ xử lý định dạng văn bản: Tiêu đề, đoạn, câu; chuẩn hoá về chính tả tiếng Việt (cách bỏ dấu, cách viết các ký tự y, i,... trong tiếng Việt). Ví dụ: Vật lý = vật lí, thời kỳ = thời kì).

Sau đó câu được chuyển sang tầng WFST. Trong tầng này tác giả xử lý thêm các vấn đề liên quan đến đặc thù của tiếng Việt, như: Từ láy, tên riêng, ...và tầng mạng Neural dùng để khử nhập nhằng về ngữ nghĩa sau khi đã tách từ (nếu có).

Sơ đồ các bước xử lý của WFST

Xét tầng WFST:

Hoạt động của WFST có thể chia thành ba bước sau:

- Bước 1: Xây dựng từ điển trọng số, trong mô hình WFST, thì việc phân đoạn từ có thể được xem như là một sự chuyển dịch trạng thái có xác xuất. Chúng ta miêu tả từ điển D là một đồ thị biến đổi trạng thái hữu hạn có trọng số.

- Bước 2: Xây dựng khả năng tách từ, bước này thống kê tất cả các khả năng tách từ của một câu. Vấn đề ở đây là để giảm sự bùng nổ các cách tách từ, thuật toán sẽ loại bỏ ngay những nhánh tách từ nào đó không phù hợp mà chứa từ không xuất hiện trong từ điển, không phải là từ láy, không phải là danh từ riêng thì loại bỏ các nhánh xuất phát từ cách tách từ đó. Thật vậy, giả sử một câu gồm n âm tiết, mà trong tiếng Việt thì một từ có tối đa 4 âm tiết tức là ta sẽ có tối đa 2n-1 cách tách từ khác nhau. Một câu tiếng Việt trung bình có 24 âm tiết thì lúc đó ta phải giải quyết 8.000.000 trường hợp tách từ có thể trong một câu.

- Bước 3: Lựa chọn khả năng tối ưu: Sau khi liệt kê tất cả các khả năng tách từ, thuật toán sẽ chọn cách tách tốt nhất, đó là tách đoạn có trọng số bé nhất.

Xét ví dụ sau: Đầu vào là câu: "Tốc độ truyền thông tin sẽ tăng cao":

Trọng số theo mỗi cách tách từ được tính là:

Id(I).D* = "Tốc độ # truyền thông # tin # sẽ # tăng # cao" (1)

= 8.68 +12.31+7.33+6.09+7.43+6.95=48.79

ID(I).D* = "TỐC ĐỘ # TRUYỀN # THÔNG TIN # SẼ # TĂNG # CAO" (2)

= 8.68 +12.31+7.24+6.09+7.43+6.95=48.70

Khi đó ta có được cách tối ưu là tách đoạn (2) "Tốc độ # truyền # thông tin # sẽ # tăng # cao".

Xét tầng Neural:

Sau khi cho câu được tách từ qua mô hình WFST. Để xác định kết quả tách từ trên có thực sự hợp lệ hay không, tác giả định nghĩa một ngưỡng giá trị t0 với ý nghĩa như sau: nếu sự chênh lệch về trọng số (giữa các cách tách từ khác nhau với cách tách từ có trọng số nhỏ nhất) lớn hơn t0 thì đó là kết quả tách từ có trọng số nhỏ nhất đó đúng của câu và được chấp nhận. Còn nếu sự chênh lệch đó không lớn hơn t0, thì cách tách từ có trọng số nhỏ nhất đó chưa được xem là kết quả tách từ đúng của câu. Lúc này, ta sẽ đưa những cách tách từ của câu này qua mô hình mạng Neural để xử lý tiếp.

Tầng nhập của mạng được kết nối hoàn toàn với một tầng ẩn gồm 10 nút với một hàm truyền. Những nút ẩn này lại được kết nối hoàn toàn với một tầng xuất chỉ gồm 1 nút. Nút xuất là một giá trị thực nằm giữa 0..1. Biểu thị cho khả năng hợp lệ của một dãy các từ loại đứng liền nhau trong một cửa

sổ trượt. Khi cửa sổ trượt từ đầu câu đến cuối câu, cộng dồn các kết quả lại với nhau và gán giá trị này vào thành trọng số của câu. Câu có trọng số lớn nhất sẽ được chọn.

Hàm truyền được chọn là hàm sigmoid:

Đánh giá phương pháp: phương pháp này là sẽ cho độ chính xác cao nếu ta xây dựng được một dữ liệu học đầy đủ và chính xác. Nó còn có thể kết hợp với các phương pháp khử nhập nhằng để cho kết quả tách rất cao (có thể chính xác đến 97%, tỉ lệ này tuỳ thuộc vào loại văn bản). Tuy nhiên, việc đánh trọng số dựa trên tần số xuất hiện của từ, nên khi tiến hành tách thì không tránh khỏi các nhập nhằng trong tiếng Việt. Hơn nữa với những văn bản dài thì phương pháp này còn gặp phải sự bùng nổ các khả năng phân đoạn của từng câu.

4.3 Phương pháp mô hình Markov ẩn

Mô hình Markov (Hidden Markov Model - HMM) được giới thiệu vào cuối những năm 1960 [6]. Cho đến hiện nay phương pháp này có một ứng dụng khá rộng như trong nhận dạng giọng nói, tính toán sinh học và xử lý ngôn ngữ tự nhiên. Mô hình Markov là mô hình trạng thái hữu hạn với các tham số biểu diễn xác suất chuyển trạng thái và xác suất sinh dữ liệu quan sát tại mỗi trạng thái.

Mô hình Markov ẩn là mô hình thống kê trong đó hệ thống được mô hình hóa được cho là một quá trình Markov với các tham số không biết trước và nhiệm vụ là xác định các tham số ẩn từ các tham số quan sát được, dựa trên sự thừa nhận này. Các tham số của mô hình được rút ra sau đó có thể sử dụng để thực hiện các phân tích kế tiếp.

Trong một mô hình Markov điển hình, trạng thái được quan sát trực tiếp bởi người quan sát, vì vậy các xác suất chuyển tiếp trạng thái là các tham số duy nhất. Mô hình Markov ẩn thêm vào các đầu ra, mỗi trạng thái có xác suất phân bổ trên các biểu hiện đầu ra có thể. Vì vậy, nhìn vào dãy của các biểu hiện được sinh ra bởi HMM không trực tiếp chỉ ra dãy các trạng thái.

Các thông số trong mô hình: xi: các trạng thái trong mô hình Markov, aij: Các xác suất chuyển tiếp, bij: các xác suất đầu ra, yi: Các dữ liệu quan sát.

Mô hình Markov ẩn thêm vào các đầu ra, mỗi trạng thái có xác suất phân bố trên các biểu hiện đầu ra có thể. Vì vậy, nhìn vào dãy của các biểu hiện được sinh ra bởi HMM không trực tiếp chỉ ra dãy các trạng thái. Ta có thể

tìm ra được chuỗi các trạng thái mô tả tốt nhất cho mỗi dữ liệu quan sát đước bằng cách tính: P(Y | X) = P(Y | X) / P(X).

Trong khi đó Yn là trạng thái thời điểm thứ t = n trong chuỗi trạng thái Y, Xn là dữ liệu quan sát được tại thời điểm thứ t = n trong chuỗi X. Do trạng thái hiện tại chỉ phụ thuộc vào trạng thái ngay trước đó với giả thiết rằng dữ liệu quan sát được tại thời điểm t chỉ phụ thuộc vào trạng thái t. Ta có thể tính P(Y, X) theo công thức:

Đánh giá phương pháp: mô hình Markov để tính được xác suất P(Y,X) thông thường ta phải liệt kê hết các trường hợp có thể của chuỗi Y và chuỗi X. Thực tế thì chuỗi Y là hữu hạn có thể liệt kê được, còn X (các dữ liệu quan sát) là rất phong phú. Để giải quyết các vấn đề này HMM đưa ra giả thiết về sự độc lập giữa các dữ liệu quan sát. Dữ liệu quan sát được tại thời điểm t chỉ phụ thuộc vào trạng thái tại thời điểm đó. Hạn chế thứ hai gặp phải là việc sử dụng xác suất đồng thời P (Y, X) đôi khi không chính xác vì với một số bài toán thì việc sử dụng xác suất điều kiện P(Y|X) cho kết quả tốt hơn rất nhiều.

4.4 Phương pháp so khớp từ dài nhất (Longest Matching):

Phương pháp so khớp từ dài nhất [7] dựa trên tư tưởng tham lam. Với mỗi câu, duyệt từ trái qua phải các âm tiết trong câu, kiểm tra xem có nhóm các âm tiết có tồn tại 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. Thuật toán chỉ đúng khi không có sự nhập nhằng những tiếng đầu của từ sau có thể ghép với từ trước tạo thành một từ có trong từ điển.

Giải thuật

Input: Chuỗi ký tự;

Output: Chuỗi từ, cụm từ (từ có chiều dài dài nhất);

V là danh sách các tiếng chưa xét;

T là bộ từ điển.

While V ≠ Ø do

Begin

Wmax= từ đầu danh sách V;

Foreach (v thuộc từ gồm các tiếng bắt đầu trong V)

If(length(v) > length(Wmax) and (v thuộc T))

Then Wmax=v;

Loại bỏ đi các từ Wmax ở đầu danh sách V;

End.

Xét ví dụ: “Tôi là sinh viên trường Đại học Kiên Giang”:

Đánh giá phương pháp: phương pháp so khớp từ dài nhất là phương pháp tách từ đơn giản chỉ cần dựa vào từ điển với đội chính xác tương đối cao. Phương pháp này sẽ không tốt nếu có hiện tượng nhập nhằng xảy ra. Độ chính xác phụ thuộc hoàn toàn vào tính đầy đủ và chính xác của từ điển.

5. THỰC NGHIỆM – KẾT QUẢ

5.1 Thực nghiệm

Mục tiêu nghiên cứu là đánh giá tính khả thi của thuật toán tách từ dựa vào phương pháp so khớp cực đại (Maximum Matching) để tách toàn bộ văn bản, với dữ liệu sử dụng là bảng âm tiết tiếng Việt và từ điển từ vựng tiếng Việt. Phần mềm Vntokenizer là phần mềm tách từ biểu diễn cho phương pháp so khớp cực đại.

Quy trình thực hiện tách từ theo phương pháp so khớp cực đại (Vntokenizer).

Input: của phần mềm Vntokenizer là một câu, một văn bản được lưu dưới dạng tệp (.txt).

Output: là một chuỗi các từ đã được tách ở dạng file (.txt).

Thuật toán sẽ duyệt từ đầu chuỗi xác định đâu là từ. Đầu tiên, ta kiểm tra xem nó có trong từ điển hay không, sau đó kiểm tra tiếp chữ kế có trong từ điển hay không, nếu chữ đầu tiên và chữ kế tiếp có trong kho dữ liệu thì chương trình sẽ đọc chữ tiếp theo, quá trình đó sẽ lập lại cho đến khi đọc chữ tiếp theo mà dãy chữ đó không có trong từ điển thì sẽ dừng lại và lấy từ. Tức là chương trình sẽ duyệt một câu từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ điển và đánh dấu từ đó. Sau đó, tiếp tục quá trình trên với tất cả các từ kế tiếp cho đến hết câu.

Các đơn vị từ bao gồm các từ trong từ điển cũng như các chuỗi số, chuỗi ký tự từ nước ngoài, các dấu câu và các chuỗi kí tự hỗn tạp khác trong văn bản. Các đơn vị từ không chỉ bao gồm các từ có trong từ điển mà còn có thể là các từ mới hoặc các từ được sinh tự do theo một quy tắc nào đó (như phương thức thêm hậu tố hay phương thức láy) hoặc các chuỗi kí hiệu không được liệt kê trong từ điển.

5.2 Kết quả

Kết quả phân tích sau khi áp dụng phương pháp so khớp cực đại, bằng cách chạy phần mềm Vntokenizer như sau: với file dữ liệu tên là 1.txt.

Giao diện màn hình khi chạy chương trình

Kết quả sau khi tách từ

Tính số lần xuất hiện của các từ

6. KẾT LUẬN

Trong bài báo này chúng tôi giới thiệu các thuật toán tách từ trong xử lý ngôn ngữ tự nhiên dựa vào phương pháp so khớp cực đại, phương pháp mô hình MarKov ẩn, phương pháp chuyển dịch trạng thái hữu hạn có trọng số, phương pháp so khớp từ dài nhất. Kết quả mà chúng tôi thu được từ nghiên cứu này là hết sức khả quan và thiết nghĩ là hoàn toàn khả thi khi ứng dụng vào thực tế.

Trên cơ sở các phương pháp tách từ đó, bài báo này sử dụng phương pháp so khớp cực đại để minh họa cho các phương pháp tách từ trên. Phần mềm vntokenizer là phần mềm tách từ để biểu diễn cho thuật toán trên. Sau khi tách từ xong, tính tầng số xuất hiện các từ được tách trong tệp dữ liệu ban đầu. Đề xuất mô hình phân loại văn bản ứng dụng vào phân loại văn bản ở phòng ban, và thư viện trường Đại học Kiên Giang. Mặc dù kết quả nghiên cứu bước đầu đã khẳng định thuật toán tách từ dựa vào phương pháp so khớp cực đại là hoàn toàn khả thi và hoàn toàn có thể áp dụng vào thực tế, tuy nhiên Bài báo chỉ mới tách được các từ trong văn bản và tính được số lần xuất hiện của từ trong đoạn văn nhằm xây dựng vector đặc trưng từ đó hình thành ma trận để so sánh với từ khóa cho trước như: công nghệ thông tin, nông nghiệp,... từ đó làm cơ sở cho bài toán phân loại văn bản.

TÀI LIỆU THAM KHẢO

[1] Hồ Tú Bảo, Lương Chi Mai (2008), Về xử lý tiếng Việt trong công nghệ thông tin, Viện Công nghệ Thông tin, Viện Khoa học và Công nghệ tiên tiến Nhật Bản.

[2] Trần Thị Oanh, Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt, Luận văn thạc sĩ trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[3] Le An Ha (2003), A method for word segmentation in Vietnamese. In Proceedings of Curpus Linguistics. Lancaster, UK.

[4] Chih-Hao Tsai (2000). MMSEG: A word identification system for Mandarin Chinese text Based on two variants of the Maximum Matching Algorithm.

[5] Dinh Dien, Hoang Kiem, Nguyen Van Toan (2001), Vietnamese Word Segmentation, the sexth 6th Natural Language Processing Pacific Rim Symposium Tokyo, Japan, pp.749 – 756.

[6] Phil Blunsom (2004), Hidden Markov Medels, pp. 1-7

[7] Chen,K.J.,&Liu, S. H. (1992), Word identification for Mandarin Chinese sentences. Proceedings of the fifteenth inrernational Conference on computational Linguistics, Nantes.