ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

**KHOA** TOÁN – CƠ – TIN HỌC

**Phùng Văn Thành**

**XÂY DỰNG MỘT HỆ THỐNG XỬ LÝ NGÔN NGỮ TỰ NHIÊN CHO VĂN BẢN TIẾNG VIỆT**

Khóa luận tốt nghiệp đại học hệ chính quy

Ngành: Máy tính và khoa học thông tin

(Chương trình đào tạo: chuẩn)

**Hà Nội – 2018**

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

**KHOA** TOÁN – CƠ – TIN HỌC

**Phùng Văn Thành**

**XÂY DỰNG MỘT HỆ THỐNG XỬ LÝ NGÔN NGỮ TỰ NHIÊN CHO VĂN BẢN TIẾNG VIỆT**

Khóa luận tốt nghiệp đại học hệ chính quy

Ngành: Máy tính và khoa học thông tin

(Chương trình đào tạo: chuẩn)

**Cán bộ hướng dẫn: Nguyễn Việt Hùng**

**Hà Nội – 2018**

# Lời cảm ơn

Trước hết cho phép em được gửi lời cảm ơn sâu sắc nhất tới người hướng dẫn đã tận tình giúp đỡ và tạo điều kiện để em có thể hoàn thành khóa luận tốt nghiệp này một cách tốt nhất có thể.

Cũng cho phép em được gửi lời cảm ơn chân thành nhất tới các thầy cô trong bộ môn tin học nói riêng và Khoa Toán – Cơ – Tin học trường Đại học Khoa học Tự Nhiên nói chung đã dạy dỗ, giúp đỡ và cho em một môi trường có thể thoải mái sáng tạo và phát triển bản thân trong suốt quá trình học tập tại trường.

Xin được gửi lời cảm ơn chân thành tới các anh chị, bạn bè, gia đình … những người đã luôn sát cánh bên tôi trong suốt năm học qua.

# Mục lục

[Lời cảm ơn 3](#_Toc514227260)

[Mục lục 4](#_Toc514227261)

[Mở đầu 5](#_Toc514227262)

[Chương 1. Đặt vấn đề 7](#_Toc514227263)

[1.1 Xử lý ngôn ngữ tự nhiên 7](#_Toc514227264)

[1.2 Các bước xử lý 7](#_Toc514227265)

[1.3 Các bài toán và ứng dụng 7](#_Toc514227266)

[1.4 Các bài toán trong xử lý tiếng Việt 9](#_Toc514227267)

[1.5 Thực trạng hiện nay và hướng giải quyết 10](#_Toc514227268)

[Chương 2. Cơ sở lý thuyết 11](#_Toc514227269)

[2.1 Bài toán tách từ 11](#_Toc514227270)

[2.2 Bài toán gán nhãn từ loại 13](#_Toc514227271)

[2.3 Bài toán phân tích cú pháp 15](#_Toc514227272)

[2.4 Bài toán phân tích tình cảm 16](#_Toc514227273)

# Mở đầu

Trong thời đại công nghệ 4.0 hiện nay, mọi thứ đều đang thiên hướng về việc xử lý các vấn đề liên quan đến internet vạn vật, xử lý ảnh, học máy, trí tuệ nhân tạo… Và một trong số đó không thể không nhắc đến việc xử lý ngôn ngữ tự nhiên. Đặc biệt là ở Việt Nam hiện nay, vấn đề này luôn được các nhà nghiên cứu cũng như số lượng lớn người quan tâm và tìm hiểu. Mặc dù đã có những mô hình xử lý các bài toán ngôn ngữ tự nhiên từ cơ bản cho đến nâng cao đã thành công ở nước ngoài. Tuy nhiên, việc áp dụng các mô hình đó vào việc xử lý ngôn ngữ tự nhiên ở Việt Nam hiện vẫn đang còn gặp nhiều khó khăn do sự bất đồng về ngôn ngữ cũng như về mặt cấu trúc ngữ pháp và ngữ nghĩa đa dạng của tiếng Việt. Và điều đặc biệt hơn cả là việc thiếu hụt về dữ liệu cũng như một hệ thống tích hợp các công cụ xử lý ngôn ngữ tự nhiên được công bố miễn phí cho cộng đồng. Nhận thấy được điều này, em đã nhận được sự chỉ dẫn từ người hướng dẫn và đã quyết định xây dựng nên một trang web trên đó tích hợp các công cụ cũng như các hệ thống xử lý độc lập để phục vụ cho việc xử lý ngôn ngữ tự nhiên cho văn bản tiếng Việt.

Trong khóa luận này, mục tiêu mà em đề ra đó là xây dựng một trang web cho phép người dùng có quyền đăng tải tài liệu lên, sử dụng các bộ công cụ được thiết kế thuận tiện cho việc chỉnh sửa và quản trị được các kho ngữ liệu của riêng mình. Ngoài ra các người dùng khác khi muốn tải tài liệu về thì phải đăng ký tài khoản trên hệ thống và phải gửi đơn xin phép đến người dùng đăng tải, khi đã được cấp phép người dùng đó sẽ được tải tài liệu về. Hệ thống còn tích hợp thêm cả các bộ công cụ về thu thập dữ liệu từ nhiều nguồn khác nhau trên các website như wikipedia, foody.vn, facebook,… Mục đích của việc thu thập dữ liệu như ở trên đã nói đó là việc phục vụ cho cộng đồng có thể chỉnh sửa và tải dữ liệu về. Ngoài ra còn là đầu vào cho các bài toán xử lý ngôn ngữ tự nhiên như bài toán tách từ, gán nhãn từ loại, phân tích cú pháp, phân tích tình cảm, xác định thực thể tên… Hệ thống không chỉ thu thập các kho dữ liệu trên các trang web mà còn chia sẻ một lượng lớn các kho ngữ liệu sẵn có như: kho ngữ liệu hai triệu âm tiết đã tách từ, mười nghìn câu được gán nhãn từ loại, mười nghìn câu ngữ pháp. Hơn thế nữa, website còn tích hợp cả công cụ bán tự động cho phép xử lý các bài toán tách từ, gán nhãn từ loại. Về cơ bản sẽ đáp ứng được nhu cầu cho nhiều người làm việc cùng một lúc do việc xử lý được gọi từ một máy chủ độc lập. Trang web được xây dựng đầy đủ các phần từ việc quản trị các bộ ngữ liệu cho đến việc hiển thị giao diện ở phía người dùng có thể tương tác được. Nếu có thêm thời gian và điều kiện thì em cũng muốn tích hợp thêm hệ thống thanh toán điện tử như paypal, nganluong,… để phục vụ cho việc giao dịch mua bán giữa các người dùng với nhau hoặc cũng có thể quy định chung một đơn vị tiền ảo của bên phát triển hệ thống để coi như là một phần phí nào đấy đáp ứng cho việc duy trì và phát triển hệ thống. Đây cũng tạm gọi là một sân chơi chung để mọi người có thể trao đổi cũng như kiếm thêm thu nhập cho việc làm mịn các kho ngữ liệu của mình.

Về mặt cấu trúc, báo cáo được chia thành 4 chương:

Chương 1: Đặt vấn đề

Chương này trước hết trình bày khái niệm cơ bản về xử lý ngôn ngữ tự nhiên, ngoài ra còn giới thiệu khái quát về ý tưởng cũng như ý nghĩa của một số bài toán xử lý ngôn ngữ tự nhiên cơ bản. Tiếp đến là nêu lên thực trạng các khó khăn hiện nay của việc nghiên cứu xử lý ngôn ngữ tự nhiên bằng tiếng Việt. Từ đó vạch định ra những công việc cần phải làm để giải quyết những vấn đề còn tồn đọng.

Chương 2: Cơ sở lý thuyết

Chương này sẽ trình bày về hướng tiếp cận cũng như các cách thức để giải quyết các bài toán. Với mỗi bài toán sẽ tập trung đi vào mục tiêu đạt được và trình bày khái quát cơ sở lý thuyết cho việc giải quyết bài toán đó.

Chương 3: Xây dựng hệ thống

Dựa vào các cơ sở lý thuyết ở chương trước, chương này sẽ tập trung vào việc phân tích, thiết kế hệ thống và từ đó đưa ra cái nhìn tổng quan về cả hệ thống. Cấu trúc các thành phần của hệ thống, cách thức hoạt đông, ý nghĩa cũng như thao tác người dùng thực hiện. Bên cạnh đó, cũng trình bày rõ quá trình triển khai các công nghệ áp dụng vào hệ thống.

Chương 4: Kết quả

Nêu lên những nội dung đã đạt được tính đến thời điểm hiện tại, ưu điểm, nhược điểm, cách khắc phục. Ngoài ra đề xuất hướng phát triển để cải thiện hiệu năng cũng như tính năng của hệ thống.

# Chương 1. Đặt vấn đề

1. Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh của Trí tuệ nhân tạo, tập trung vào việc nghiên cứu sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người. Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người như: tương tác giữa người và máy, cải thiện hiệu quả giao tiếp giữa con người với con người, hoặc đơn giản là nâng cao hiệu quả xử lý văn bản và lời nói.

Trong xử lý ngôn ngữ tự nhiên có 2 quan điểm cơ bản:

1. Xử lý các từ ngữ bằng máy tính.

2. Làm cho máy tính hiểu được các từ ngữ.

Hiện tại, cả 2 hướng này đều đang được tích cực nghiên cứu và phát triển, nhờ đó rất nhiều các hệ thống hiệu quả đã và đang được tạo ra.

1. Các bước xử lý

**Phân tích hình thái** - Trong bước này từng từ sẽ được phân tích và các ký tự không phải chữ (như các dấu câu) sẽ được tách ra khỏi các từ. Trong tiếng Anh và nhiều ngôn ngữ khác, các từ được phân tách với nhau bằng dấu cách. Tuy nhiên trong tiếng Việt, dấu cách được dùng để phân tách các tiếng (âm tiết) chứ không phải từ. Cùng với các ngôn ngữ như tiếng Trung, tiếng Hàn, tiếng Nhật phân tách từ trong tiếng Việt là một công việc không hề đơn giản.

**Phân tích cú pháp** - Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ này. Sẽ có những dãy từ bị loại do vi phạm các luật văn phạm.

**Phân tích ngữ nghĩa** - Thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bộ phân tích cú pháp.

**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** - Cấu trúc thể hiện điều được phát ngôn sẽ được thông dịch lại để xác định nó thật sự có nghĩa là gì.

Tuy nhiên, ranh giới giữa 5 bước xử lý này cũng rất mong manh. Chúng có thể được tiến hành từng bước một, hoặc tiến hành cùng lúc - tùy thuộc vào giải thuật và ngữ cảnh cụ thể.

1. Các bài toán và ứng dụng

**Nhận dạng chữ viết**: Có hai kiểu nhận dạng, thứ nhất là nhận dạng chữ in, ví dụ nhận dạng chữ trên sách giáo khoa rồi chuyển nó thành dạng văn bản điện tử như dưới định dạng doc của Microsoft Word chẳng hạn. Phức tạp hơn là nhận dạng chữ viết tay, có khó khăn bởi vì chữ viết tay không có khuôn dạng rõ ràng và thay đổi từ người này sang người khác. Với chương trình nhận dạng chữ viết in có thể chuyển hàng ngàn đầu sách trong thư viện thành văn bản điện tử trong thời gian ngắn. Nhận dạng chữ viết của con người có ứng dụng trong khoa học hình sự và bảo mật thông tin (nhận dạng chữ ký điện tử).

**Nhận dạng tiếng nói:** Nhận dạng tiếng nói rồi chuyển chúng thành văn bản tương ứng. Giúp thao tác của con người trên các thiết bị nhanh hơn và đơn giản hơn, chẳng hạn thay vì gõ một tài liệu nào đó bạn đọc nó lên và trình soạn thảo sẽ tự ghi nó ra. Đây cũng là bước đầu tiên cần phải thực hiện trong ước mơ thực hiện giao tiếp giữa con người với robot. Nhận dạng tiếng nói có khả năng trợ giúp người khiếm thị rất nhiều.

**Tổng hợp tiếng nói:** Từ một văn bản tự động tổng hợp thành tiếng nói. Thay vì phải tự đọc một cuốn sách hay nội dung một trang web, nó tự động đọc cho chúng ta. Giống như nhận dạng tiếng nói, tổng hợp tiếng nói là sự trợ giúp tốt cho người khiếm thị, nhưng ngược lại nó là bước cuối cùng trong giao tiếp giữa robot với người.

**Dịch từ động** (machine translate): Như tên gọi đây là chương trình dịch tự động từ ngôn ngữ này sang ngôn ngữ khác. Một phần mềm điển hình về tiếng Việt của chương trình này là Evtrans của Softex, dịch tự động từ tiếng Anh sang tiếng Việt và ngược lại, phần mềm từng được trang web vdict.com mua bản quyền, đây cũng là trang đầu tiên đưa ứng dụng này lên mạng. Tháng 10 năm 2008 có hai công ty tham gia vào lĩnh vực này cho ngôn ngữ tiếng Việt là công ty Lạc Việt (công ty phát hành từ điển Lạc Việt) và Google, một thời gian sau đó Xalo\_vn cũng đưa ra dịch vụ tương tự.

**Tìm kiếm thông tin** (information retrieval): Đặt câu hỏi và chương trình tự tìm ra nội dung phù hợp nhất. Thông tin ngày càng đầy lên theo cấp số nhân, đặc biệt với sự trợ giúp của internet việc tiếp cận thông tin trở lên dễ dàng hơn bao giờ hết. Việc khó khăn lúc này là tìm đúng nhất thông tin mình cần giữa bề bộn tri thức và đặc biệt thông tin đó phải đáng tin cậy. Các máy tìm kiếm dựa trên giao diện web như Google hay Yahoo hiện nay chỉ phân tích nội dung rất đơn giản dựa trên tần suất của từ khoá và thứ hạng của trang và một số tiêu chí đánh giá khác để đưa ra kết luận, kết quả là rất nhiều tìm kiếm không nhận được câu trả lời phù hợp, thậm chí bị dẫn tới một liên kết không liên quan gì do thủ thuật đánh lừa của các trang web nhằm giới thiệu sản phẩm (có tên tiếng Anh là SEO viết tắt của từ search engine optimization). Thực tế cho đến bây giờ chưa có máy tìm kiếm nào hiểu được ngôn ngữ tự nhiên của con người trừ trang www.ask.com được đánh giá là "hiểu" được những câu hỏi có cấu trúc ở dạng đơn giản nhất. Mới đây cộng đồng mạng đang xôn xao về trang Wolfram Alpha, được hứa hẹn là có khả năng hiểu ngôn ngữ tự nhiên của con người và đưa ra câu trả lời chính xác. Lĩnh vực này hứa hẹn tạo ra bước nhảy trong cách thức tiếp nhận tri thức của cả cộng đồng.

**Tóm tắt văn bản:** Từ một văn bản dài tóm tắt thành một văn bản ngắn hơn theo mong muốn nhưng vẫn chứa những nội dung thiết yếu nhất.

**Khai phá dữ liệu** (data mining) và phát hiện tri thức: Từ rất nhiều tài liệu khác nhau phát hiện ra tri thức mới. Thực tế để làm được điều này rất khó, nó gần như là mô phỏng quá trình học tập, khám phá khoa học của con người, đây là lĩnh vực đang trong giai đoạn đầu phát triển. Ở mức độ đơn giản khi kết hợp với máy tìm kiếm nó cho phép đặt câu hỏi để từ đó công cụ tự tìm ra câu trả lời dựa trên các thông tin trên web mặc cho việc trước đó có câu trả lời lưu trên web hay không (giống như trang Yahoo! hỏi và đáp, nơi chuyên đặt các câu hỏi để người khác trả lời), nói một cách nôm na là nó đã biết xử lý dữ liệu để trả lời câu hỏi của người sử dụng, thay vì máy móc đáp trả những gì chỉ có sẵn trong bộ nhớ.

1. Các bài toán trong xử lý tiếng Việt

**Bài toán Tách từ:** Tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép… có trong câu. Đối với xử lý ngôn ngữ, để có thể xác định cấu trúc ngữ pháp của câu, xác định từ loại của một từ trong câu, yêu cầu nhất thiết đặt ra là phải xác định được đâu là từ trong câu. Vấn đề này tưởng chừng đơn giản với con người nhưng đối với máy tính, đây là bài toán rất khó giải quyết.[1]

**Bài toán Gán nhãn từ loại:** Một trong các vấn đề nền tảng của phân tích ngôn ngữ là việc phân loại các từ thành các lớp từ loại dựa theo thực tiễn hoạt động ngôn ngữ. Mỗi từ loại tương ứng với một hình thái và một vai trò ngữ pháp nhất định. Các bộ chú thích từ loại có thể thay đổi tuỳ theo quan niệm về đơn vị từ vựng và thông tin ngôn ngữ cần khai thác trong các ứng dụng cụ thể [19]. Mỗi từ trong một ngôn ngữ nói chung có thể gắn với nhiều từ loại, và việc giải thích đúng nghĩa một từ phụ thuộc vào việc nó được xác định đúng từ loại hay không. Công việc gán nhãn từ loại cho một văn bản là xác định từ loại của mỗi từ trong phạm vi văn bản đó.[2]

**Bài toán phân tích cú pháp:** Việc phân tích cú pháp câu có thể chia làm hai mức chính. Mức thứ nhất là tách từ và xác định thông tin từ loại. Mức thứ hai là sinh cấu trúc cú pháp cho câu dựa trên các từ và từ loại do bước trước cung cấp

**Bài toán phân tích tình cảm:** Phân tích tình cảm nhằm mục đích xác định thái độ của người nói, nhà văn, hoặc chủ đề khác liên quan đến một số chủ đề hoặc phân cực ngữ cảnh tổng thể hoặc phản ứng cảm xúc với tài liệu, tương tác hoặc sự kiện. Thái độ có thể là một phán quyết hoặc đánh giá, trạng thái cảm xúc (có nghĩa là trạng thái cảm xúc của tác giả hoặc diễn giả), hoặc giao tiếp cảm xúc dự định (có nghĩa là, tác động tình cảm của tác giả hoặc người đối thoại).[3]

**Bài toán phân tích cú pháp phụ thuộc:** Cú pháp phụ thuộc là cấu trúc cú pháp chứa các mục từ vựng nối với nhau bởi các quan hệ nhị phân không đối xứng gọi là sự phụ thuộc [4]. Quan hệ phụ thuộc này có thể được đặt tên để làm rõ liên hệ giữa hai mục từ.

1. Thực trạng hiện nay và hướng giải quyết

Về thực trạng tình hình nghiên cứu xử lý ngôn ngữ tự nhiên tiếng Việt hiện nay ở Việt Nam còn gặp 1 số khó khăn như: không có dữ liệu đủ lớn, thiếu những nghiên cứu nền tảng, ngoài ra không thể không nhắc đến việc khác nhau về cấu trúc, ngữ pháp, ngữ nghĩa giữa tiếng Việt và tiếng Anh nói riêng cũng như các ngôn ngữ khác nói chung do vậy dẫn đến sự nhập nhằng về mặt ngữ nghĩa, ngữ cảnh mà nó xuất hiện.  Lớn nhất phải kể đến sự khó khăn về nhân sự, những người nắm giữ những kiến thức về xử lý ngôn ngữ tự nhiên tiếng Việt quả thực không nhiều, và cũng không có được 1 mạng lưới liên kết, trao đổi và hỗ trợ một cách hiệu quả.

Với những khó khăn kể ra ở trên, em mong muốn góp một phần nhỏ nào đấy trong việc xây dựng một hệ thống tích hợp các công cụ xử lý ngôn ngữ tự nhiên cho tiếng Việt. Ở đấy chia sẻ một số kho ngữ liệu sẵn có và một hệ thống các công cụ giúp cho người dùng có thể làm việc cùng nhau qua mạng, thu thập dữ liệu và xử lý dữ liệu với các thao tác đơn giản, dễ sử dụng trên giao diện web.

Tuy nhiên hệ thống vẫn còn 1 số giới hạn như: chỉ hỗ trợ 1 số công cụ về việc tách từ, gán nhãn từ loại, phân tích cú pháp, phân tích tình cảm. Do khả năng thực thi của em còn giới hạn cũng như thời gian khá gấp nên còn nhiều sai sót không tránh khỏi.

# Chương 2. Cơ sở lý thuyết

1. Bài toán tách từ

Tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép… có trong câu. Đối với xử lý ngôn ngữ, để có thể xác định cấu trúc ngữ pháp của câu, xác định từ loại của một từ trong câu, yêu cầu nhất thiết đặt ra là phải xác định được đâu là từ trong câu. Vấn đề này tưởng chừng đơn giản với con người nhưng đối với máy tính, đây là bài toán rất khó giải quyết.  
Chính vì lý do đó tách từ được xem là bước xử lý quan trọng đối với các hệ thống Xử Lý Ngôn Ngữ Tự Nhiên, đặc biệt là đối với các ngôn ngữ thuộc vùng Đông Á theo loại hình ngôn ngữ đơn lập, ví dụ: tiếng Trung Quốc, tiếng Nhật, tiếng Thái, và tiếng Việt. Với các ngôn ngữ thuộc loại hình này, ranh giới từ không chỉ đơn giản là những khoảng trắng như trong các ngôn ngữ thuộc loại hình hòa kết như tiếng Anh…, mà có sự liên hệ chặt chẽ giữa các tiếng với nhau, một từ có thể cấu tạo bởi một hoặc nhiều tiếng. Vì vậy đối với các ngôn ngữ thuộc vùng Đông Á, vấn đề của bài toán tách từ là khử được sự nhập nhằng trong ranh giới từ.

Để hiểu đơn giản hơn về bài toán, ta có thể lấy 1 ví dụ đơn giản sau:

* Input: Tôi đã mua quyển sách mà thầy giáo giới thiệu.
* Output: Tôi đã mua quyển sách mà thầy\_giáo giới\_thiệu .

Như chúng ta đã biết, văn bản tiếng Việt đặt dấu cách giữa các âm tiết chứ không phải giữa các từ. Một từ có thể có một, hai hoặc nhiều âm tiết nên có nhiều cách phân chia các âm tiết thành các từ, gây ra nhập nhằng. Việc phân giải nhập nhằng này gọi là bài toán tách từ. Tiêu chí quan trọng nhất trong bài toán tách từ đương nhiên là độ chính xác. Hiện tại người ta đã đạt được độ chính xác lên đến 97% tính theo từ. Tuy nhiên nếu tính theo câu (số câu được tách hoàn toàn đúng/tổng số câu) thì độ chính xác chỉ khoảng 50%. Đây là vấn đề nghiêm trọng đối với các bước xử lý sau như phân tích ngữ pháp, ngữ nghĩa vì một từ bị tách sai có ảnh hưởng toàn bộ đến cách phân tích cả câu.

Ngoài ra tiêu chí độ chính xác tách từ mới cũng quan trọng với các ứng dụng thực tế. Tiếng Việt là một sinh ngữ – nó luôn luôn biến đổi. Các từ mới thuần Việt cũng như vay mượn được tạo ra hàng ngày. Nếu một ứng dụng không xử lý được những từ này thì hiệu năng của nó sẽ giảm dần theo thời gian.

Hiện tại bài toán tách từ có 3 phương pháp tiếp cận chính như sau:

* Tiếp cận dựa vào từ điển cố định.
* Tiếp cận dựa vào thống kê thuần túy.
* Tiếp cận dựa trên cả hai phương pháp trên.

Các phương pháp được sử dụng rất đa dạng và phong phú:

* So khớp từ dài nhất (Longest Matching)
* So khớp cực đại (Maximum Matching)
* Mô hình Markov ẩn (Hidden Markov Models- HMM)
* Học dựa trên sự cải biến (Transformation-based Learning – TBL)
* Chuyển đổi trạng thái trọng số hữu hạn (Weighted Finite State Transducer – WFST)
* Độ hỗn loạn cực đại (Maximum Entropy – ME)
* Máy học sử dụng vectơ hỗ trợ (Support Vector Machines)
* Trường xác xuất có điều kiện (CRFs)

Ngoài ra còn có thể kết hợp những phương pháp trên, ở đây em sẽ trình bày thuật toán phổ biến nhất đang được sử dụng đó là trọng số cực tiểu được quy về bài toán đồ thị như sau:

* Tạo ra 2 đỉnh ảo là start, và end (đầu và cuối câu).
* Lần lượt so sánh các đoạn với độ dài bất kì với 1 từ điển ngôn ngữ có sẵn.
* Các đoạn có xuất hiện trong từ điển sẽ tạo thành 1 đỉnh mới trên đồ thị.
* Trọng số giữa 2 đỉnh (2 đoạn phải liên tiếp nhau trong câu) được tính theo công thức f(i,j) với i,j là 2 từ.
* Tìm đường đi từ đỉnh start đến đỉnh end có trọng số nhỏ nhất trên đồ thị đó.

Trong bước 4, công thức f(i,j) thường được tính theo giá trị uni-gram(khả năng xuất hiện của 1 từ) và bi-gram(khả năng 2 từ xuất hiện liên tiếp nhau). Ngoài ra còn có thể có thêm 1 số yếu tố khác như từ loại, khả năng liên kết từ loại, ... cũng được sử dụng trong hàm f.

Trước đây, các giá trị này (trừ uni-gram và bi-gram được lấy từ cách thống kê corpus) thường được đánh giá bằng tay (do người thực hiện).  Nhưng với sự phát triển của lý thuyết máy học như Markov ẩn, CRFs, ... các giá trị này thường được tính 1 cách tự động.

Trong bước 5, thuật toán tìm đường đi từ đỉnh start đến đỉnh end thường sử dụng thuật toán Viterbi với độ phức tạp thuật toán O(n) với n là độ dài câu cần tách.

Ngoài ra, có 1 vấn đề nảy sinh trong khi tách từ là việc xuất hiện các từ mời (các từ không được định nghĩa trong từ điển).

Đây là 1 vấn đề không thể bỏ qua khi ngôn ngữ là luôn luôn thay đổi và sinh ra các từ mới, trong khi từ điển (dành cho xử lý ngôn ngữ tự nhiên) không thể cập nhật hết được.

Các nghiên cứu đối với tiếng Nhật và tiếng Trung đã xử lý vấn đề này khá tốt. Với sự gần gũi về ngữ pháp và đặc điểm ngôn ngữ, ta có thể ứng dụng những nghiên cứu đó đối với tiếng Việt.

Dưới đây là một số công cụ tách từ có kết quả tốt cho tiếng Việt:

* vnTokenizer được viết trên ngôn ngữ Java và có độ chính xác lên đến 97%
* DongDu được viết trên ngôn ngữ C++ với độ chính xác cũng rất cao (>98%)
* JVnSegmenter

1. Bài toán gán nhãn từ loại

Một trong các vấn đề nền tảng của phân tích ngôn ngữ là việc phân loại các từ thành các lớp từ loại dựa theo thực tiễn hoạt động ngôn ngữ. Mỗi từ loại tương ứng với một hình thái và một vai trò ngữ pháp nhất định. Các bộ chú thích từ loại có thể thay đổi tuỳ theo quan niệm về đơn vị từ vựng và thông tin ngôn ngữ cần khai thác trong các ứng dụng cụ thể [5]. Mỗi từ trong một ngôn ngữ nói chung có thể gắn với nhiều từ loại, và việc giải thích đúng nghĩa một từ phụ thuộc vào việc nó được xác định đúng từ loại hay không. Công việc gán nhãn từ loại cho một văn bản là xác định từ loại của mỗi từ trong phạm vi văn bản đó. Khi hệ thống văn bản đã được gán nhãn, hay nói cách khác là đã được chú thích từ loại thì nó sẽ được ứng dụng rộng rãi trong các hệ thống tìm kiếm thông tin, trong các ứng dụng tổng hợp tiếng nói, các hệ thống nhận dạng tiếng nói cũng như trong các hệ thống dịch máy.

Quá trình gán nhãn từ loại có thể chia làm 3 bước [6].

* Phân tách xâu kí tự thành chuỗi các từ. Giai đoạn này có thể đơn giản hay phức tạp tuỳ theo ngôn ngữ và quan niệm về đơn vị từ vựng. Chẳng hạn đối với tiếng Anh hay tiếng Pháp, việc phân tách từ phần lớn là dựa vào các kí hiệu trắng. Tuy nhiên vẫn có những từ ghép hay những cụm từ công cụ gây tranh cãi về cách xử lí. Trong khi đó với tiếng Việt thì dấu trắng càng không phải là dấu hiệu để xác định ranh giới các đơn vị từ vựng do tần số xuất hiện từ ghép rất cao.
* Gán nhãn tiên nghiệm, tức là tìm cho mỗi từ tập tất cả các nhãn từ loại mà nó có thể có. Tập nhãn này có thể thu được từ cơ sở dữ liệu từ điển hoặc kho văn bản đã gán nhãn bằng tay. Đối với một từ mới chưa xuất hiện trong cơ sở ngữ liệu thì có thể dùng một nhãn ngầm định hoặc gắn cho nó tập tất cả các nhãn. Trong các ngôn ngữ biến đổi hình thái người ta cũng dựa vào hình thái từ để đoán nhận lớp từ loại tương ứng của từ đang xét.
* Quyết định kết quả gán nhãn, đó là giai đoạn loại bỏ nhập nhằng, tức là lựa chọn cho mỗi từ một nhãn phù hợp nhất với ngữ cảnh trong tập nhãn tiên nghiệm. Có nhiều phương pháp để thực hiện việc này, trong đó người ta phân biệt chủ yếu các phương pháp dựa vào quy tắc ngữ pháp mà đại diện nổi bật là phương pháp Brill ([7]) và các phương pháp xác suất ([8]). Ngoài ra còn có các hệ thống sử dụng mạng nơ-ron ([9]), các hệ thống lai sử dụng kết hợp tính toán xác suất và ràng buộc ngữ pháp [10], gán nhãn nhiều tầng [11].

Về mặt ngữ liệu, các phương pháp phân tích từ loại thông dụng hiện nay dùng một trong các loại tài nguyên ngôn ngữ sau:

* Từ điển và các văn phạm loại bỏ nhập nhằng [12].
* Kho văn bản đã gán nhãn [8], có thể kèm theo các quy tắc ngữ pháp xây dựng bằng tay [7].
* Kho văn bản chưa gán nhãn, có kèm theo các thông tin ngôn ngữ như là tập từ loại và các thông tin mô tả quan hệ giữa từ loại và hậu tố [13].
* Kho văn bản chưa gán nhãn, với tập từ loại cũng được xây dựng tự động nhờ các tính toán thống kê [11]. Trong trường hợp này khó có thể dự đoán trước về tập từ loại.

Dưới đây là ví dụ đơn giản về bài toán gán nhãn từ loại cũng như một số phương pháp để giải quyết:

* Input: Đi tìm tàu phải không?
* Output: Đi/V tìm/V tàu/N phải/V không/R ?/?

*Phương pháp gán nhãn xác suất:*

Ý tưởng của phương pháp gán nhãn từ loại xác suất là xác định phân bố xác suất trong không gian kết hợp giữa dãy các từ *Sw* và dãy các nhãn từ loại *St*. Sau khi đã có phân bố xác suất này, bài toán loại bỏ nhập nhằng từ loại cho một dãy các từ được đưa về bài toán lựa chọn một dãy từ loại sao cho xác suất điều kiện *P(St | Sw)* kết hợp dãy từ loại đó với dãy từ đã cho đạt giá trị lớn nhất.

Theo công thức xác suất Bayes ta có:

*P(St | Sw) = P(Sw | St).P(St)/P(Sw)*.

Ở đây dãy các từ *Sw* đã biết, nên thực tế chỉ cần cực đại hoá xác suất

*P(Sw | St).P(St)*. Với mọi dãy *St = t1t2 ... tN* và với mọi dãy *Sw = w1w2 ... wN* :

*P(w1w2... wN | t1t2...tN) = P(w1 | t1t2...tN) P(w2 | w1,t1t2...tN)...P(wN | w1... wN-1, t1t2...tN)*

*P(t1t2...tN) = P(t1)P(t2 | t1) P(t3 | t1t2) ... P(tN | t1...tN-1)*

Người ta đưa ra các giả thiết đơn giản hoá cho phép thu gọn mô hình xác suất về một số hữu hạn các tham biến.

Đối với mỗi *P(wi | w1... wi-1, t1t2...tN)*, giả thiết khả năng xuất hiện một từ khi cho một nhãn từ loại là hoàn toàn xác định khi biết nhãn đó, nghĩa là *P(wi | w1... wi-1, t1t2...tN) = P(wi | ti)*.

Như vậy xác suất *P(w1w2... wN|t1t2...tN)* chỉ phụ thuộc vào các xác suất cơ bản có dạng *P(wi| ti)*:

*P(w1w2... wN | t1t2...tN) = P(w1 | t1)P(w2 | t2) ... P(wN | tN)*

Đối với các xác suất *P(ti | t1...ti-1)*, giả thiết khả năng xuất hiện của một từ loại là hoàn toàn xác định khi biết các nhãn từ loại trong một lân cận có kích thước *k* cố định, nghĩa là: *P(ti | t1...ti-1)= P(ti | ti-k...ti-1)*. Nói chung, các bộ gán nhãn thường sử dụng giả thiết *k* bằng 1 (bigram) hoặc 2 (trigram).

Như vậy mô hình xác suất này tương đương với một mô hình Markov ẩn, trong đó các trạng thái ẩn là các nhãn từ loại (hay các dãy gồm *k* nhãn nếu *k* > 1), và các trạng thái hiện (quan sát được) là các từ trong từ điển. Với một kho văn bản đã gán nhãn mẫu, các tham số của mô hình này dễ dàng được xác định nhờ thuật toán Viterbi.

Ngoài ra còn có một số phương pháp để giải quyết bài toán tách từ như:

* Phương pháp cực đại hóa Entropy (Maxent)
* Phương pháp Conditional Random Fields (CRFs)

Dưới đây là một số công cụ có kết quả tốt cho việc gán nhãn từ loại tiếng Việt:

* vnTagger – Lê Hồng Phương
* vnQTag – Nguyễn Thị Minh Huyền
* JvnTagger – Phan Xuân Hiếu

1. Bài toán phân tích cú pháp

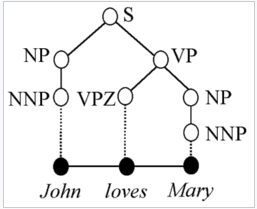
Phân tích cú pháp là bước xử lý quan trọng trong các bài toán hiểu ngôn ngữ tự nhiên. Nó cung cấp một nền tảng vững chắc cho việc xử lý văn bản thông minh như các hệ thống hỏi đáp, khai phá văn bản và dịch máy.

Việc phân tích cú pháp câu có thể chia làm hai mức chính. Mức thứ nhất là tách từ và xác định thông tin từ loại. Mức thứ hai là sinh cấu trúc cú pháp cho câu dựa trên các từ và từ loại do bước trước cung cấp. Do tiếng Việt là ngôn ngữ đơn âm tiết nên chúng ta thường gặp phải vấn đề nhập nhằng ở cả hai mức.

Để hiểu đơn giản hơn về ý tưởng của bài toán chúng ta có thể xem ví dụ dưới đây:

* Input: John loves Marry
* Output: (S (NP (NNP John)) (VP (VPZ loves) (NP (NNP Mary))) (. .))

Và với kết quả đầu ra này chúng ta có thể xây dựng được cây như sau (treebank)



1. Bài toán phân tích tình cảm

**Cơ sở lý thuyết : (Chính 1 15 trang)**

Hệ thống tập trung công cụ xử lý ngôn ngữ nào thì trình bày về bài toán đấy.

Bài toán 1 : tách từ

* Mô tả bài toán : mục tiêu
* Input
* Output :
* Một số các tiếp cận để giải quyết(thuật toán…)
* Một số công cụ có kết quả tốt cho TV.

Bài toán 2 : gán nhãn từ loại

**Xây dựng (Phân tích – thiết kế) hệ thống (chính 2 20 trang)**

Mô tả về hê thống

Cấu trúc (các thành phần hệ thống) : bao nhiêu module :

Trình bày từng module

Chức năng

* Mô tả rõ cách thức hoạt động + ý nghĩa (nghiệp vụ)
* Mô tả rõ thao tác người dùng thực hiện

Mô tả quá trình triển khai :

Ngôn ngữ ?

Framework ?

API nào ?

Dùng openSource nào ?(Không để mã nguồn ở đây)

**Kết quả (5 trang)**

Đã làm được đến đâu :

Ưu điểm, nhược điểm, khắc phục.

Phụ lục

Mã nguồn

Thông tin bổ sung

[1] <http://viet.jnlp.org/kien-thuc-co-ban-ve-xu-ly-ngon-ngu-tu-nhien/thuat-toan-tach-tu-tokenizer/thuat-toan-tach-tu>

[2] <http://www.vietlex.com/images/tulieu/ITCra03POSTagging.pdf>

[3] Wikipedia (truy cập ngày 24/5/2008). Sentiment analysis. <https://en.wikipedia.org/wiki/Sentiment_analysis>

[4] Ryan McDonald, Joakim Nivre (2007). Introduction to Data-Driven Dependency Parsing. Introductory Course, ESSLLI 2007.

[5] Vergnes J., Giguet E., "Regards théoriques sur le tagging", 5e conférence sur le Traitement Automatique du Langage Naturel (TALN98), Paris, 10-12 juin, 1998.

[6] Paroubek P., Rajman M., "Etiquetage morpho-syntaxique", Ingénierie des langues, chapitre 5, Hermes Science Europe, 2000.

[7] Brill E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", Computational Linguistics, 21(4), December 199, p.543-565.

[8] Dermatas E., Kokkinakis G., "Automatic Stochastic Tagging of Natural Language Texts", Computational Linguistics 21.2, 1995, p. 137 - 163.

[9] Schmid H., "Part-of-Speech Tagging with Neural networks", International Conference on Computational Linguistics, Japan, 1994, p. 172-176, Kyoto.

[10] El-Bèze M, Spriet T., "Etiquetage probabiliste et contraintes syntaxiques", Actes de la conférence sur le Traitement Automatique du Langage Naturel (TALN95), Marseille, France,14-16/6/1995.

[11] Tufis D., "Tiered Tagging and combined classifier", In Jelineck F. and Nörth E. (Eds), Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999.

[12] Oflazer K., "Error-tolenrant finite-state recognition with applications to morphological analysis and spelling correction", Computational Linguistics, 22(1), 1996, p. 73-89.

[13] Levinger M., Ornan U., Itai A., "Learning morpho-lexical probabilities from an untagged corpus with an application to Hebrew", Comutational Linguistics, 21(3), 1995, p. 383-404.

[14] MacMahon J.G., Smith F.J., "Improving statistical language model performance with automatically generated word hierarchies", Computational Linguistics, 19(2), 1993, p. 313-330