BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

ĐINH THỊ MỸ HẠNH

TÌM HIỂU HIỆN TƯỢNG NHẬP NHẰNG TRONG TIẾNG VIỆT VÀ KHẢ NĂNG KHẮC PHỤC TRONG SOẠN THẢO VĂN BẢN

Chuyên ngành : KHOA HỌC MÁY TÍNH
Mã số : 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS.TS. PHAN HUY KHÁNH

Phản biện 1: PGS.TS. Võ Trung Hùng

Phản biện 2: TS. Trương Công Tuấn

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 19 tháng 6 năm 2011.

** Có thể tìm hiểu luận văn tại:*

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Trong hệ thống ngôn ngữ trên thế giới hiện nay, tiếng Việt được xem là một trong những ngôn ngữ có sự phong phú và đa dạng. Chính sự phong phú và phức tạp của tiếng Việt đã dẫn đến những khó khăn cho cả người sử dụng theo cách thông thường và khi xử lý trên máy tính.

Do những ảnh hưởng của lịch sử hình thành và phát triển, tiếng Việt có tính lai tạp về mặt ngôn ngữ rất cao, đặc biệt ảnh hưởng từ tiếng Hán và tiếng Pháp. Sự đa nghĩa trong tiếng Việt cũng chính là một đặc điểm nổi bật. Ngoài ra, do thói quen sử dụng của mỗi người hoặc mỗi vùng miền, tiếng Việt lại có những sự biến đổi nhất định, thậm chí sự thiếu nhất quán trong cách nói, cách viết.

Tiếng Việt ngày nay còn bị ảnh hưởng bởi thói quen sử dụng ngôn ngữ trên Internet hoặc các thiết bị truyền thông hiện đại như điện thoại di động, điều này làm xuất hiện thêm nhiều từ mới được người dùng Internet hoặc điện thoại di động chấp nhận, đặc biệt giới trẻ như câu *"Buổi sinh nhật hôm nay vui wá!"*.

Sự nhập nhằng trong khi nói, viết hoặc diễn đạt ý nghĩ đã dẫn đến những sự hiểu lầm ở nhiều mức độ khác nhau. Và cũng chính những đặc điểm nói trên đã làm cho tiếng Việt vốn đã phức tạp lại càng phức tạp hơn, đặc biệt là một số yếu tố đã làm mất đi tính trong sáng và giá trị tốt đẹp vốn có của tiếng Việt.

Gìn giữ sự trong sáng của tiếng Việt là một yêu cầu và nhiệm vụ quan trọng, cần thiết của cả giới phê bình, nghiên cứu văn hóa, ngôn ngữ lẫn những người sử dụng thông thường.

Xuất phát từ những phân tích và quan sát trên, nhiệm vụ nghiên cứu của đề tài ***"Tìm hiểu hiện tượng nhập nhằng trong***

tiếng Việt và khả năng khắc phục trong soạn thảo văn bản" là tìm hiểu về những vấn đề cơ bản trong xử lý ngôn ngữ, xử lý tiếng Việt, đặc biệt là vấn đề "nhập nhằng" trong tiếng Việt, từ đó đề xuất một giải pháp để khắc phục trong quá trình soạn thảo văn bản cho một số trường hợp cụ thể của hiện tượng nhập nhằng.

2. Mục tiêu của đề tài

Đề tài tập trung nghiên cứu về xử lý ngôn ngữ tự nhiên, xử lý tiếng Việt. Tác giả cũng bỏ nhiều thời gian nghiên cứu về các hiện tượng nhập nhằng thường xảy ra trong tiếng Việt. Đề tài còn nghiên cứu các khả năng xử lý nhập nhằng và xây dựng ứng dụng hỗ trợ xử lý nhập nhằng tiếng Việt trong một phạm vi hẹp.

3. Phạm vi và giới hạn của đề tài

Vấn đề nhập nhằng trong tiếng Việt có rất nhiều trường hợp, tuy nhiên trong phạm vi của đề tài này tác giả giới hạn lại một số nội dung sau đây:

Về mặt lý thuyết: Tìm hiểu lý thuyết về XLNN và XLTV, lịch sử hình thành và phát triển của tiếng Việt; Tìm hiểu lý thuyết về các vấn đề liên quan đến hiện tượng nhập nhằng trong tiếng Việt; Tìm hiểu những vấn đề cơ bản về soạn thảo văn bản, phần mềm soạn thảo văn bản; Đề xuất giải pháp để giải quyết HTNN do viết sai lỗi chính tả tiếng Việt (giới hạn những lỗi chính tả ở cấp độ âm tiết) và HTNN do xác định sai phạm vi, ranh giới của từ tiếng Việt.

Về mặt chương trình: Xây dựng ứng dụng hỗ trợ xử lý nhập nhằng gây ra do lỗi chính tả về mặt âm tiết, đồng thời chương trình hỗ trợ việc tách văn bản thành các từ độc lập để người sử dụng dễ dàng hiểu nội dung văn bản. Kết quả của việc tách từ sẽ được sử dụng phục vụ cho việc phát triển ứng dụng, giải quyết vấn đề phân tích nhập nhằng về phân loại từ và cú pháp câu.

4. Phương pháp nghiên cứu

Thu thập, tìm hiểu, phân tích các tài liệu và thông tin có liên quan đến đề tài; Phân tích và thiết kế hệ thống chương trình; Triển khai xây dựng chương trình; Kiểm thử, nhận xét và đánh giá kết quả.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Ý nghĩa khoa học: Hiểu được những vấn đề cơ bản trong xử lý tiếng Việt, xử lý nhập nhằng trong tiếng Việt; Đề xuất được giải pháp để hỗ trợ xử lý một số hiện tượng nhập nhằng trong soạn thảo văn bản tiếng Việt.

Ý nghĩa thực tiễn: Hiểu và ứng dụng được những kiến thức nền tảng trong xử lý tiếng Việt để xử lý nhập nhằng trong tiếng Việt; Có thể ứng dụng chương trình này để hỗ trợ xử lý nhập nhằng trong soạn thảo văn bản tiếng Việt; Có ý nghĩa trong việc bảo tồn và phát huy các giá trị của tiếng Việt.

6. Bố cục luận văn

Mở đầu

Chương 1: Cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên

Chương 2: Soạn thảo văn bản và hiện tượng nhập nhằng trong soạn thảo văn bản

Chương 3: Đề xuất giải pháp khắc phục nhập nhằng

Kết luận.

CHƯƠNG 1:
CƠ SỞ LÝ THUYẾT VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN

1.1. XỬ LÝ NGÔN NGỮ TỰ NHIÊN

1.1.1. Khái niệm

1.1.2. Các bước xử lý

1.1.3. Các bài toán và ứng dụng

1.2. TÌM HIỂU VỀ TIẾNG VIỆT VÀ VẤN ĐỀ XỬ LÝ TIẾNG VIỆT

1.2.1. Sự hình thành của tiếng Việt

1.2.2. Đặc điểm của tiếng Việt

1.2.2.1. Đặc điểm ngữ âm

1.2.2.2. Đặc điểm từ vựng

1.2.2.3. Đặc điểm ngữ pháp

1.2.3. Từ trong tiếng Việt

1.2.3.1. Khái niệm

Từ là đơn vị nhỏ nhất có nghĩa, có kết cấu vỏ ngữ âm bền vững, hoàn chỉnh, có chức năng gọi tên, được vận dụng độc lập, tái hiện tự do trong lời nói để tạo câu[7].

1.2.3.2. Đơn vị cấu tạo

Đơn vị cơ sở để cấu tạo từ tiếng Việt là các *tiếng*, cái mà ngữ âm học vẫn gọi là các *âm tiết*.

1.2.3.3. Phương thức cấu tạo

1.2.4. Biến thể của từ

1.3. HIỆN TƯỢNG NHẬP NHẰNG TRONG XỬ LÝ VĂN BẢN TIẾNG VIỆT

1.3.1. Khái niệm

Nhập nhằng là hiện tượng mà khi nói, viết hoặc diễn tả những từ ngữ, ý nghĩ mơ hồ, không rõ nghĩa hoặc có nhiều nghĩa làm cho

người đọc hoặc người nghe không phân biệt rõ ràng, gây ra sự hiểu lầm.

Khái niệm *nhập nhằng* cũng có thể hiểu theo một cách khác như sau: trong mỗi lĩnh vực, các vấn đề thường được đề cập, trình bày hoặc được hiểu theo một chuẩn nhất định, chuẩn này có thể được quy ước bằng văn bản hoặc quy ước ngầm định. Khi đó tất cả những cách hiểu, cách đề cập hoặc trình bày vấn đề nằm ngoài phạm vi chuẩn đó được xem là nhập nhằng.

1.3.2. Một số hiện tượng nhập nhằng

1.3.2.1. Hiện tượng nhập nhằng do viết sai chính tả tiếng Việt

Việc viết sai lỗi chính tả tiếng Việt đang ở mức đáng báo động, hiện tượng này không những diễn ra phổ biến trong giới trẻ, đặc biệt trong giới trẻ sử dụng internet mà còn xuất hiện nhiều trên các phương tiện thông tin đại chúng và các văn bản của Nhà nước. Chính những sai sót về chính tả này có thể gây ra những nhập nhằng trong việc tiếp nhập thông tin.

Trong cộng đồng sử dụng mạng Internet cũng hình thành một lớp từ mới chưa có trong từ điển tiếng Việt, điều này cũng dẫn đến những thói quen sử dụng từ ngữ không tốt trong người sử dụng, đặc biệt giới trẻ.

Có những trường hợp sai chính tả có thể dẫn đến những hệ quả xấu như hình thành thói quen nói sai, viết sai và hiểu sai vấn đề.

1.3.2.2. Hiện tượng nhập nhằng về phạm vi, ranh giới giữa các từ

Trong một số ngôn ngữ như tiếng Anh, việc xác định ranh giới, phạm vi giữa các từ khá dễ dàng, mỗi từ riêng lẻ đã mang trọn vẹn một nghĩa và ranh giới của chúng được xác định thông qua

khoảng trắng. Tiếng Việt thì khác, do là ngôn ngữ đơn lập nên từ vựng chủ yếu là các từ ghép vì thế khoảng trắng không phải luôn luôn là ranh giới chính xác.

Trong tiếng Việt, việc xác định chính xác phạm vi, ranh giới giữa các từ có thể hỗ trợ rất nhiều cho quá trình khử nhập nhằng, đặc biệt đối với ngôn ngữ viết. Đây cũng chính là mục đích chính mà báo cáo này muồn đề cập đến.

1.3.2.3. Hiện tượng nhập nhằng do tính đa nghĩa của từ

Bất cứ ngôn ngữ nào cũng có từ đa nghĩa, nguyên nhân là vì rất nhiều khái niệm có các sắc thái ý nghĩa tuy không hoàn toàn trùng khớp nhau nhưng lại có nhiều nét tương đồng. Hiện tượng này gây cản trở cho việc dịch tự động, chương trình không biết dịch từ đa nghĩa theo nghĩa nào trong nhóm nghĩa của nó.

1.3.2.4. Hiện tượng nhập nhằng ngữ nghĩa khi sử dụng các từ đồng âm

Hai từ *đồng âm* với nhau nghĩa là hai từ có âm giống nhau nhưng mang nghĩa khác nhau, còn *đồng tự* là hai từ về mặt ký tự là giống nhau nhưng nghĩa khác nhau. Do đặc điểm của tiếng Việt từ đồng âm cũng thường là từ đồng tự, ở các ngôn ngữ khác hai hiện tượng này không trùng khớp nhau.

Cũng phải phân biệt từ đồng tự với từ đa nghĩa, trong từ đa nghĩa, các nghĩa đều có chung một nguồn gốc và do vậy luôn có nét tương đồng trong khi đó trong từ đồng tự chúng không có liên hệ về nguồn gốc với nhau, nghĩa của chúng khác nhau rõ rệt.

Ví dụ 1

Từ *"kiếm"* trong hai câu sau đây là hai từ đồng tự:

Anh ta sử dụng kiếm rất điêu luyện.

Kiếm ăn bây giờ khó lắm.

1.3.2.5. Hiện tượng nhập nhằng trong cách phân biệt từ loại

Từ loại là một yếu tố quan trọng trong việc xác định nghĩa chính xác của từ và sắp xếp các từ thành câu hoàn chỉnh trong dịch tự động.

Từ loại giúp khử nhập nhằng, nhưng chính bản thân nó trong một số trường hợp cũng nhập nhằng. Với các *ngôn ngữ không biến hình* như tiếng Việt, vấn đề xác định từ loại yêu cầu các thuật toán phức tạp hơn, bắt buộc phải phân tích cú pháp. Mặt khác, ngay trong nội bộ ngành ngôn ngữ vẫn chưa có sự thống nhất về phân loại từ loại cho tiếng Việt.

1.3.2.6. Hiện tượng nhập nhằng khi sử dụng tiếng Việt không dấu

Ngày nay, việc gõ tiếng Việt không dấu trở nên phổ biến hơn, đặc biệt trên các ứng dụng Internet hoặc điện thoại di động như email, chat… Gõ tiếng Việt không dấu giúp người sử dụng thao tác nhanh hơn, nhưng trong một số trường hợp nó lại gây ra những sự hiểu nhầm tai hại đối với người đọc.

1.3.2.7. Hiện tượng nhập nhằng về sự vận dụng

Cùng một câu nhưng khi sử dụng trong các hoàn cảnh khác nhau của ngôn ngữ nói hoặc ngôn ngữ viết, nếu không biết cách sử dụng một cách phù hợp cũng sẽ gây ra sự "nhập nhằng", sự hiểu lầm cho người đọc hoặc người nghe. Hiện tượng này đặc biệt phổ biến trong tiếng Việt, vì tiếng Việt vốn đa nghĩa, đa sắc thái và có tính biểu cảm rất cao. Điều này đòi hỏi người sử dụng ngôn ngữ một sự khéo léo và tinh tế nhất định, có sự hiểu biết ở một mức độ cần thiết để có thể tận dụng hết những giá trị biểu đạt của ngôn ngữ.

1.3.2.8. Hiện tượng nhập nhằng trong phân tích cú pháp tiếng Việt

Trong phân tích cú pháp tiếng Việt, hiện tượng nhập nhằng xảy ra ở nhiều mức, từ mức từ, từ loại đến mức cú pháp câu. Điều này dẫn đến một câu có thể được phân tích theo nhiều cách khác nhau, trong khi chỉ có một vài cách phân tích trong số đó đúng.

1.4. KẾT LUẬN CHƯƠNG

Chương này trình bày khái niệm và các bước để xử lý ngôn ngữ tự nhiên, các bài toán liên quan đến xử lý ngôn ngữ tự nhiên như nhận dạng tiếng nói, tổng hợp tiếng nói, dịch tự động, tìm kiếm văn bản, tóm tắt văn bản… Ngoài ra còn trình bày về sự hình thành, phát triển và một số đặc điểm nổi bật của tiếng Việt. Chương 1 còn dành một số lượng lớn các trang để trình bày khái niệm cũng như những hiện tượng nhập nhằng phổ biến nhất trong xử lý văn bản tiếng Việt.

CHƯƠNG 2:
SOẠN THẢO VĂN BẢN VÀ HIỆN TƯỢNG NHẬP NHẰNG TRONG SOẠN THẢO VĂN BẢN

2.1. MỘT SỐ VẤN ĐỀ VỀ SOẠN THẢO VĂN BẢN

2.1.1. Đặt vấn đề

2.1.2. Khái niệm ký tự, từ, câu, dòng, đoạn

2.1.3. Nguyên tắc tự xuống dòng của từ

2.1.4. Một số quy tắc gõ văn bản cơ bản

2.1.5. Phần mềm soạn thảo văn bản

2.2. HIỆN TƯỢNG NHẬP NHẰNG TRONG SOẠN THẢO VĂN BẢN

Những mức độ nhập nhằng trong STVB: Trong quá trình soạn thảo văn bản, hiện tượng nhập nhằng có thể xảy ra ở nhiều mức độ khác nhau:

Mức một, nhập nhằng xảy ra do sai sót về từ, cụm từ, sai sót chữ viết tắt, cách viết ngày tháng năm, viết các ký hiệu. *Mức hai*, nhập nhằng ở mức độ cú pháp câu. *Mức ba*, nhập nhằng về mặt ngữ nghĩa.

2.3. CÁCH PHÁT HIỆN HIỆN TƯỢNG NHẬP NHẰNG TRONG SOẠN THẢO VĂN BẢN

Thứ nhất, phát hiện HTNN trước khi tiến hành STVB. Quá trình này chính là khử nhập nhằng trong tư duy, suy nghĩ của người soạn thảo, nói chính xác hơn thì trong trường hợp này, bản thân người soạn thảo phải tự tìm cách để khử nhập nhằng bằng cách nắm vững các quy tắc về chính tả tiếng Việt, quy tắc STVB, hiểu biết về ngôn ngữ tiếng Việt và biết cách vận dụng phù hợp. Nếu bản thân người soạn thảo không thể tự tìm và khử được những nhập nhằng

trong tư duy, suy nghĩ thì có thể trao đổi với người khác để có cách trình bày vấn đề chính xác hơn.

Thứ hai, phát hiện HTNN trong quá trình STVB. Nếu chỉ sử dụng phần mềm hỗ trợ STVB tiếng Việt (mà không sử dụng kèm một chương trình hỗ trợ tìm và khử nhập nhằng cho văn bản tiếng Việt nào khác) thì chỉ có một cách để phát hiện nhập nhằng là người sử dụng phải tự làm thủ công. Tuy nhiên cách làm này sẽ không đem lại nhiều hiệu quả và độ chính xác không cao. Do đó nhất thiết phải có một chương trình hỗ trợ phát hiện và khử nhập nhằng đi kèm.

Thứ ba, phát hiện HTNN sau khi việc STVB hoàn tất. Nghĩa là người sử dụng sẽ mở tệp văn bản đã soạn thảo, sau đó gọi chức năng phát hiện nhập nhằng để xử lý.

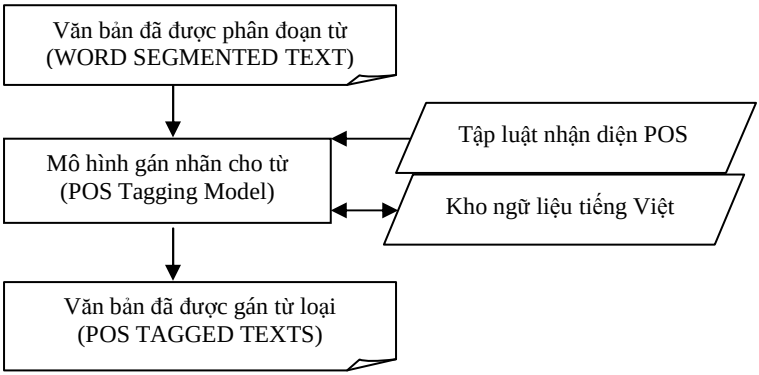
Theo tìm hiểu của tác giả, hiện nay vẫn chưa có một hệ thống hay chương trình nào có thể xử lý được tất cả các HTNN trong STVB tiếng Việt. Các kết quả đã có cũng chỉ mới là những giải pháp cho một số trường hợp cụ thể.

2.4. CÁC GIẢI PHÁP XỬ LÝ NHẬP NHẰNG ĐÃ CÓ TRƯỚC ĐÂY

Trong bài báo *"Phân tích cú pháp tiếng Việt sử dụng văn phạm phi ngữ cảnh từ vựng hóa kết hợp xác xuất"* [25], nhóm tác giả đã nghiên cứu biện pháp xử lý hiện tượng nhập nhằng và các hiện tượng cú pháp phụ thuộc từ trong phân tích cú pháp tiếng Việt. Bài báo đề xuất việc xây dựng một công cụ phân tích cú pháp dựa trên văn phạm phi ngữ cảnh với luật có chứa thông tin về xác suất và từ vựng.

Trong tài liệu [24], nhóm tác giả đã trình bày chi tiết các thử nghiệm về gán nhãn từ loại cho các văn bản tiếng Việt bằng cách áp dụng bộ gán nhãn QTAG. 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. Để tiến hành việc gán nhãn từ loại, nhóm tác giả đã tiến hành ba bước: Phân tách xâu ký tự thành các từ, gán nhãn tiên nghiệm, quyết định kết quả gán nhãn, tức loại bỏ nhập nhằng.



Hình 2.1. Mô hình tổng quát bài toán gán nhãn từ loại

2.5. KẾT LUẬN CHƯƠNG

Chương 2 trình bày những vấn đề cơ bản về soạn thảo văn bản, khái niệm về ký tự, từ, câu, dòng, đoạn. Trong chương này còn trình bày khái niệm về hiện tượng nhập nhằng trong tiếng Việt. Ngoài ra còn trình bày một số hiện tượng nhập nhằng phổ biến trong tiếng Việt, qua đó chúng ta có thể thấy rằng hiện tượng nhập nhằng khá phổ biến và rất khó để xử lý một cách triệt để. Phần cuối chương, tác giả đã trình bày những kết quả nghiên cứu về vấn đề xử lý nhập nhằng trong ngôn ngữ tự nhiên nói chung và tiếng Việt nói riêng.

CHƯƠNG 3:
ĐỀ XUẤT GIẢI PHÁP KHẮC PHỤC NHẬP NHẰNG

3.1. GIỚI THIỆU VÀ PHÂN TÍCH BÀI TOÁN

Trong số các hiện tượng nhập nhằng mà tác giả đã đề cập đến trong chương 2, trong phần này, tác giả chỉ chọn một số hiện tượng nhập nhằng cụ thể để đề xuất giải pháp khắc phục. Đó là **nhập nhằng do viết sai lỗi chính tả tiếng Việt ở cấp độ âm tiết (viết những âm tiết không có trong tiếng Việt) và nhập nhằng do không xác định được phạm vi, ranh giới giữa các từ trong văn bản**.

Nếu xem những quy tắc về chính tả tiếng Việt là miền chuẩn, và những gì nằm trong miền chuẩn ấy được chấp nhận và không gây nhập nhằng thì những trường hợp viết sai chính tả tiếng Việt nằm ngoài miền chuẩn (tức viết sai chính tả) đều được xem là nhập nhằng.

Trong phạm vi báo cáo này, tác giả xử lý một phần các lỗi chính tả tiếng Việt có thể mắc phải dẫn đến hiện tượng nhập nhằng, đó là xử lý lỗi chính tả ở mức âm tiết tiếng Việt. Ví dụ có thể phát hiện ra lỗi chính tả của từ và đưa ra một loạt gợi ý để người sử dụng chỉnh sửa lỗi.

Xét một ví dụ về hiện tượng nhập nhằng do không xác định được phạm vi, ranh giới giữa các từ.

Ví dụ 32

Người dân thuộc địa bàn đô thị có mức thu nhập bình quân đầu người cao hơn vùng nông thôn.

Trong ví dụ 32, một số đối tượng, ví dụ trẻ em có thể sẽ xác định không đúng phạm vi giữa các từ sẽ dẫn đến hiểu sai (nhập nhằng) nội dung câu. Cụm từ *thuộc địa bàn* sẽ có hai cách phân tách,

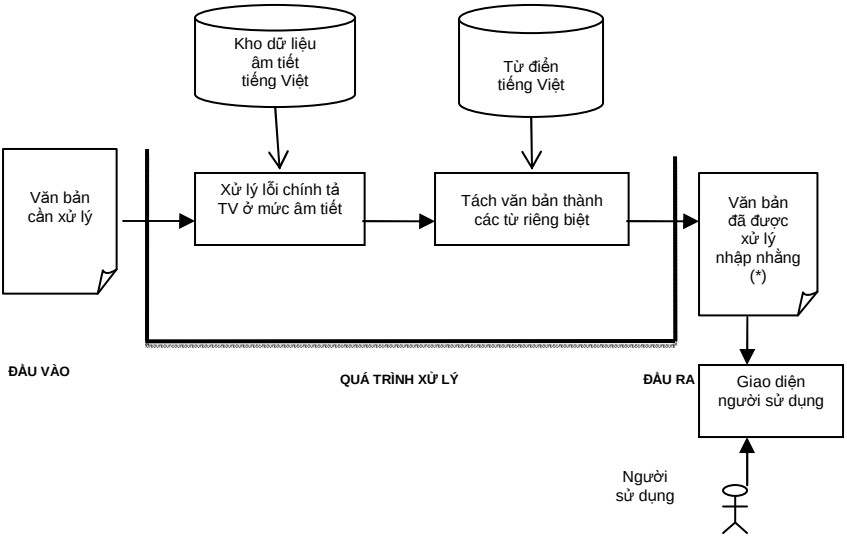
thứ nhất là *thuộc/ địa bàn* (nghĩa là nằm trên địa bàn nào đó), thứ hai là *thuộc địa/ bàn* (nghĩa là người dân ở thuộc địa bàn bạc về điều gì đó, đây là cách phân tách sai trong ngữ cảnh trên). Trong tiếng Việt tồn tại khá nhiều những cụm từ tương tự như trên.

Do đó, khi xác định đúng phạm vi từ sẽ đọc đúng và hiểu đúng, nghĩa là tránh được nhập nhằng. Giải quyết được bài toán về xác định phạm vi, ranh giới từ sẽ là cơ sở quan trọng để thực hiện gán nhãn từ loại cho từ, phân tích cú pháp câu tiếng Việt.

Giải quyết bài toán:

Bài toán bây giờ được chia làm hai bước xử lý:

- Xử lý lỗi chính tả tiếng Việt ở mức âm tiết
- Phân tách văn bản thành các từ độc lập



Hình 3.1. Mô hình kiến trúc tổng quan của quá trình xử lý

Giới hạn phạm vi giải quyết của bài toán:

Khái niệm *văn bản* được đề cập đến trong báo cáo này là văn bản chỉ chứa chữ, không chứa hình vẽ.

Chương trình cũng chưa xử lý các định dạng của văn bản đầu vào.

Thời điểm xử lý văn bản: Chương trình được tích hợp trong ứng dụng Microsoft Word và chương trình sẽ lấy nội dung văn bản được soạn thảo sẵn để xử lý. Người sử dụng có thể tùy chọn tiến hành kiểm lỗi chính tả mức âm tiết đối văn bản đầu vào, sau đó tiến hành phân tách từ. Chương trình không được gọi thực thi một cách tự động mà cần có thao tác của người sử dụng.

Loại nhập nhằng được xử lý: bài toán giải quyết sự nhập nhằng gây ra do sai sót về lỗi chính tả tiếng Việt ở mức âm tiết và hỗ trợ xử lý nhập nhằng có thể có do không xác định được phạm vi, ranh giới giữa các từ trong tiếng Việt.

Kho dữ liệu tiếng Việt:

Chương trình dùng 2 kho dữ liệu hỗ trợ cho quá trình xử lý:

Kho dữ liệu âm tiết tiếng Việt: lưu hơn 10.000 âm tiết tiếng Việt, hỗ trợ cho chức năng tìm và sửa lỗi, đồng thời có thể được cập nhập thêm từ mới thông qua bước xử lý lỗi chính tả. Các âm tiết được lưu ở bảng mã Unicode.

Từ điển tiếng Việt hỗ trợ chức năng tách từ gồm gần 24.000 từ tiếng Việt. Ngoài ra, để hỗ trợ tốt hơn cho việc tách từ, tác giả còn bổ sung vào kho dữ liệu từ tiếng Việt một số danh từ riêng phổ biến.

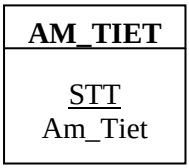
3.2. THIẾT KẾ CƠ SỞ DỮ LIỆU VÀ CÁC THUẬT TOÁN CHÍNH

3.2.1. Thiết kế cơ sở dữ liệu

Cơ sở dữ liệu cho bài toán tương đối đơn giản, dữ liệu được chia thành 2 phần riêng biệt, một phần phục vụ cho chức năng kiểm lỗi chính tả tiếng Việt ở mức âm tiết, một phần phục vụ cho chức năng tách từ trong văn bản.

3.2.1.1. Dữ liệu cho chức năng kiểm lỗi chính tả tiếng Việt mức âm tiết

Dữ liệu cho chức năng xử lý nhập nhằng do sai lỗi chính tả tiếng Việt ở mức âm tiết gồm 1 bảng AmTiet (Âm tiết) chứa tất cả các âm tiết có trong tiếng Việt.



Hình 3.4 Dữ liệu lưu các âm tiết tiếng Việt

Bảng 3.1. Bảng từ điển dữ liệu

Tên trường	Kiểu dữ liệu	Kích thước	Giải thích
STT	Autonumber	Integer	Thứ tự của mục từ
Am_Tiet	Text	10	Âm tiết tiếng Việt

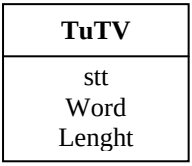
3.2.1.2. Dữ liệu cho chức năng tách từ trong văn bản

Để phục vụ cho giải thuật này, ta cần xây dựng cơ sở dữ liệu chứa tất cả các từ có trong tiếng Việt. Tác giả xây dựng kho dữ liệu này trên cơ sở tập tin dữ liệu của phần mềm VietDict của tác giả Hồ Ngọc Đức, tải miễn phí tại địa chỉ *http://vietdict.viet.net*. Tập tin này được lưu với đuôi *.txt, chứa gần 24.000 từ và giải thích từ (Việt – Việt), cấu trúc trình bày gần giống các quyển từ điển tiếng Việt.

Tác giả đã viết một thủ tục đơn giản để tiến hành tìm và tách lấy tất cả các từ tiếng Việt trong tập tin để lưu vào cơ sở dữ liệu. Đây chưa phải là tất cả các từ có trong tiếng Việt, chỉ là dữ liệu để demo chương trình.

Dữ liệu của chức năng tách từ là 1 bảng dữ liệu chứa các từ có trong tiếng Việt (căn cứ vào từ điển tiếng Việt), gồm 3 trường dữ liệu

là số thứ tự, mục từ Word và kích thước của từ Length. Ví dụ từ *ban mai* có Length =2, từ *sạch sành sanh* có Length =3. Trường Length dùng phục vụ cho một số giải thuật của chương trình.



Hình 3.5 Dữ liệu chứa các từ tiếng Việt

Mỗi bảng đều có 2 trường dữ liệu:

Bảng 3.2. Từ điển dữ liệu

Tên trường	Kiểu dữ liệu	Kích thước	Giải thích
stt	Autonumber	Integer	Thứ tự mục từ
Word	Text	30	Từ tiếng Việt
Length	Number	Byte	Kích thước từ

Bảng dữ liệu này chỉ có mục đích là lưu trữ dữ liệu. Dữ liệu được lưu ở bảng mã Unicode, kiểu gõ Telex để thống nhất với dữ liệu của phần kiểm lỗi chính tả tiếng Việt đã trình bày ở phần trên của báo cáo.

3.2.2. Các giải thuật chính

3.2.2.1. Giải thuật tìm và hỗ trợ sửa lỗi chính tả tiếng Việt ở mức âm tiết

3.2.2.2. Thuật toán xác định từ trong văn bản

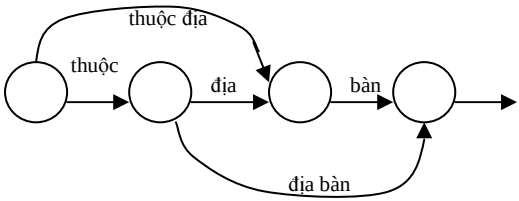
3.2.2.3. Thuật toán tách đoạn văn bản thành các từ riêng biệt

Đây là chức năng chính của chương trình. Trước khi trình bày giải thuật tách từ của mình, tác giả xin trình bày sơ qua một số giải thuật mà tác giả đã tìm hiểu được liên quan đến vấn đề này.

Theo tài liệu [23], nhóm tác giả này đã tiến hành xây dựng otomat đoán nhận từ vựng, phục vụ cho việc tách từ vựng trong văn bản tiếng Việt. Tư tưởng của thuật toán tách từ vựng là quy việc phân tách câu về việc tìm đường đi trên một đồ thị có hướng, không trọng số.

Giả sử câu ban đầu là một dãy gồm n+1 âm tiết s0, s1, ..., sn. Ta xây dựng một đồ thị có n+2 đỉnh v0, v1, ..., vn, vn+1, sắp thứ tự trên một đường thẳng từ trái sang phải; trong đó, từ đỉnh vi đến đỉnh vj có cung (i < j) nếu các âm tiết si, si+1, ..., sj-1 theo thứ tự lập thành một từ. Khi đó mỗi cách phân tách câu khác nhau tương ứng với một đường đi trên đồ thị từ đỉnh đầu v0 đến đỉnh cuối vn+1. Trong thực tế, cách phân tích câu đúng đắn nhất thường ứng với đường đi qua ít cung nhất trên đồ thị.

Trong trường hợp câu có sự nhập nhằng thì đồ thị sẽ có nhiều hơn một đường đi ngắn nhất từ đỉnh đầu đến đỉnh cuối, ta liệt kê toàn bộ các đường đi ngắn nhất trên đồ thị, từ đó đưa ra tất cả các phương án tách câu có thể và để người dùng quyết định sẽ chọn phương án nào, tuỳ thuộc vào ngữ nghĩa hoặc văn cảnh. Ví dụ, xét một câu có cụm "*thuộc địa bàn*", ta có đồ thị như hình 3.7 sau:



Hình 3.7 Otomat đoán nhận cụm từ "thuộc địa bàn"

Cụm này có sự nhập nhằng giữa *thuộc địa* và *địa bàn* và ta sẽ có hai kết quả phân tách là *"thuộc địa / bàn"* và *"thuộc / địa bàn"*. Ta có thể chỉ ra rất nhiều những cụm nhập nhằng trong tiếng Việt, chẳng

hạn "*tổ hợp âm tiết*", "*bằng chứng có*",... Trường hợp trong câu có âm tiết không nằm trong từ điển thì rõ ràng ôtômát âm tiết không đoán nhận được âm tiết này. Kết quả là đồ thị ta xây dựng từ câu đó là không liên thông.

Dựa vào tính chất này, ta thấy rằng nếu đồ thị không liên thông thì dễ dàng phát hiện ra rằng đơn vị âm tiết không đoán nhận được không nằm trong từ điển âm tiết, tức nó bị viết sai chính tả hoặc là một đơn vị âm tiết (từ vựng) mới.

Để triển khai được thuật toán nói trên cần có một cơ sở dữ liệu lớn và hoàn chỉnh, đặc biệt cần xây dựng được đồ thị nối giữa các từ tiếng Việt. Với số lượng gần 74.000 từ tiếng Việt, đây là một công việc đòi hỏi sự đầu tư thời gian và trí tuệ của nhiều người. Do đó, trong thời gian hạn chế của việc thực hiện luận văn tốt nghiệp, tác giả chọn một giải pháp khác để đảm bảo xây dựng được một chương trình demo hỗ trợ xử lý một số hiện tượng nhập nhằng cụ thể. Trên cơ sở ý tưởng của thuật toán được đề cập trong tài liệu **Error! Reference source not found.**, tác giả xây dựng cho mình một giải thuật khác, giải thuật này cũng dựa trên tính chất "*cách phân tách tối ưu nhất là tách được những từ có nhiều âm tiết nhất*".

Trước tiên chương trình sẽ tiến hành kiểm tra lỗi chính tả ở mức âm tiết để đảm bảo rằng văn bản đã được viết đúng chính tả tiếng Việt ở mức thấp nhất là mức âm tiết, sau đó thay vì tiến hành đọc vào từng âm tiết và kiểm tra tính liên thông (như thuật toán đã đề cập trên) thì sẽ đọc vào một âm tiết (gọi là âm tiết X) là âm tiết đầu tiên của phần văn bản sẽ được xử lý, sau đó kiểm tra sự tồn tại của cụm từ dài nhất chứa âm tiết vừa đọc (gọi là từ Y) có tồn tại trong tiếng Việt hay không, nếu tồn tại thì xem như đây là cách tách từ tối ưu nhất và không chia nhỏ cụm từ Y, nếu không tồn tại thuật toán sẽ

tiến hành kiểm tra tương tự với các từ ngắn hơn (bằng cách chia nhỏ cụm từ Y).

Xét cụm từ *"thuộc địa bàn"*, X = *"thuộc"*, MaxLen (của từ bắt đầu bằng âm tiết *"thuộc"*) = 2, ta lấy được từ *"thuộc địa"*, theo trình tự trình bày trên đây, âm tiết tiếp theo được xét sẽ là *"bàn"*, như thế ta đã bỏ qua từ *"địa bàn"*, kết quả không tối ưu. Do đó, thuật toán phải lưu vết hiện tại và quay về xét các trường hợp có thể xảy ra với từ *"địa"*.

Chương trình sẽ đưa ra tất cả các khả năng có thể sau đó đánh giá xem kết quả nào là tối ưu nhất trên cơ sở từ chứa nhiều âm tiết hơn luôn có độ ưu tiên cao hơn. Trong một số trường hợp mà chương trình không thể tự đưa ra quyết định phân tách được, như từ *"thuộc địa"* và *"địa bàn"* trên đây, chương trình sẽ đưa gợi ý để người sử dụng lựa chọn tùy theo ngữ cảnh của văn bản.

Theo thuật toán trình bày trên đây, sẽ không xảy ra trường hợp xuất hiện âm tiết không có trong từ điển tiếng Việt, vì văn bản đầu vào của thuật toán này buộc phải kiểm tra lỗi chính tả tiếng Việt ở mức âm tiết rồi.

3.3. CÀI ĐẶT

3.3.1. Môi trường làm việc

3.3.2. Khái quát vắn tắt về VB6

3.3.2.1. Điều khiển các ứng dụng Microsoft Office

3.3.2.2. Tạo một COM Add-In với Visual Basic

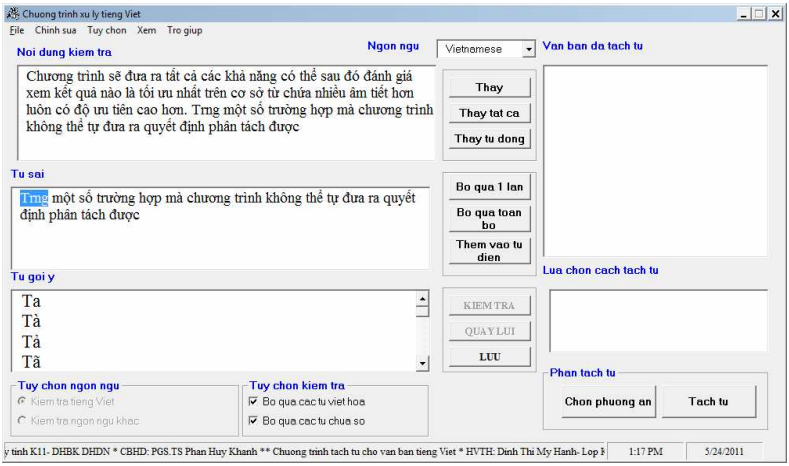
3.3.2.3. Kiểm tra COM Add-In

3.4. GIỚI THIỆU GIAO DIỆN CHƯƠNG TRÌNH VÀ HƯỚNG DẪN SỬ DỤNG

3.4.1. Giao diện chính của chương trình

3.4.2. Chức năng kiểm lỗi chính tả tiếng Việt mức âm tiết

Người sử dụng chọn chức năng **KIỂM TRA** để bắt đầu kiểm lỗi chính tả tiếng Việt mức âm tiết. Trong phần **Từ sai**, những từ không có trong dữ liệu âm tiết tiếng Việt sẽ được bôi xanh, đồng thời mục **Từ gợi ý** sẽ đưa ra danh sách các từ gợi ý để thay thế, đây là những từ được lấy từ cơ sở dữ liệu của chương trình. Người sử dụng sẽ lựa chọn các nút lệnh để thực hiện sửa lỗi hoặc bỏ qua từ bị lỗi.



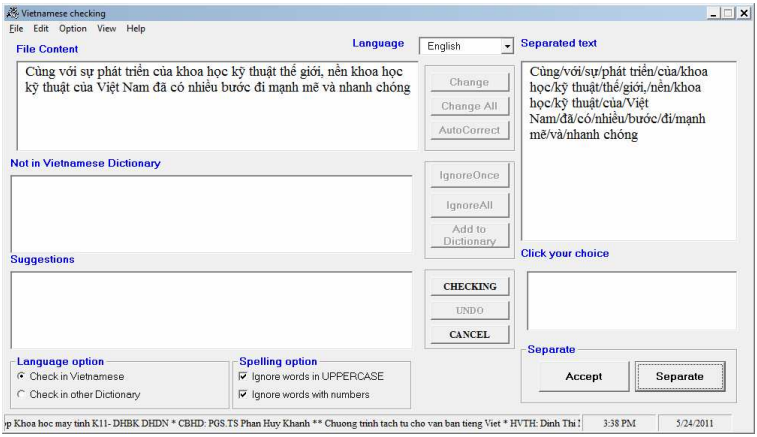
Hình 3.9 Giao diện chính của chương trình

3.4.3. Chức năng tách từ

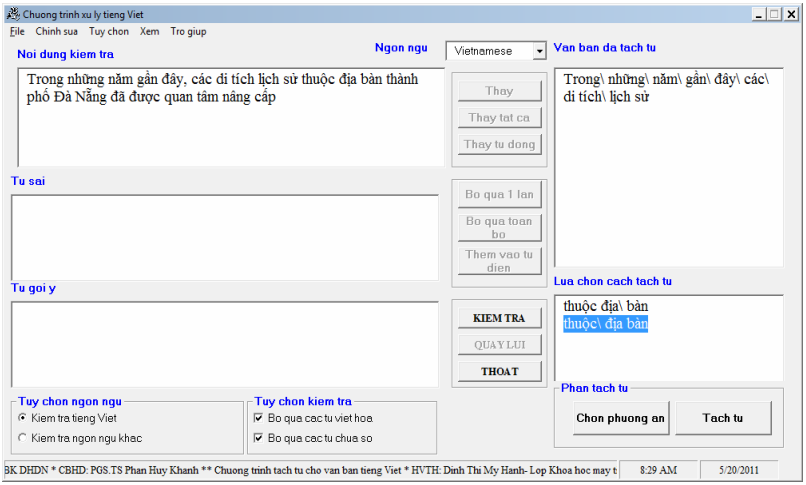
Để sử dụng chức năng tách từ cho văn bản, đầu tiên người sử dụng phải lấy phần văn bản muốn tách đưa vào phần **Nội dung kiểm tra**, sau đó bấm nút **Tách từ**.

Nếu trong đoạn văn bản cần tách không có trường hợp nào có nhiều cách phân tách (như trường hợp cụm từ "*thuộc địa bàn*" đã nêu ở trên) thì chương trình sẽ tự động tách và đưa ra phần văn bản đã được tách thành các từ riêng biệt trong phần **Văn bản đã được tách từ** (xem hình 3.10)

Ngược lại, nếu văn bản xuất hiện những cụm từ có thể phân tách theo nhiều cách khác nhau, chương trình sẽ dừng lại ở cụm từ đó và đưa ra tất cả các cách phân tách có thể trong phần **Lựa chọn cách tách từ** (xem hình 3.11), lúc đó người sử dụng sẽ lựa chọn cách tách phù hợp với ngữ cảnh văn bản và bấm nút **Chọn phương án** để chấp nhận, chương trình sẽ lưu lựa chọn này và tiếp tục xử lý phần văn bản còn lại.



Hình 3.10 Giao diện chức năng tách từ (giao diện tiếng Anh)



Hình 3.11 Người sử dụng lựa chọn phương án tách văn bản

3.5. KẾT LUẬN CHƯƠNG

Chương 3 của báo cáo này tập trung trình bày những đề xuất để khắc phục HTNN trong STVB. Trong phần này tác giả cũng đã nhắc lại những kết quả mà một số công trình nghiên cứu đã đạt được trong lĩnh vực xử lý nhập nhằng tiếng Việt, đồng thời đưa ra một mô hình tổng quan để xử lý bài toán. Phạm vi xử lý là hỗ trợ khắc phục hiện tượng nhập nhằng gây ra do những lỗi chính tả tiếng Việt ở cấp độ âm tiết, đồng thời hỗ trợ xử lý nhập nhằng do không xác định được ranh giới giữa các từ.

KẾT LUẬN

1. Đánh giá kết quả chương trình

Sau một thời gian nghiên cứu và thực hiện đề tài, tác giả đã đạt được một số kết quả về mặt lý thuyết và ứng dụng, cụ thể như sau:

Cơ sở lý thuyết

Nghiên cứu được những vấn đề cơ bản liên quan đến xử lý ngôn ngữ tự nhiên nói chung và xử lý tiếng Việt nói riêng. Nghiên cứu về những hiện tượng nhập nhằng thường xuất hiện trong tiếng Việt, tìm hiểu những nội dung chính về soạn thảo văn bản và hiện tượng nhập nhằng trong soạn thảo văn bản. Tìm hiểu về các công trình, các bài báo nghiên cứu về xử lý nhập nhằng trong ngôn ngữ tự nhiên nói chung và tiếng Việt nói riêng.

Xây dựng ứng dụng

Xây dựng được một chương trình hỗ trợ xử lý nhập nhằng với một số chức năng chính sau:

Sửa lỗi chính tả tiếng Việt cho văn bản ở mức độ âm tiết; Chức năng chính: tách đoạn văn bản cho trước thành các từ riêng biệt, chức năng này đã giải quyết được những nhập nhằng về ranh giới từ trong văn bản, đồng thời kết quả này có thể dùng tiếp cho các ứng dụng phát triển về sau.

Về cơ sở dữ liệu:

Tác giả đã xây dựng được một kho dữ liệu gồm khoảng hơn 10.000 âm tiết tiếng Việt. Dữ liệu này phục vụ cho chức năng kiểm lỗi chính tả tiếng Việt ở mức âm tiết. Ngoài ra để phục vụ cho chức năng tách từ trong văn bản tiếng Việt, tác giả đã xây dựng được một kho dữ liệu gồm khoảng 24.000 từ có trong tiếng Việt và bổ sung một số danh từ riêng phổ biến.

Nhận xét

Ưu điểm:

Chương trình được tích hợp sẵn trên ứng dụng Microsoft Word nên người sử dụng dễ gọi thực thi. Giao diện chương trình đơn giản, thân thiện nên dễ sử dụng, hệ thống menu và các nút lệnh được thiết kế rõ ràng, logic giúp người dùng dễ thích nghi. Kết quả thể hiện rõ ràng, gợi ý hỗ trợ cụ thể.

Một số hạn chế:

Chương trình chỉ mới hỗ trợ xử lý hiện tượng nhập nhằng về phạm vi, ranh giới từ và một phần của hiện tượng nhập nhằng gây ra do sai chính tả tiếng Việt chứ chưa giải quyết được tất cả các hiện tượng nhập nhằng của tiếng Việt. Kết quả thực thi còn mang tính chất hỗ trợ chứ chưa giải quyết triệt để hiện tượng nhập nhằng.

Kho dữ liệu từ tiếng Việt chưa đầy đủ và dù tác giả đã bổ sung các danh từ riêng phổ biến không có trong từ điển tiếng Việt nhưng chưa thể đầy đủ tất cả nên ở một chừng mực nào đó, kết quả chương trình vẫn chưa chính xác 100%. Chương trình chưa hỗ trợ xử lý trực tiếp khi đang soạn thảo văn bản và chưa xử lý các định dạng văn bản.

2. Hướng phát triển của đề tài

Hoàn thiện kho dữ liệu từ tiếng Việt để kết quả phân tích của chương trình có độ chính xác cao hơn. Phát triển ứng dụng có khả năng lấy xử lý được các định dạng của văn bản. Xử lý thêm các trường hợp lỗi chính tả tiếng Việt ở cấp độ cao hơn. Với các từ đã phân tách được, tác giả sẽ tiến hành gán nhãn từ loại, hỗ trợ phân tích cú pháp câu tiếng Việt. Trợ giúp người sử dụng ngay trong quá trình soạn thảo văn bản.