Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt

TRẦN THỊ OANH

Master Thesis

Giảng viên hướng dẫn: TS. Lê Anh Cường

PGS.TS. Hà Quang Thụy

2008

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của bản thân. Các số liệu, kết quả trình bày trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào trước đây.

LỜI CẢM ƠN

Trước tiên, tôi xin gửi lời cảm ơn chân thành và sự biết ơn sâu sắc tới PGS.TS Hà Quang Thuỵ và TS Lê Anh Cường (Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội) đã tận tình hướng dẫn tôi trong suốt quá trình thực hiện khoá luận này.

Tôi xin chân thành cảm ơn các thày cô giáo đã giảng dạy tôi trong suốt các năm tôi theo học tại trường Đại học công nghệ đã cho tôi những kiến thức quí báu để tôi có thể vững bước trên con đường đi của mình.

Tôi xin gửi lời cảm ơn các anh chị em trong nhóm seminar về khai phá dữ liệu ở phòng thí nghiệm Các hệ tích hợp thông minh (SISLAB) trường Đại học Công nghệ đã nhiệt tình chỉ bảo trong quá trình tôi tham gia nghiên cứu khoa học và thực hiện luận văn này.

Và lời cuối cùng, tôi xin bày tỏ lòng chân thành và biết ơn vô hạn tới cha mẹ, và các anh chị tôi, những người luôn ở bên cạnh tôi những lúc tôi khó khăn nhất, giúp tôi vượt qua khó khăn trong học tập cũng như trong cuộc sống.

Hà Nội, ngày 30 tháng 12 năm 2008

Sinh viên

Trần Thị Oanh

MỞ ĐẦU

Phân đoạn từ (Word segmentation) và gán nhãn từ loại(Part-of-speech tagging – POS tagging) là hai bài toán đặc biệt quan trọng trong xử lý ngôn ngữ tự nhiên. Tuy nhiên, các nghiên cứu tại Việt Nam về hai vấn đề này vẫn còn ở giai đoạn ban đầu. Do đó, nhu cầu là rất lớn cả về cơ sở khoa học và xây dựng công cụ thực hiện sẵn dùng. Thực tế cho thấy hai quá trình này có liên quan với nhau và ảnh hưởng đến chất lượng của một hệ chung. Vì thế, trong luận văn này chúng tôi nghiên cứu **"Mô hình tách từ, gán nhãn từ loại và hướng tiếp cận tích hợp cho tiếng Việt"**. Đóng góp của luận văn là việc tìm hiểu, nghiên cứu và đề xuất mô hình thực hiện tách từ, gán nhãn từ loại POS tiếng Việt; xây dựng công cụ thực hiện liên quan; bên cạnh đó để huấn luyện và kiểm thử mô hình chúng tôi cũng tiến hành xây dựng một corpus tiếng Việt về tách từ và gán nhãn từ loại xấp xỉ 8000 câu. Luận văn cũng tiến hành thực nghiệm một hướng tiếp cận tích hợp cho hai bài toán này. Các kết quả này có sẽ đặc biệt hữu ích cho các nghiên cứu ở mức cao hơn như dịch máy, tóm tắt văn bản, phân tích cú pháp ...
Cấu trúc của luận văn được trình bày như sau:

- **Chương 1: Khái quát về tách từ và gán nhãn từ loại tiếng Việt** : Trong chương này, luận văn giới thiệu hai bài toán cơ bản trong xử lý ngôn ngữ tự nhiên là bài toán tách từ và bài toán gán nhãn từ loại tiếng Việt - các hướng tiếp cận cho mỗi bài toán và tình hình nghiên cứu chung ở Việt Nam cũng như trên thế giới. Chương cũng trình bày các hướng tiếp cận tích hợp hai bài toán này để nâng cao hiệu quả của cả hai mô hình đã được áp dụng thành công cho tiếng Trung.

- **Chương 2: Mô hình tách từ tiếng Việt** : Chương này nghiên cứu và đề xuất một mô hình cho bài toán tách từ tiếng Việt tận dụng thông tin từ tri thức từ nhiều nguồn khác nhau nhằm làm tăng độ chính xác của bộ tách từ.

- **Chương 3: Mô hình gán nhãn từ loại tiếng Việt:** Chương này nghiên cứu và đề xuất mô hình gán nhãn từ loại tiếng Việt, các công việc mà luận văn đã tiến hành nhằm xây dựng một mô hình gán nhãn POS hiệu quả như thiết kế corpus gán

nhãn từ loại, đề xuất mô hình sử dụng Maximum Entropy Markov Model (MEM) và thiết kế các tập đặc trưng khác nhau để tìm ra các đặc trưng hữu ích cho tiếng Việt.

- **Chương 4: Mô hình tích hợp tách từ và gán nhãn từ loại tiếng Việt:** Từ các nghiên cứu đã trình bày trong chương 2 và chương 3 và đặc điểm tiếng Việt, chương này trình bày một mô hình tích hợp áp dụng cho tiếng Việt.

- **Kết luận:** Phần này tóm tắt lại nội dung của luận văn và những đóng góp chính mà luận văn đã thực hiện.

- **Phụ lục A: Một số thuật ngữ Anh - Việt :** Một số thuật ngữ tiếng Anh hay dùng và chữ viết tắt.

- **Phụ lục B: Chú giải tập từ loại vnPOS :** Mô tả cụ thể để giải thích thêm và các nhãn từ loại mà luận văn đề xuất để xây dựng corpus gán nhãn từ loại cho tiếng Việt.

Mục lục

Danh sách hình vẽ

Danh sách bảng

Chương 1

Khái quát về tách từ và gán nhãn từ loại tiếng Việt

1.1 Khái quát về tách từ tiếng Việt

Tiếng Việt là một ngôn ngữ đơn lập, không biến hình, các ký tự được dựa trên hệ chữ cái Latin. Cũng giống như với các thứ tiếng khác như tiếng Trung, tiếng Nhật, tiếng Hàn thì từ trong tiếng Việt không được xác định bởi khoảng trắng. Một từ tiếng Việt có thể được tạo bởi một hoặc nhiều hình vị và mỗi hình vị phân tách nhau bởi các khoảng trắng. Từ là đơn vị cơ bản để phân tích cấu trúc của ngôn ngữ, do vậy để tiến tới những ứng dụng xa hơn về xử lý ngôn ngữ tiếng Việt như gán nhãn chức năng cú pháp, phân tích cú pháp thì việc đầu tiên ta phải giải quyết bài toán phân đoạn từ.

Các nhà nghiên cứu đã đề xuất một số hướng tiếp cận để giải bài toán phân đoạn từ [3, 5, 9, 10, 12, 13, 11] . Nhìn chung, các hướng tiếp cận đó được chia thành 2 loại: tiếp cận dựa trên từ điển và tiếp cận dựa trên thống kê. Hai phương pháp tiêu biểu của hướng tiếp cận dựa vào từ điển là Longest Matching và Maximal Matching. Hầu hết những nghiên cứu khởi thủy về phân đoạn từ đều dựa trên cách tiếp cận này như trong [13][18]. Hướng tiếp cận này có đặc điểm là đơn giản, dễ hiểu tuy nhiên hiệu quả mang lại không cao. Lý do là bởi nó không xử lý được rất nhiều trường hợp nhập nhằng cũng như không có khả năng phát hiện từ mới trong văn bản. Chính vì vậy mà các hệ thống phân đoạn từ có chất lượng cao hiện nay thường sử dụng hướng tiếp cận dựa trên thống kê. Ví dụ như đối với tiếng Trung thì có các nghiên cứu liên quan như [9, 12], tiếng Thái [13] cũng như

tiếng Việt [3, 8]. Cụ thể, đối với tiếng Việt thì tác giả Lê An Hà[10] đã xây dựng corpus 10M và sử dụng thông tin N-gram để tối ưu tổng các xác suất phân đoạn cho mỗi phân cụm (chunk). Kết quả thực nghiệm tuy không cao nhưng cũng đã cho thấy N-gram sẽ trở nên hữu ích nếu ta biết cách sử dụng thông tin này khi liên kết với các nguồn thông tin khác. Hiện nay, rất nhiều hệ thống phân đoạn từ phổ biến sử dụng hướng tiếp cận lai. Ví dụ, nhóm tác giả Cẩm Tú [3] đã nghiên cứu ứng dụng các mô hình CRF và SVM để phân đoạn từ tiếng Việt. Hoặc như nhóm tác giả Đinh Điền [8] đã sử dụng mô hình MEM sử dụng giải thuật tối ưu GIS để huấn luyện bộ phân đoạn trên corpus gán nhãn. Trong nghiên cứu đó, tác giả đã phân tách hai quá trình nhận dạng từ mới (unknown word recognition) và phân đoạn từ đã biết (known word segmentation) như hai tiến trình độc lập nhau. Tuy nhiên, chúng tôi nhận thấy rằng hai tiến trình này nên được tiến hành đồng thời để nâng cao độ chính xác. Một ví dụ điển hình của hướng tiếp cận như vậy cho tiếng Trung được J.Gao đề cập trong [9].

Trong các phương pháp lai, các tác giả đã tận dụng thông tin từ điển và một số thông tin khác nhằm phát hiện tên thực thể. Tuy nhiên, trong các nghiên cứu đó đều chưa quan tâm thích đáng tới việc đánh giá ảnh hưởng của từng nguồn tri thức và đặc biệt là chưa có những nghiên cứu để phát hiện từ mới (không chỉ là tên thực thể và các dạng factoid). Nghiên cứu các phương pháp phát hiện từ mới ta thấy thông tin Ngram đóng vai trò hữu ích giúp ta phát hiện từ mới khi mà corpus thống kê đủ lớn và xác định được độ đo phù hợp. Câu hỏi đặt ra là làm cách nào để tận dụng được từ tất cả các nguồn tri thức đó. Đây cũng chính là động lực cho luận văn này.

1.2 Khái quát về gán nhãn từ loại - POS tagging

1.2.1 Giới thiệu về bài toán gán nhãn từ loại

Gãn nhãn từ loại là một công việc quan trọng và bắt buộc phải có đối với mọi hệ xử lý ngôn ngữ tự nhiên. Công việc gán nhãn từ loại cho một văn bản là xác định từ loại của mỗi từ trong phạm vi văn bản đó, tức là phân loại các từ thành các lớp từ loại dựa trên thực tiễn hoạt động ngôn ngữ trong đó:

- **Input:** Một chuỗi các từ và tập nhãn từ loại (Ví dụ đối với tiếng Anh: “Book that flight.”, và tập thẻ Penn Treebank)

- **Output:** Một nhãn tốt nhất cho từng từ trong câu (Ví dụ: Book/VB that/DT flight/NN ./.)

Quá trình gán nhãn từ loại có thể chia làm 3 bước như sau:

1. **Giai đoạn tiền xử lý:** Phân tách xâu ký 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 ký hiệu trắng. Tuy nhiên vẫn có những từ ghép hay những cụm từ gây tranh cãi về cách xử lý. 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.

2. **Khởi tạo gán nhãn:** 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 ngữ liệu 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.

3. **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 khởi tạo nói trên. 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 và các phương pháp xác suất. Ngoài ra còn có các hệ thống sử dụng mạng nơ-ron, các hệ thống lai sử dụng kết hợp tính toán xác suất và ràng buộc ngữ pháp, gán nhãn nhiều tầng, ...

Việc gán nhãn từ loại đã được quan tâm từ rất sớm, cùng với nó là sự xuất hiện của rất nhiều phương pháp giải quyết. Tới nay, các phương pháp mới vẫn đang tiếp tục được nghiên cứu nhằm hoàn thiện hơn nữa các kết quả đã đạt được.

Hiện nay, bài toán gán nhãn từ loại cho tiếng Anh đã được giải quyết khá tốt, đạt kết quả rất khả quan. Bên cạnh việc hoàn thiện hơn nữa các bộ gán nhãn đã có, ngày càng nhiều bộ gán nhãn mới ra đời, đem lại kết quả gần như tối ưu. Tuy nhiên, đối với các ngôn ngữ khác, đặc biệt là các ngôn ngữ tượng hình (như tiếng Trung Quốc, Nhật, Hàn Quốc ..), các ngôn ngữ của Ấn Độ, Thái Lan, A Rập, Nga cũng như đối với tiếng Việt

thì bài toán gán nhãn từ loại vẫn còn là một thách thức lớn. Các phương pháp và công cụ đã được xây dựng gần như hoàn thiện cho Tiếng Anh khi đem áp dụng cho các ngôn ngữ khác loại trên thường đưa lại kết quả thấp. Như vậy, yêu cầu đặt ra với từng ngôn ngữ là phải kế thừa, tận dụng được các phương pháp sẵn có, tiến hành hiệu chỉnh hoặc là đề xuất ra các hướng tiếp cận mới sao cho phù hợp với các đặc điểm riêng biệt của ngôn ngữ mình.

1.2.2 Các hướng tiếp cận bài toán POS tagging

Theo [4], hầu hết các thuật toán gán nhãn từ loại rơi vào một trong hai lớp: gán nhãn dựa trên luật (rule-based) hoặc bộ gán nhãn xác suất (stochastic taggers).

Các bộ gán nhãn dựa trên luật thường liên quan tới một cơ sở dữ liệu lớn các luật được viết bằng tay. Ví dụ một từ nhập nhằng đang xét có xu hướng là một danh từ hơn là một động từ nếu nó đi sau một từ chỉ định. Phần tiếp sau sẽ mô tả một bộ gán nhãn dựa trên luật mẫu, ENGTWOL, dựa trên kiến trúc cú pháp ràng buộc của Karlson năm 1995.

Bộ gán nhãn xác suất thường giải quyết nhập nhằng bằng cách sử dụng một corpus huấn luyện để tính toán xác suất của một từ cho sẵn sẽ được gán một thẻ nào đó trong ngữ cảnh cho trước. Phần sau sẽ mô tả một bộ gán nhãn HMM (HMM Tagger), hay còn được gọi là Maximum Likelihood Tagger, hoặc một bộ gán nhãn Markov Model, cũng dựa trên mô hình Markov ẩn.

Ngoài ra còn có các hướng tiếp cận khác gồm bộ gán nhãn dựa trên biến đổi transformation-based tagger hoặc bộ gán nhãn Brill (Brill tagger). Bộ gán nhãn Brill sẽ sử dụng các đặc tính của cả 2 kiến trúc gán nhãn trên. Giống như bộ gán nhãn dựa trên luật, nó dựa vào luật để xác định khi một từ nhập nhằng thì nó có khả năng là một thẻ nào nhất. Giống như bộ gán nhãn xác suất, nó có một thành phần học máy để tạo ra các luật một cách tự động từ một corpus huấn luyện đã được gán nhãn trước. Tuy nhiên, trong phạm vi luận văn này chúng tôi không trình bày cụ thể việc nghiên cứu 2 phương pháp này (xem thêm trong [4]).

Gán nhãn chức năng cú pháp dựa trên luật

Các thuật toán khởi thủy gán nhãn tự động từ loại thường gồm hai giai đoạn. Giai đoạn một nó sử dụng một từ điển để gán cho mỗi từ một danh sách các từ loại có thể có. Giai

đoạn 2 nó sử dụng một danh sách gồm tập các luật không có nhập nhằng thường được soạn bằng tay để gán cho mỗi từ chỉ một từ loại phù hợp nhất. Mộ bộ gán nhãn điển hình áp dụng cho tiếng Anh là bộ gán nhãn **ENGTWOL**[4].

Gán nhãn từ loại xác suất

Phần này trình bày một bộ gán nhãn xác suất điển hình sử dụng mô hình Markov ẩn. Thuật toán này lựa chọn chuỗi nhãn tốt nhất cho toàn bộ câu. Và thông thường người ta hay sử dụng thuật toán Viterbi để tìm chuỗi thẻ tốt nhất đó. Giả sử với câu đầu vào là W ta cần tìm một chuỗi thẻ $T{=}t_1, ..., t_n$ thỏa mãn công thức 1.1:

$$\widehat{T} = argmax_{T \in \tau} P(T|W) \tag{1.1}$$

Sử dụng luật Bayes, P(T|W) được viết theo công thức 1.2

$$P(T|W) = \frac{P(T)P(W|T)}{P(W)} \tag{1.2}$$

Ta đang quan tâm tới tìm chuỗi thẻ phù hợp nhất làm cực đại công thứ 3.1 nên mẫu số trong tất cả các trường hợp là giống nhau. Do vậy, bài toán trở thành tìm chuỗi thẻ thỏa mãn công thức 1.3

$$\widehat{T} = argmax_{T \in \tau} P(T)P(W|T) \tag{1.3}$$

Áp dụng luật chuỗi xác suất ta có công thức 1.4:

$$P(T)P(W|T) = \prod_{i=1}^{n} P(w_i|w_1t_1...w_{i-1}t_{i-1}t_i)P(t_i|w_1t_1...w_{i-1}t_{i-1}) \tag{1.4}$$

Vẫn không có phương pháp hiệu quả để tính toán xác xuất của chuỗi này một cách chính xác, nó yêu cầu quá nhiều dữ liệu. Tuy nhiên, xác suất có thể được xấp xỉ bởi một xác suất đơn giản hơn bằng cách áp dụng các giả thiết độc lập điều kiện. Mặc dù các giả thiết này là không thực tế nhưng trong thực hành thì việc đánh giá đó là vẫn hợp lý. Ở đây, ta sử dụng giả thiết N-gram để mô hình hóa xác suất chuỗi từ. Cụ thể ta dùng mô hình phổ biến nhất là mô hình tri-gram. Đầu tiên, ta làm đơn giản hóa rằng xác suất của một từ thì chỉ phụ thuộc vào thẻ của nó (xem công thức 1.5):

$$P(w_i|w_1t_1...w_{i-1}t_{i-1}t_i) = P(w_i|t_i) \tag{1.5}$$

Tiếp đến, ta giả thiết rằng các thẻ phía trước có thể được xấp xỉ bởi 2 thẻ gần nó nhất (xem công thức 1.6):

$$P(t_i|w_1t_1...w_{i-1}t_{i-1}) = P(t_i|t_{i-2}t_{i-1}) \tag{1.6}$$

Vì vậy cuối cùng ta lựa chọn chuỗi thẻ làm cực đại công thức 1.7:

$$P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2}t_{i-1})[\prod_{i=1}^{n} P(w_i|t_i)] \tag{1.7}$$

Các thành phần thừa số trong công thức 1.7 có thể được tính toán từ corpus huấn luyện của mô hình. Chú ý rằng để có thể tránh xác suất bằng 0 ta cần sử dụng các kỹ thuật làm trơn.

1.2.3 Các nghiên cứu gán nhãn từ loại cho tiếng Việt

Đối với tiếng Anh thì bài toán này gần như đã được giải quyết xong đạt độ chính xác rất cao lên tới >96% [1]. Tuy nhiên, đối với các văn bản Việt ngữ, việc gán nhãn từ loại có nhiều khó khăn, đặc biệt là bản thân việc phân loại từ tiếng Việt cho đến nay vẫn là một vấn đề còn nhiều tranh cãi, chưa có một chuẩn mực thống nhất.

Hiện nay đã có một số nghiên cứu gán nhãn từ loại cho tiếng Việt và đạt được một số thành tựu nhất định. Điển hình là bộ gán nhãn từ loại xác suất vnQTAG của nhóm tác giả Nguyễn Thị Minh Huyền [16]. Ý tưởng của phương pháp 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. Nhóm tác giả Nguyễn Quang Châu [15] trình bày một hướng tiếp cận cho bài toán gán nhãn từ loại trong văn bản tiếng Việt trên cơ sở vận dụng các mô hình thống kê dựa vào kho ngữ liệu, từ điển, cú pháp và ngữ cảnh. Ngoài ra còn một hướng tiếp cận khác sử dụng kho ngữ liệu song ngữ Anh-Việt [6]. Các hướng tiếp cận này có ưu điểm là tận dụng được các công cụ đã phát triển gần

như hoàn thiện dùng cho tiếng Anh tuy nhiên nhược điểm của nó là: Do sự khác nhau về hình thái giữa tiếng Anh và tiếng Việt nên phép chiếu trực tiếp không đơn giản là phép chiếu 1-1 mà thường là phép chiếu phức tạp m-n. Sự khác nhau về cơ bản giữa hai ngôn ngữ này là nguyên nhân của không ít nhập nhằng cần phải giải quyết, nó còn có thể tiềm tàng nhiều trường hợp mà nhóm tác giả cũng như các độc giả chưa tính tới.

1.3 Vấn đề tích hợp tách từ và gán nhãn từ loại

Ở Việt Nam chưa có một công trình nghiên cứu nào về tích hợp hai bài toán rất quan trọng trong xử lý ngôn ngữ tự nhiên là bài toán tách từ và bài toán gán nhãn từ loại tiếng Việt. Các nghiên cứu này chủ yếu mới có nghiên cứu cho tiếng Trung như [22][14][20]. Ý tưởng của phương pháp tích hợp là có thể kết hợp hai tiến trình lại với nhau nhằm nâng cao hiệu quả của chúng. Các hướng tích hợp có thể chia làm 2 loại: Một là loại tích hợp giả pseudo-integration và một loại là tích hợp thực sự true-integration.
Hướng tích hợp giả: [19] mô tả một phương pháp gồm 3 bước chính:

1. Tạo ra N chuỗi tách từ tốt nhất(N-best word sequences) đối với một câu cho sẵn.
2. Thực hiện gán nhãn POS cho mỗi chuỗi từ đó, sau đó chọn ra N chuỗi thẻ POS tốt nhất tương ứng.
3. Sử dụng đánh giá có trọng số của (1) và (2) để chọn giải pháp tách từ và gán nhãn POS tốt nhất cho câu đầu vào đó.

Trong hệ thống này, việc giải mã cho tách từ và gán POS vẫn được thực hiện riêng rẽ, và sự suy luận chính xác cho cả hai là điều có thể. Tuy nhiên, sự tương tác giữa POS và segmentation bị hạn chế bởi reranking: thông tin POS được sử dụng để cải tiến chất lượng phân đoạn đối với chỉ N segmentor output.

Hướng tích hợp thực sự tiêu biểu trong hai công trình [19, 14] Trong [19] các tác giả đề xuất một phương pháp dựa trên CRFs hai tầng sử dụng giải mã đồng thời tách từ và gán POS. Trong phương pháp này, tác giả mô hình bài toán tách từ và gán nhãn bằng một CRFs hai tầng. Lúc giải mã, đầu tiên thực hiện giải mã riêng ở mỗi tầng. Sau đó, một khung xác suất được xây dựng để tìm ra giải mã kết hợp tốt nhất cho cả hai bài toán.

Còn khi huấn luyện, tác giả huấn luyện một lần các CRF riêng đó cho hai bài toán, đối với phạm vi ứng dụng này thì huấn luyện đồng thời sẽ tốn công hơn. Kết quả đánh giá tách từ và POS tag thu được kết quả state-of-the-art trên cả tập PCT và First SIGHAN Bakeoff datasets. Trong cả hai bài toán, phương pháp đề xuất cải tiến so với phương pháp baseline không thực hiện giải mã đồng thời.

Trong [14] trình bày một nghiên cứu tích hợp khá công phu. Để xây dựng một bộ gán nhãn POS, có hai câu hỏi được đặt ra:

1. Thực hiện gán nhãn sau khi tách từ theo hai pha riêng biệt (one-at-a-time), hoặc thực hiện liên kết gán nhãn từ loại và tách từ thành một bước đơn đồng thời nhau (all-at-one approach).

2. Gán thẻ POS dựa trên nền tảng từ (giống English), tận dụng các đặc trưng mức từ của ngữ cảnh (word-based), hoặc dựa trên nền tảng ký tự với các đặc trưng của ký tự (character-based)?

Bài báo trình bày một nghiên cứu tỉ mỉ về kiến trúc xử lý và biểu diễn đặc trưng cho gán POS tiếng Trung với khung Maximum Entropy. Họ phân tích hiệu quả của từng tiếp cận nhằm tìm ra hướng tiếp cận phù hợp nhất. Kết quả thực nghiệm cho thấy tiếp cận character-based tốt hơn so với tiếp cận dựa trên word-based đối với bài toán POS tag là không có gì đáng ngạc nhiên. Khác với English mà mỗi English letter không có nghĩa, thì nhiều character tiếng Trung lại mang nghĩa. Hơn nữa, tỷ lệ OOV đối với Chinese words thì cao hơn so với Chinese characters, đối với unknown words, việc sử dụng các character thành phần trong từ giúp để dự đoán chính xác nhãn POS là một heuristic tốt. Tiếp cận all-at-once xem xét tất cả các khía cạnh của thông tin sẵn có theo một khung tích hợp đồng nhất cho kết quả tốt hơn nhưng cũng yêu cầu chi phí tính toán cao hơn. Tuy nhiên, điểm bất lợi của phương pháp này là sự khó khăn khi tích hợp toàn bộ thông tin về từ vào việc gán POS. Ví dụ, đặc trưng chuẩn "word + POS tag" sẽ không thể ứng dụng rõ ràng được.

Chương 2

Mô hình tách từ tiếng Việt

Trong luận văn này, chúng tôi chọn mô hình maximum entropy làm phương pháp học máy trong đó các đặc trưng của mô hình được lựa chọn dựa trên những nghiên cứu về tri thức của các mô hình khác và các đặc điểm của ngôn ngữ tiếng Việt. Cụ thể, chúng tôi sử dụng thông tin có được từ ba nguồn là mô hình phân đoạn từ dựa vào từ điển, mô hình N-gram và mô hình nhận dạng thực thể. Chúng tôi cũng làm những thực nghiệm để đánh giá tính hiệu quả của hệ thống dựa trên tập dữ liệu đã gán nhãn. Bên cạnh đó, chúng tôi đánh giá ảnh hưởng của từng nguồn tri thức đó đối với mô hình cuối cùng. Trong thực nghiệm, chúng tôi lấy phương pháp Longest Matching làm mô hình cơ sở (baseline) để so sánh.

2.1 Các mô hình liên quan

2.1.1 Mô hình dựa vào từ điển

Hai phương pháp kinh điển của hướng tiếp cận dựa trên từ điển là: Longest Matching (LM) và Maximal Matching (MM).

Phương pháp LM duyệt câu đầu vào tuần tự từ trái qua phải và chọn từ dài nhất nếu từ đó có trong từ điển. Rõ ràng là phương pháp này rất đơn giản nhưng bị phân lớp sai trong nhiều trường hợp nhập nhằng. Ví dụ câu *"Đó là cách để truyền thông tin"*, nếu áp dụng phương pháp LM thì câu này sẽ bị phân tách sai thành *"Đó là cách để truyền_ thông tin"*.

Phương pháp MM sẽ tạo ra tất cả các phân đoạn có thể cho một câu bất kỳ, sau đó câu

được phân đoạn đúng được chọn là câu chứa ít từ nhất. Giống như phương pháp trên phương pháp này cũng có yếu điểm là không thể đưa ra phân đoạn đúng trong trường hợp nhiều kết quả phân đoạn lại chứa cùng một số lượng từ ít nhất. Ví dụ câu *"Học sinh học sinh học"* có hai ứng cử là *"Học_ sinh học sinh_ học"* và *"Học sinh_ học sinh_ học"*. Trong những trường hợp này, ta cần áp dụng các phương pháp học máy trên một cơ sở dữ liệu lớn để xác định được phân đoạn đúng.

2.1.2 Mô hình nhận dạng tên thực thể - Named Entity Recognition

Bài toán nhận dạng tên thực thể là bài toán gán nhãn mỗi từ trong văn bản vào một trong các lớp được định nghĩa trước như tên người, tên địa danh, tên tổ chức, ngày tháng, số, tiền tệ, ... Một ví dụ là:
"[PERSON Ông Nguyễn Hữu Minh] được đề cử chức tổng giám đốc của [ORG Công ty Đại Á] nhiệm kỳ [DTIME 2002-2006]."
Nhiều phương pháp học máy đã được áp dụng thành công cho bài toán nhận dạng này, trong đó các phương pháp chủ yếu dựa vào các đặc trưng ngôn ngữ và thông tin ngữ cảnh của từ để xác định lớp cho mỗi từ. Ví dụ, Tri Tran Q. [21] đã nghiên cứu sử dụng SVM để giải bài toán này và kết quả đạt được là khả quan. Hoặc như J.Gao cùng đồng tác giả [9] đã đề xuất một khung toán học thực hành để vừa thực hiện phân đoạn các từ đã biết cũng như phát hiện từ mới. Những nghiên cứu như vậy đã chỉ ra rằng bài toán nhận dạng thực thể có một mối liên hệ gần gũi với bài toán phân đoạn từ.

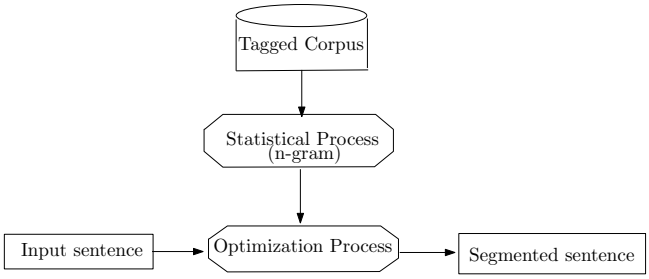
2.1.3 Mô hình N-gram

Mô hình ngôn ngữ N-gram thể hiện khá tốt mối quan hệ ngữ cảnh của từ. Trong mô hình đó, mỗi từ được coi như phụ thuộc xác suất vào n-1 từ trước nó.

$$P(W) = P(w_1 w_2 ... w_n) = \prod_{i=1}^{n} P(w_i | w_{i-n+1} ... w_{i-1}). \tag{2.1}$$

Mô hình N-gram được ứng dụng để phân đoạn từ trong đó với mỗi câu thì phân đoạn tốt nhất theo mô hình này là phân đoạn có xác suất P(W) được tính theo công thức 2.1 là lớn nhất. Trong đó, các xuất suất về sự phụ thuộc của một từ vào n từ trước đó được

thống kê dựa trên một corpus đủ lớn. Tùy vào giả thiết về tính phụ thuộc mà ta có các mô hình 2-gram hoặc 3-gram tương ứng. Phương pháp này là một trong những phương pháp thống kê chính để giải bài toán phân đoạn từ khi không có thông tin từ điển và dữ liệu gán nhãn. Mô hình phân đoạn từ sử dụng N-gram được biểu diễn như hình bên dưới (hình 2.1). Khi áp dụng phương pháp này đòi hỏi chúng ta phải xác định một độ đo tốt



Hình 2.1: Word segmentation using N-gram model.

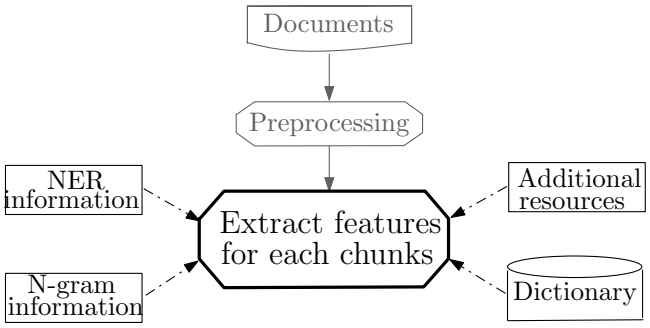
phù hợp với bài toán để đánh giá khả năng mỗi cụm hình vị có là một từ hay không? Có rất nhiều độ đo có thể sử dụng như: đơn giản chỉ sử dụng thông tin về tần suất xuất hiện của từ, hoặc có thể sử dụng thông tin mutual information hoặc t-score, ... Ví dụ, Maosong và các đồng tác giả [12] đã sử dụng độ đo mutual information và t-scores và một số kỹ thuật khác để xác định từ cho tiếng Trung và đã thu được kết quả khá cao (>90%). Đối với tiếng Việt tác giả Lê An Hà[10] đơn giản sử dụng tần suất N-gram để tối ưu xác suất của mỗi chunk. Kết quả thực nghiệm tuy không cao nhưng đã chứng tỏ rằng N-gram là một phương pháp phù hợp có thể ứng dụng cho bài toán phân đoạn từ tiếng Việt nói riêng.

2.2 Phân tích các mô hình

Hiệu quả của các phương pháp dựa trên từ điển như Longest Matching và Maximal Matching phụ thuộc phần lớn vào độ bao phủ của từ điển. Tuy nhiên, trên thực tế không tồn tại một từ điển hoàn thiện nào có khả năng bao phủ hết các mục từ của một ngôn ngữ bất kỳ bởi vì những từ mới luôn luôn xuất hiện. Theo thống kê, corpus SIGHAN's PK có xấp xỉ 30% OOVs [9]. Corpus tiếng Việt mà chúng tôi chọn để đánh giá mô hình cũng chứa 11.6% OOVs[3]. Đây là những tỷ lệ khá cao.

OOVs thường có hai loại: Một là các dạng tên thực thể hoặc dạng factoid; Hai là những từ mới không thuộc loại 1. Để nâng cao chất lượng phân đoạn từ thì các giải pháp cần kết hợp thông tin từ điển và các kỹ thuật để phát hiện từ mới. Xem xét dạng 1, chúng tôi thấy factoid có thể được nhận diện dễ dàng nhờ dùng biểu thức chính qui. Tuy nhiên, các NE không dễ nhận diện như vậy mà cần nghiên cứu các phương pháp phát hiện thực thể và đặc điểm riêng của các NE tiếng Việt. Kết quả của những nghiên cứu này sẽ được trình bày rõ hơn trong phần lựa chọn đặc trưng để phát hiện NE sử dụng mô hình MEM. Còn các từ mới thuộc loại 2 thường là những thuật ngữ chuyên ngành, từ nước ngoài được Việt hóa, ... Với những từ này thì không có qui tắc riêng nào để phát hiện mà cách thường được sử dụng nhất là thống kê tần suất từ. Nếu từ đó được dùng trên một ngưỡng nào đó thì ta có thể coi đó là một từ. Do vậy, chúng tôi sẽ sử dụng thông tin N-gram để đánh giá khả năng một cụm hình vị có phải là từ hay không?.

Từ những thông tin liên quan đó, chúng tôi trích đặc trưng cho mô hình Maximum Entropy Markov Model để huấn luyện bộ phân lớp. Cụ thể các mô hình gồm: mô hình dựa trên từ điển, mô hình nhận diện thực thể, mô hình N-gram và một số nguồn dữ liệu khác (xem biểu diễn ở hình dưới đây).



Hình 2.2: Các mô hình liên quan cần để trích các đặc trưng.

2.3 Thiết kế tập đặc trưng

Dựa trên các phân tích ở trên, chúng tôi đưa ra thiết kế chi tiết các đặc trưng chia ra làm 3 tập như sau:

Bảng 2.1: Table of feature sets

No	Model	Type of Features	Detailed Features
FS1	Tách dựa vào từ điển	Sự liên kết âm tiết SC	Mỗi SC có phải là entry của từ điển?
FS2	NER model	External Resource - Dictionary - Name List - Location List	Mỗi SC có phải là valid Name? In Location List? Is-Regular-Expression(0,0) Is-Initial-Capitalization(0,0) Is_All_Capitalization(0,0) Is_First_Observation(0,0) Is_Marks(0,0))
		Factoid	Is_Regex
FS3	N-gram Model	N-gram information	The log of probability (2-gram, 3-gram)

Bảng 2.2: Example of feature set 1

Syllable	Features set 1	
...	...	...
thoại	SC(-3,0)	In_dictionary: 0
	SC(-2,0)	In_dictionary: 0
	SC(-1,0)	In_dictionary: 1
	SC(0,0)	In_dictionary: 0
...	...	...

2.3.1 FS1: Đặc trưng trích từ mô hình tách từ dựa vào từ điển

Khác với các tiếp cận trước [3][5][8], thay vì sử dụng thông tin của các âm tiết trước và sau âm tiết hiện tại, chúng tôi chỉ sử dụng thông tin của các âm tiết đứng trước. Đây cũng là ý tưởng tạo từ ứng cử trong phương pháp Longest Matching. Xét ví dụ câu *"Thị trường điện thoại di động đang rất nóng"* , giả sử ta trích đặc trưng cho âm tiết hiện tại *"thoại"* thì các đặc trưng thuộc tập FS1 gồm có các đặc trưng được mô tả trong bảng 2.2.

2.3.2 FS2: Đặc trưng dựa vào mô hình nhận dạng tên thực thể

Như đã thảo luận ở phần trên, các dạng factoid sẽ được nhận biết nhờ sử dụng biểu thức chính qui. Do đó, trong tập đặc trưng này sẽ có một đặc trưng isRegex để nhận biết các dạng như ngày tháng, thời gian, tiền tệ, số, email, số điện thoại, fax và địa chỉ web. Để nhận dạng tên người ta sẽ dựa vào một danh sách tên tiếng Việt gồm khoảng 21.000 tên. Từ danh sách và đặc điểm tên tiếng Việt ta nhận thấy tên người hợp lệ thường tuân theo qui tắc:

Tên người hợp lệ = Họ + Tên đệm + Tên

Do vậy, dựa vào danh sách ta liệt kê 3 tập danh sách tương ứng gồm: danh sách chứa các họ, danh sách chứa các loại tên đệm và danh sách các tên riêng. Và để nhận biết tên riêng thì tương ứng với mỗi cụm liên kết hình vị trong phần FS1, ta sẽ có thêm một đặc trưng tương ứng để kiểm tra xem cụm đó có phải là một tên hợp lệ trong tiếng Việt hay không dựa vào qui tắc trên. Đặc trưng này cũng nhận giá trị:

- 1 nếu SC tuân theo luật
- 0 nếu ngược lại

Một dạng tên thực thể nữa được xét ở đây là tên địa danh hoặc tên của các công ty. Để phát hiện các thực thể thuộc loại này ta sẽ dựa vào một danh sách địa danh gồm khoảng 800 tên. Tương ứng với mỗi liên kết hình vị ta sẽ có một đặc trưng nhận giá trị:

- 1 nếu SC có trong danh sách địa danh
- 0 nếu ngược lại

Một điểm cần lưu ý là: Các tên thực thể được xét thường có ký tự đầu tiên của mỗi hình vị được viết hoa. Do đó, các hình vị ở đầu mỗi câu rất dễ bị nhầm với tên thực thể. Để tránh nhầm lẫn này ta cần thêm một đặc trưng nữa là Is_First_Observation(0,0) nhận giá trị 1 nếu hình vị này đứng đầu câu và 0 nếu ngược lại.

2.3.3 FS3: Đặc trưng dựa vào mô hình Ngram

Các mô hình phân đoạn từ dựa vào N-gram sử dụng xác suất của từng n-gram như một đơn vị thông tin cơ sở. Các xác suất này được tính dựa vào thống kê corpus lớn có độ bao phủ hình vị và độ bao phủ từ đủ tin cậy. Khi sử dụng N-gram để phân đoạn từ tác giả đã xây dựng corpus 10M hình vị, còn trong nghiên cứu này chúng tôi thu thập 14M corpus từ www.wikipedia.com . Chúng tôi thống kê xác suất mức 2-gram và 3-gram. Vì do corpus chưa lớn lắm thế nên một số cụm hình vị có tần suất xuất hiện nhỏ. Thế nên, khi sử dụng thông tin xác suất các n-gram này chúng tôi không sử dụng trực tiếp những xác suất đó mà sẽ ánh xạ chuyển chúng về đoạn [0,1] theo các công thức 2.2 và2.3.

$$mi = Log(P(N-gram)) = Log(f) - Log(14000000). \tag{2.2}$$

Bảng 2.3: Ví dụ một câu được tách từ

Thị	trường	chứng	khoán	đang	đi	xuống
B_W	I_W	B_W	I_W	B_W	B_W	B_W
The market		stock		being	go	down

$$Info(N-gram) = (1 - \frac{|mi + |max\_N - gram||}{|min\_N - gram|}). \quad (2.3)$$

Theo thống kê từ corpus thô(14M-syllable Wiki), ta có:

- P(2-gram) : min_2-gram ≈ -41, max_2-gram ≈ -8.00
- P(3-gram) : min_3-gram ≈ -41, max_3-gram ≈ -10.00

2.4 Kết quả thực nghiệm

Mô hình được sử dụng là mô hình maximum entropy [1] với giải thuật tối ưu BLMVM [2] có hỗ trợ giá trị là số thực. Khi sử dụng mô hình này, bài toán phân đoạn từ tiếng Việt được chuyển về bài toán phân lớp trong đó mỗi âm tiết sẽ được phân về một trong hai lớp là B_W (Begin of word) hoặc I_W (inner of word). Một ví dụ câu phân đoạn được cho trong bảng 2.3: Công cụ MEM được dùng trong các thực nghiệm được lấy từ http://www-tsujii.is.s.u-tokyo.ac.jp/ tsuruoka/maxent/. Về corpus, chúng tôi thực nghiệm trên corpus được công bố trong bài báo [3] tại địa chỉ http://www.jaist.ac.jp/ hieuxuan/vnwordseg/data. Corpus dùng để thống kê thông tin N-gram được lấy từ trang wikipedia.

2.4.1 Đánh giá các đặc trưng FS1 và FS2 so với các mô hình trước đó

Các nghiên cứu trước cũng thiết kế các đặc trưng dựa trên từ điển và mô hình NER, tuy nhiên các đặc trưng được thiết kế ở đây khác so với các đề xuất trong [8][9]. Kết quả thực nghiệm dưới đây sẽ so sánh và đánh giá tính phù hợp của cách chọn đặc trưng này. Đối với các xét thông tin dựa vào từ điển, chúng tôi đã tiến hành thực nghiệm và kết quả cho thấy cách tiếp cận của mô hình này cho kết quả cao hơn cách tiếp cận trước đó (xem bảng 2.4):

Với các đặc trưng dựa vào NER, thay vì kiểm tra từng âm tiết có trong danh sách tên đệm, tên họ, tên hay không (như tiếp cận trình bày trong [3]), chúng tôi sẽ kiểm tra từng

Bảng 2.4: Kết quả đánh giá hiệu quả của đặc trưng dựa vào từ điển

No	Features dựa trên	Precision	Recall	F1 measure
1	Left and right syllables	94.03	93.64	**93.84**
2	Only left syllables	94.95	94.2	**94.58**

Bảng 2.5: Result to estimate the importance of NER-based features

No	Features based on	Precision	Recall	F1 measure
1	Old	94.92	94.22	**94.57**
2	Our approach	95.15	94.43	**94.79**

liên kết âm tiết trong FS1 có phải là một tên hợp lệ? Kết quả thực nghiệm trong bảng 2.5 chứng minh đặc trưng này hiệu quả hơn hẳn. Lý do được giải thích là: do tiếng Việt có đặc điểm là tên họ, tên đệm và tên riêng có thể trùng nhau nên khi sử dụng kiểm tra riêng rẽ như [3] thì sẽ gây nhầm lẫn và dẫn tới dự đoán sai.

2.4.2 Đánh giá tầm quan trọng của từng tập thuộc tính

Luận văn cũng trình bày kết quả thực nghiệm đánh giá ảnh hưởng của từng tập đặc trưng tới hiệu quả phân đoạn cũng như chứng minh tính hiệu quả của mô hình cuối cùng. Để đánh giá chúng tôi sử dụng 3 độ đo là: Độ chính xác, độ hồi tưởng và độ đo F1 trên 5-fold test. Các kết quả được chỉ ra trong bảng 2.6 và3.1. Khi so sánh chúng tôi lấy phương pháp Longest Matching làm cơ sở. Đánh giá tác động của từng thuộc tính tới mô hình cuối cùng chúng tôi thiết kế hai loại thực nghiệm sử dụng các liên kết đặc trưng khác nhau cho mô hình MEM. Trong loại đầu tiên, ta sẽ lần lượt bỏ đi từng đặc trưng một với kết quả thực nghiệm cho trong bảng 6. Loại thực nghiệm thứ 2 ta sẽ thực nghiệm với từng tập đặc trưng một và kết quả được cho trong bảng 3.1.

Nhìn vào kết quả thực nghiệm ta dễ dàng nhận thấy rằng tập đặc trưng dựa trên từ điển có ảnh hưởng lớn nhất: Nếu chỉ sử dụng từ điển thì độ đo F1 là 94.58%, còn nếu từ điển đi thì kết quả là tệ nhất so với thực nghiệm bỏ đi mỗi đặc trưng khác (độ đo F1 là 87.5%). Điều này dễ giải thích bởi từ điển chứa một lượng từ khá ổn định và thông tin từ là chuẩn xác. Đặc trưng có tầm quan trọng thứ 2 là các đặc trưng giúp phát hiện Nes (ta có F1 là 93.55% nếu bỏ đặc trưng này đi và F1 là 91.32% nếu chỉ sử dụng NEs). Đặc trưng có ảnh hưởng ít nhất là đặc trưng của mô hình N-gram.

Khi các đặc trưng được kết hợp với nhau thì kết quả thu được là cao nhất: 95.30%

Bảng 2.6: Kết quả thực nghiệm khi bỏ đi lần lượt từng tập đặc trưng.

No	Not use (Reject)	Method	Precision	Recall	F1 measure
1	-	Longest Matching	81.07	87.97	**84.52**
2	Dict.-based feature set	MEM	96.99	77.1	**87.05**
3	NER-based feature set	MEM	97.21	89.88	**93.55**
4	N-gram-based feature set	MEM	95.15	94.43	**94.79**
5		MEM	96.71	93.89	**95.30**

Bảng 2.7: Kết quả thực nghiệm sử dụng từng loại đặc trưng riêng.

No	Only use	Method	Precision	Recall	F1 measure
1	-	Longest Matching	81.07	87.97	**84.52**
2	Dict.-based feature set	MEM	94.95	94.2	**94.58**
3	NER-based feature set	MEM	90.89	91.74	**91.32**
4	N-gram-based feature set	MEM	97.98	60.5	**79.24**
5	All	MEM	96.71	93.89	**95.30**

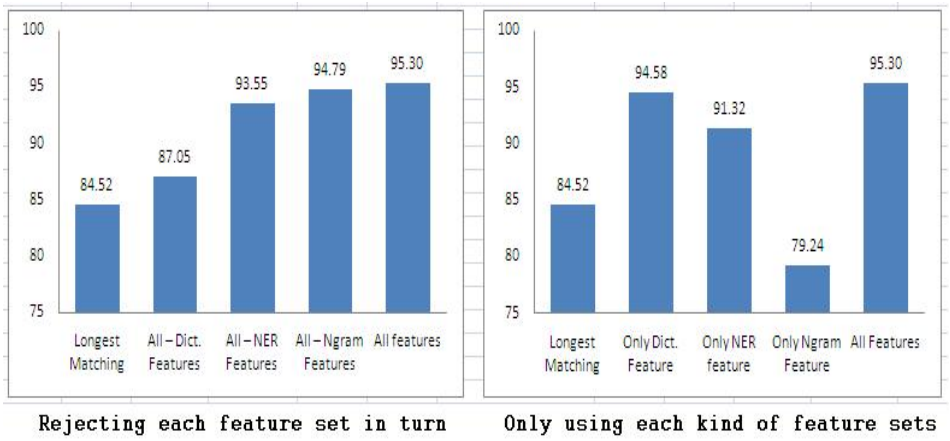
độ đo F1. Điều đó chứng tỏ rằng các tri thức về ngôn ngữ và ngữ cảnh của từ được cung cấp càng nhiều thì chất lượng phân đoạn của giải pháp đề xuất càng chính xác. Một biểu diễn trực quan của độ đo F1 cho hai loại thực nghiệm được trình bày trong hình 2.3.

2.5 Đánh giá kết quả tách từ

Luận văn trình bày một hướng tiếp cận mới cho bài toán phân đoạn từ tiếng Việt trong đó mô hình được chọn là mô hình Maximum Entropy Markov Model với giải thuật tối ưu BLMVM có hỗ trợ giá trị thực. Luận văn đã kết hợp rất nhiều đặc trưng hữu ích từ các mô hình khác gồm: mô hình phân đoạn từ dựa vào từ điển, mô hình nhận dạng tên thực thể và mô hình N-gram. Khác với các tiếp cận trước [3, 8], luận văn nghiên cứu cách trích chọn đặc trưng hữu ích hơn từ các mô hình dựa vào từ điển và mô hình nhận dạng tên thực thể. Một điểm nữa khác với các nghiên cứu trước đó là chúng tôi dùng thêm thông tin N-gram để nhằm phát hiện thêm các từ mới.

Kết quả thực nghiệm chỉ ra rằng mô hình sử dụng cả 3 loại tập đặc trưng nói trên đã làm tăng đáng kể chất lượng phân đoạn (95.30% độ đo F1). Thực nghiệm cũng đánh giá tầm quan trọng của từng loại thuộc tính đối với mô hình phân đoạn và kết quả cho thấy đặc trưng có ảnh hưởng lớn nhất là của mô hình dựa vào từ điển, tiếp đó là mô hình NE và mô hình N-gram có ảnh hưởng ít nhất.

Để đánh giá khả năng tách từ của mô hình so với các mô hình tốt nhất hiện nay, chúng tôi



Hình 2.3: Biểu đồ độ đo F1

tiến hành so sánh trên cùng corpus, kết quả tách từ đạt độ chính xác state of the art với 95.30% F1 (cao hơn tiếp cận trong [3]). Cũng với mô hình đó, chúng tôi cũng làm các thực nghiệm kiểm thử trên corpus của Trung tâm từ điển học Việt Nam www.vietlex.com.vn và đo độ đo F1 đạt 94.76% (>94.44% như báo cáo trong [8]).

Chương 3

Mô hình gán nhãn từ loại tiếng Việt

Trước khi xây dựng và kiểm thử mô hình gán nhãn từ loại, chúng tôi đã tiến hành xây dựng tập thẻ từ loại sau đó gán nhãn corpus từ loại tiếng Việt 8000 câu. Xuất phát từ thành công của mô hình Maximum Entropy Markov Model (MEM) đã được áp dụng cho tiếng Anh, tiếng Trung, ... luận văn cũng đề xuất xây dựng mô hình gán nhãn từ loại tiếng Việt dựa trên mô hình đó. Với mô hình lựa chọn này, luận văn tiến hành nghiên cứu và thử nghiệm các đặc trưng khác nhau nhằm tìm ra tập đặc trưng hữu ích đối với tiếng Việt.

3.1 Xây dựng corpus gán nhãn từ loại cho tiếng Việt

Xây dựng corpus là một công việc rất tốn thời gian và công sức. Trong luận văn này, chúng tôi đã cố gắng xây dựng một corpus tiếng Việt đủ dùng cho các ứng dụng về sau. Corpus này được xây dựng dựa vào corpus đã tách từ của nhóm tác giả Cẩm Tú được công bố trong [3]. Để xây dựng corpus, chúng tôi đã thực hiện các công việc sau:

- Thiết kế bộ VnPOS tag cho tiếng Việt gồm 14 nhãn từ và >10 nhãn ký hiệu (symbols).
- Xây dựng tool trợ giúp gán nhãn POS cho văn bản sau khi đã tách từ theo đúng định dạng và tài liệu đi kèm.
- Gán nhãn POS cho 8000 câu thuộc nhiều lĩnh vực khác nhau.

3.1.1 Thiết kế tập thẻ VnPOSTag

Chỉ xét riêng đối với tiếng Anh đã tồn tại rất nhiều tập thẻ POS khác nhau điển hình (theo [4]) là:

- Brown corpus: 87 nhãn
- Penn Treebank: 45 nhãn
- Lancaster UCREL C5: 61 nhãn

Chọn tập nhãn lớn sẽ làm tăng độ khó nhưng tập nhãn nhỏ hơn có thể không đủ đáp ứng cho một mục đích nhất định nào đó. Việc chọn tập nhãn nào sẽ tùy thuộc vào từng ứng dụng cụ thể, nói cách khác là tùy thuộc vào số lượng thông tin mà ứng dụng đó đòi hỏi. Do đó, cần phải có sự cân đối giữa:

- Có được lượng thông tin rõ ràng hơn (Tức là phạm vi phân lớp từ loại nhỏ hơn, chia thành nhiều từ loại hơn dựa trên nhiều yếu tố thể hiện sự khác biệt)
- Có khả năng tiến hành thực hiện việc gán nhãn (Tức là số lượng các từ loại càng ít càng dễ tiến hành)

Tức là cần phải có một sự thoả hiệp để đạt được một bộ chú thích từ loại không quá lớn và có chất lượng. Với tiếng Việt thì việc thiết kế tập thẻ POS càng khó khăn hơn bởi ngay trong tiếng Việt thì từ loại vẫn còn là vấn đề gây nhiều tranh cãi. Dựa theo một số tài liệu tham khảo về cú pháp tiếng Việt thì các mục từ trong tiếng Việt nhìn chung được phân chia thành các nhóm, mỗi nhóm lại được phân chia sâu hơn tùy loại [7]. Theo Diệp Quang Ban [7], việc tập hợp và quy loại các từ thường có ba tiêu chuẩn phân loại sau:

1. **Tiêu chuẩn 1:** *ý nghĩa khái quát. Các từ loại là những nhóm từ rất to lớn về khối lượng mà mỗi nhóm có một đặc trưng phân loại: tính vật thể, phẩm chất, hành động hoặc trạng thái,v.v... Ví dụ, những từ như: nhà, bàn, học sinh, con, quyển, sự v.v... được phân vào lớp danh từ, vì ý nghĩa từ vựng của chúng được khái quát hóa và trừu tượng hóa thành ý nghĩa thực thể - ý nghĩa phạm trù ngữ pháp của danh từ.*
2. **Tiêu chuẩn 2:** *khả năng kết hợp. Với ý nghĩa khái quát, các từ có thể có khả năng tham gia vào một kết hợp có nghĩa. Ở mỗi vị trí của kết hợp có thể xuất hiện những*

từ có khả năng lần lượt thay thế nhau, trong khi đó, ở các vị trí khác trong kết hợp, các từ còn lại tạo ra bối cảnh cho sự xuất hiện khả năng thay thế của những từ nói trên. Những từ cùng xuất hiện trong cùng một bối cảnh, có khả năng thay thế nhau ở cùng một vị trí, có tính chất thường xuyên, được tập hợp vào một lớp từ. Vận dụng vào tiếng Việt, những từ: nhà, bàn, cát, đá v.v... có thể xuất hiện và thay thế nhau trong kết hợp kiểu: nhà này, bàn này, cát này, đá này, v.v... và được xếp vào lớp danh từ. Chúng không thể xuất hiện và thay thế cho nhau trong kết hợp kiểu : hãy ăn, hãy mua, ăn xong, mua xong v.v ..., vốn là kiểu kết hợp của lớp động từ.

3. **Tiêu chuẩn 3:** *chức năng cú pháp. Tham gia vào cấu tạo câu, các từ có thể đứng ở một hay một số vị trí nhất định trong câu, hoặc có thể thay thế nhau ở vị trí đó, và cùng biểu thị một mối quan hệ về chức năng cú pháp với các thành phần khác trong cấu tạo câu, có thể phân vào một từ loại. Ví dụ, các từ: nhà, bàn, cát, đá ... có thể đứng ở nhiều vị trí trong câu. Chúng có thể thay thế nhau ở những vị trí đó, và có quan hệ về chức năng giống nhau với các thành phần khác trong câu ở mỗi vị trí, nhưng thường ở vị trí chủ ngữ trong quan hệ với vị ngữ (là hai chức năng cơ bản trong cấu tạo câu). Chức năng chủ ngữ là chức năng cú pháp chủ yếu để phân loại các từ nói trên vào lớp danh từ. Còn chức năng vị ngữ lại là chức năng cú pháp chủ yếu của các động từ (và tính từ), v.v ...*

Trong đó, tiêu chuẩn (2) và (3) làm trọng tâm trong sự phân định các tập thẻ. Việc xác định tập thẻ tùy thuộc vào từng loại ứng dụng xem cần thông tin cú pháp từ vựng tới mức nào mà có nhiều cách phân chia thô, mịn khác nhau. Để thuận tiện cho việc làm corpus và phục vụ một số ứng dụng nhất định trong nghiên cứu của nhóm như (Question Answering System, Text Summarization, ...), chúng tôi xác định tập thẻ vnTagSet ở mức thô gồm các thẻ được liệt kê như ở bảng 3.1 với ý nghĩa mỗi loại xin xem thêm ở phần phụ lục B.

3.1.2 Mô tả bộ dữ liệu làm vnPOS corpus

Bộ dữ liệu dùng để xây dựng corpus từ loại tiếng Việt chính là bộ dữ liệu được sử dụng trong phần tách từ đã trình bày trên. Kích cỡ của corpus cỡ 8000 câu được thu thập từ các báo điện tử của tiếng Việt thuộc nhiều chủ đề khác nhau như công nghệ thông tin, kinh tế, chính trị, xã hội, pháp luật, đời sống, ...

Bảng 3.1: Tập thẻ vnPOSTag của từ loại tiếng việt.

STT	Tên thẻ	Ý nghĩa của thẻ
1	NN	Danh từ thường
2	NC	Danh từ chỉ loại
3	NP	Danh từ riêng
4	VB	Động từ
5	JJ	Tính từ
6	PP	Đại từ
7	D	Định từ và số từ
8	AD	Phụ từ
9	IN	Giới từ
10	CC	Liên từ
11	UH	Thán từ
12	RB	Trợ từ
13	TN	Thành ngữ
14	X	Các từ không thể phân loại được
15++	Symbols	Các ký hiệu đặc biệt khác (, #, $, ...)

3.1.3 Xây dựng vnPOS corpus

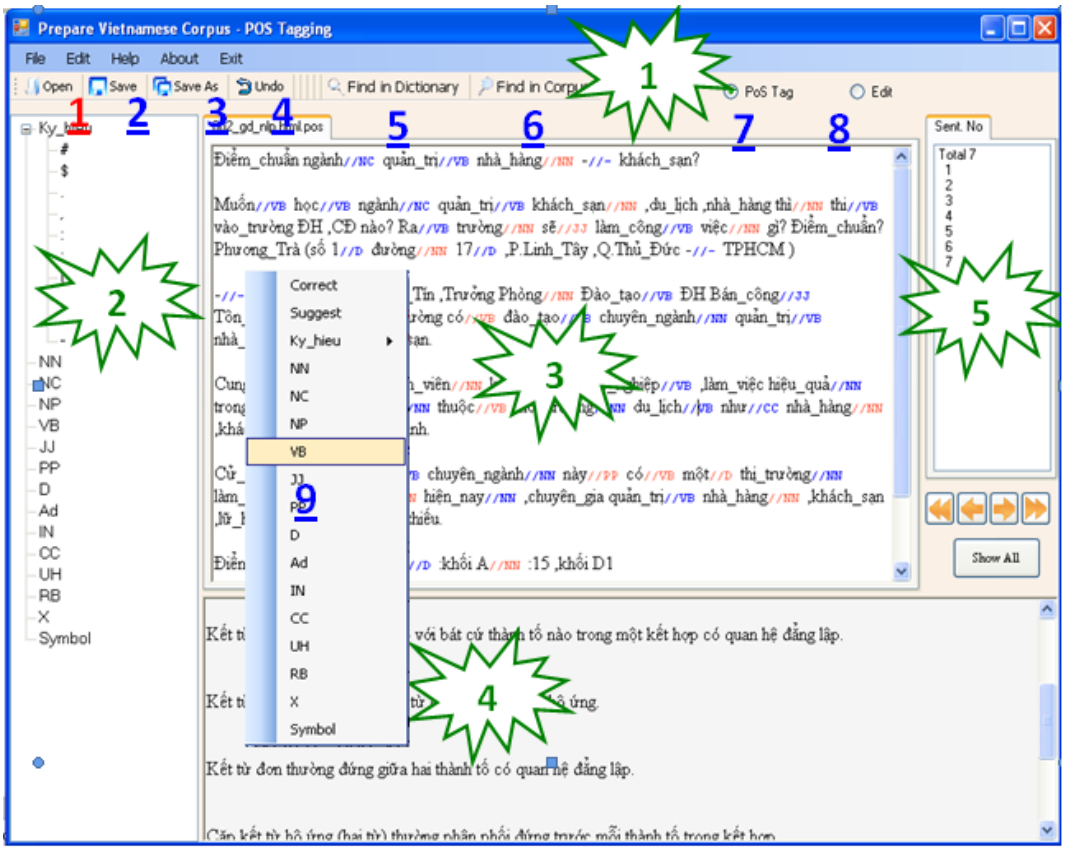
Xây dựng tool trợ giúp gán nhãn vnPOS

Để giúp cho việc gán nhãn chức năng cú pháp POS được thuận lợi và nhanh chóng, chúng tôi đã xây dựng một công cụ trợ giúp gán nhãn có giao diện như hình 3.1.

Khi file được tải vào RichTextBox ở phần trung tâm thì tool có đưa ra gợi ý về thẻ PoS của từng từ như sau:

- Các từ trong từ điển sẽ được gán nhãn dạng:
  - Màu xanh BLUE (ngành//NC) tức là trong từ điển có sẵn nó chỉ giữ một chức năng đó.
  - Màu đỏ RED (có//VB) tức là trong từ điển nó giữ nhiều hơn một chức năng cú pháp.
- Các từ còn lại thì để trống nhãn

Ngoài ra, tool còn có chức năng tìm kiếm thẻ của từng từ đã được gán nhãn trước đó để đưa gợi ý thêm trong quá trình làm dữ liệu.



Hình 3.1: Giao diện công cụ trợ giúp gán nhãn vnPOS.

Thực hiện gán nhãn vnPOS

Kết quả của quá trình này là một corpus đã gán nhãn POS tiếng Việt gồm xấp xỉ 8000 câu lấy từ các báo điện tử thuộc nhiều chủ đề khác nhau gồm khoa học công nghệ, kinh tế, chính trị, xã hội, mô tô xe máy, đời sống, pháp luật. Đây là một ví dụ câu được gán nhãn trong corpus.

Với//IN khoảng//D 8//D triệu//D thuê_bao//NN GSM//NP thì//IN thị_trường//NN trong//IN nước//NC là//RB rất//AD lớn//JJ với//IN những//D nhà//NC khai_thác//VB ,//, cung_cấp//VB dịch_vụ//NN GTGT//NN trên//IN điện_thoại_di_động//NN .//. Họ//PP đều//AD hoan_nghênh//VB sự//NC đổi_mới//VB cả//PP hai//D phương_diện//NN kinh_tế//NN và//CC chính_trị//NN .//.

3.2 Gán nhãn từ loại bằng phương pháp Maximum Entropy Markov Model

3.2.1 Mô hình xác suất

Theo [1] mô hình xác suất được định nghĩa trên không gian HxT, trong đó H là tập từ có thể và ngữ cảnh từ loại, hoặc còn gọi là "lịch sử", và T là tập các thẻ có thể có. Xác suất mô hình cuar lịch sử h cùng với thẻ t được định nghĩa theo công thức 3.1:

$$p(h,t) = \Pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)} \tag{3.1}$$

trong đó Π là hằng số chuẩn hóa, $\{\mu, \alpha_1, ..., \alpha_k\}$ là các tham số mang giá trị dương của mô hình và $\{f_1, ..., f_k\}$ chính là các đặc trưng "features", thỏa $f_j(h,t) \in \{0,1\}$. Chú ý rằng mỗi tham số α_j tương ứng với một đặc trưng f_j.

Cho trước một tập các từ $\{w_1, ..., w_n\}$ và một chuỗi thẻ $\{t_1, ..., t_n\}$ được xem là dữ liệu huấn luyện, ta định nghĩa h_i là lịch sử khi dự đoán thẻ t_i. Các tham số $\{\mu, \alpha_1, ..., \alpha_k\}$ được chọn sao cho làm cực đại likelihood dữ liệu huấn luyện sử dụng p theo công thức 3.2:

$$L(p) = \prod_{i=1}^{n} p(h_i, t_i) = \prod_{i=1}^{n} \Pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_i)} \tag{3.2}$$

Mô hình này được xem xét dưới dạng Maximum Entropy, trong đó mục tiêu là cực đại entropy của một phân phối dưới những ràng buộc nhất định. Ở đây, entropy của phân phối p được định nghĩa theo công thức 3.3

$$H(p) = -\sum_{h \in H, t \in \tau} p(h,t) \log p(h,t) \tag{3.3}$$

và các ràng buộc được cho bởi công thức 3.4

$$Ef_j = \widetilde{E}f_j, 1 \leq j \leq k \tag{3.4}$$

trong đó kỳ vọng đặc trưng của mô hình là 3.5

$$Ef_j = \sum_{h \in H, t \in \tau} p(h,t) f_j(h,t) \tag{3.5}$$

và kỳ vọng đặc trưng quan sát là 3.6

$$\widetilde{E}f_j = \sum_{i=1}^{n} \widetilde{p}(h_i,t_i) f_j(h_i,t_i) \tag{3.6}$$

trong đó $\widetilde{p}(h_i,t_i)$ là xác suất của (h_i,t_i) trong dữ liệu huấn luyện. Vì thế, các ràng buộc này sẽ ép buộc mô hình phải phù hợp (match) các kỳ vọng đặc trưng đó với kỳ vọng đặc trưng quan sát trong dữ liệu huấn luyện.

3.2.2 Các đặc trưng của POS tagging

Xác suất đồng thời của lịch sử h là thẻ t được xác định bằng các tham số mà các đặc trưng tương ứng của nó là hữu ích, ví dụ, α_j thỏa mãn $f_j(h,t) = 1$. Khi cho trước (h,t), một đặc trưng phải tồn tại trên bất cứ word hoặc tag trong lịch sử h, và phải chứa thông tin giúp dự đoán thẻ t, ví dụ như thông tin chính tả của từ hiện tại, hoặc thông tin về hai thẻ trước từ hiện tại. Ngữ cảnh word và tag xác định đối với một feature được cho bằng định nghĩa của lịch sử h_i như công thức 3.7:

$$h_i = \{w_i, w_{i+1}, w_{i+2}, w_{i-1}, w_{i-2}, t_{i-1}, t_{i-2}\} \tag{3.7}$$

Ví dụ,

$$f_j(h_i, t_i) = \begin{cases} 1, & \text{if suffix}(w_i) = "ing"; \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

Nếu đặc trưng trên tồn tại trong tập đặc trưng của mô hình, thì tham số mô hình tương ứng sẽ đóng góp vào xác suất đồng thời $p(h_i, t_i)$ khi w_i kết thúc có đuôi "ing" và t_i = VBG. Do vậy, một tham số mô hình α_j xem như một trọng số hiệu quả cho một bộ dự đoán ngữ cảnh nhất định, trong trường hợp suffix "ing", hướng tới xác suất quan sát một thẻ nhất định, trong trường hợp VBG. Mô hình sẽ tạo ra một không gian đặc trưng bằng cách quét mỗi cặp (h_i, t_i) trong dữ liệu huấn luyện với "templates" được cho sẵn (Xin xem chi tiết thêm trong tài liệu [1]).

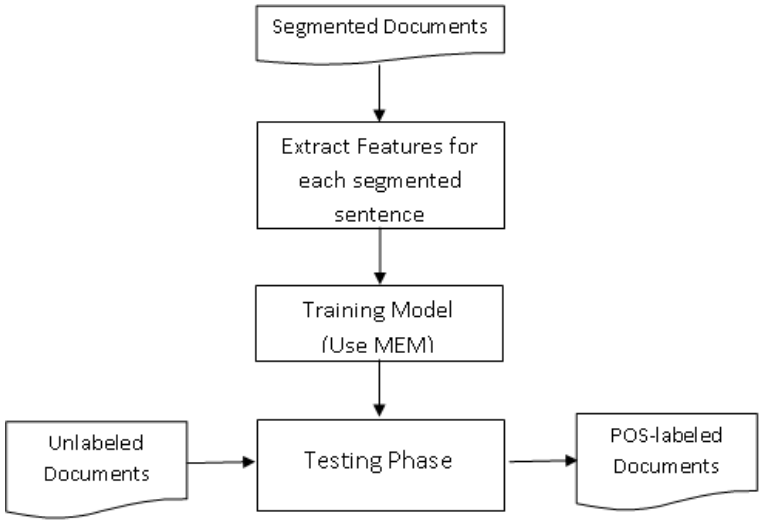
3.3 Đề xuất mô hình gán nhãn từ loại cho tiếng Việt

Trong phần trên chúng tôi đã đề xuất mô hình xây dựng bộ tách từ cho tiếng Việt, tiếp theo chúng tôi sẽ nghiên cứu về kiến trúc xử lý và xây dựng bộ gán nhãn từ loại cho tiếng Việt. Như đã trình bày trong chương 1, chúng ta có rất nhiều cách tiếp cận khác nhau cho bài toán gán nhãn từ loại. Tuy nhiên, chúng tôi nhận thấy rằng các phương pháp học máy cho kết quả tốt hơn cả. Do vậy, để thực hiện gán nhãn POS, chúng tôi sử dụng phương pháp học máy MEM [2] [1] đã được sử dụng thành công cho tách từ tiếng Anh và một số thứ tiếng khác (xem trình bày trong phần 3.2 ở trên). Khi đó, bài toán POS được xem là bài toán phân lớp với các lớp chính là các nhãn từ loại mô tả ở bảng 3.1.
Trong phần này, chúng tôi quan tâm tới kiến trúc theo kiểu pipeline, nghĩa là việc gán nhãn từ loại được thực hiện sau khi đã có thông tin về từ vựng. Kiến trúc tổng thể gán nhãn POS được thể hiện trong hình 3.2:

Trong đó, có hai pha chính là pha huấn luyện mô hình và pha giải mã.

- **Pha huấn luyện mô hình:** Đầu vào là văn bản đã được tách từ đưa qua bộ trích chọn đặc trưng (các đặc trưng hữu ích cho tiếng Việt sẽ được trình bày cụ thể trong các phần sau) rồi đưa vào mô hình MEM để huấn luyện.

- **Pha giải mã:** văn bản đầu vào sẽ được qua pha giải mã theo thuật toán beam search trình bày dưới đây, kết quả sẽ cho ra chuỗi thẻ tốt nhất ứng với mỗi câu đầu vào (chuỗi thẻ phải thuộc vào tập thẻ được chọn).



Hình 3.2: Kiến trúc gán nhãn POS.

Phần tiếp theo sẽ trình bày các cách trích chọn đặc trưng hữu ích cho bài toán này.

3.3.1 Gán nhãn từ loại dựa vào thông tin từ

Để tìm các đặc trưng hữu ích cho tiếng Việt, trước hết chúng tôi đã nghiên cứu cách trích chọn đặc trưng của một số mô hình gán nhãn cho tiếng Anh [1]. Phương pháp của Ratnaparkhi giả thiết rằng một câu đã được tách từ trước và gán nhãn POS dựa trên mức từ sử dụng các đặc trưng ngữ cảnh xung quanh từ đang xét.

Các đặc trưng - Features.

Các mẫu đặc trưng được mô tả như ở dưới đây, trong đó W đề cập tới từ còn POS đề cập tới nhãn từ loại của từ.

- Từ W_i (i= -2,-1,0,1,2)
- Sự liên kết từ với từ hiện tại với window size = 2
- Thẻ của từ đằng trước POS(W_{-1})
- Thẻ của 2 từ đằng trước từ hiện tại POS(W_{-2})POS(W_{-1})
- Từ đang xét có phải dấu câu?
- Từ đang xét có phải First Observation?

- Từ đang xét có Captitalize?

Giải mã - Testing

Kho ngữ liệu kiểm thử được gán nhãn theo từng câu một, thủ tục đòi hỏi thuật toán tìm kiếm để liệt kê các chuỗi nhãn ứng cử viên cho câu và chuỗi nhãn với xác suất cao nhất được chọn là đáp án. Thuật toán tìm kiếm giải mã được trình bày tiếp sau: Thủ tục kiểm thử tương tự với thuật toán mà Ratnaparkhi đã mô tả đó là sử dụng thuật toán beam seach. Sau khi đã huấn luyện mô hình entropy cực đại, ta có thể sử dụng nó để gán nhãn từ loại cho một câu mới. Quá trình gán nhãn cho câu mới tiến hành cho các từ từ trái sang phải. Tại mỗi thời điểm sẽ lưu lại k chuỗi nhãn tốt nhất (xác suất lớn nhất) và sử dụng nó để làm ngữ cảnh gán nhãn cho từ tiếp theo. Cho trước một câu $w_1, ..., w_n$, một chuỗi nhãn ứng cử viên có xác suất điều kiện như phương trình 3.9

$$P(a_1...a_n|w_1...w_n) = \prod_{i=1}^{n} p(a_i|b_i) \tag{3.9}$$

trong đó b_i là lịch sử tương ứng với từ thứ i. Thay vì phải tính tích các thừa số nhỏ ta logarit hai vế của phương trình trên và đưa về phép lấy tổng.
Thuật toán BEAM SEARCH: beamsize = N

- Sinh các tag cho từ w_i, tìm ra N tag có xác suất cao nhất gắn vào N chuỗi tag kí hiệu là S_j (j = 1, ..., N)
- For i = 2 to n (n là độ dài của câu)
  - For j = 1 to N
    - Sinh các tag cho w_i với S_j là chuỗi tag trước đó.
    - Gắn tag này vào đuôi của S_j
  - Từ các chuỗi tag đang có tìm N chuỗi có xác suất cao nhất là S_j (j = 1, ..., N)
- Trả về chuỗi tag có xác suất cao nhất S_1

Trong thực nghiệm, chúng tôi chọn N = 3.

Bảng 3.2: Kết quả gán nhãn POS dựa vào thông tin mức từ

Fold	**Precision**
1	85.17
2	85.64
3	85.51
4	85.71
5	85.81
Averg.	**85.57**

Kết quả thực nghiệm

Thực nghiệm được tiến hành trên corpus có kích thước 8000 câu như mô tả ở phần 2. Toàn bộ corpus được chia làm 5 fold sau đó kiểm thử theo phương pháp cross validation 5-fold test. Kết quả thực nghiệm được mô tả ở bảng 3.2:

Kết quả thực nghiệm cho thấy độ chính xác trung bình đạt được chỉ là 85.57%, thấp hơn nhiều so với kết quả 96% mà Rat sử dụng khi gán nhãn cho English. Các đặc trưng tỏ ra hữu ích với bộ POS tiếng Anh thì dường như không ứng dụng được cho tiếng Việt trong MEM. Sự khác biệt ngôn ngữ giữa tiếng Anh và tiếng Việt khiến cho việc áp dụng phương pháp tiếng Anh cho tiếng Việt trở nên không hiệu quả.

3.3.2 Gán nhãn từ loại dựa vào âm tiết

Vì bộ gán nhãn POS dựa trên thông tin về từ cho kết quả không được cao, còn cách xa độ chính xác mà Ratnaparkhi thực hiện POS cho tiếng Anh, chúng tôi tiếp tục nghiên cứu một cách biểu diễn đặc trưng khác đã được xây dựng thành công cho tiếng Trung [14]. Với đặc điểm tiếng Việt rằng âm tiết cấu tạo nên từ, chúng tôi thiết kế tập đặc trưng mới dựa trên thông tin về âm tiết như sau:

Đặc trưng - Features

Chúng tôi đã tiến hành thử nghiệm với loại đặc trưng dựa trên âm tiết như mô tả trong phần dưới đây:

- Âm tiết S_i (i= -2,-1,0,1,2)
- Sự liên kết âm tiết với âm tiết hiện tại với window size = 2
- Thẻ của âm tiết đằng trước POS(S_{-1W_0})

- Thẻ của 2 âm tiết đằng trước từ hiện tại POS(S_{-2W_0})POS(S_{-1W_0})
- Âm tiết đang xét có phải dấu câu?
- Âm tiết đang xét có phải First Observation?
- Âm tiết đang xét có Captitalize?

Trong đó với chú ý thêm là đặc trưng POS(S_{-1W_0}) chính là nhãn POS của âm tiết trước của từ ngay trước từ hiện tại. Và POS(S_{-2W_0})POS(S_{-1W_0}) chính là nhãn POS của âm tiết trước của từ ngay trước và từ ngay trước nữa của từ hiện tại. Giả sử xét ví dụ câu đầu vào là: *Từ lâu ông được biết đến là nhà quản_ lý tài_ ba .*
Giả sử xét âm tiết *ba* thì 2 đặc trưng tương ứng cho âm tiết này sẽ nhận giá trị là: VB và NC_VB (Với giả thiết rằng "nhà" được gán nhãn là NC và "quản_lý" được gán nhãn là VB).

Giải mã - Testing

Thủ tục giải mã tương tự như đã mô tả ở phần trên, chú ý một điểm là xác suất của một từ được gán nhãn POS t được tính bằng tích xác suất của mỗi âm tiết trong từ đó được gán nhãn POS t tương ứng. Giả sử khi đánh giá xác suất của từ *"tài_ ba"* được gán nhãn JJ thì đầu tiên ta tính xác suất của mỗi âm tiết *"tài"* và *"ba"* được gán nhãn JJ sau đó nhân tích hai xác suất này ta được xác suất của từ *"tài_ ba"* được gán nhãn JJ. Đây cũng chính là ràng buộc để tất cả các âm tiết trong một từ của một câu đã được tách từ phải có cùng một nhãn POS.

Kết quả thực nghiệm

Chúng tôi cũng tiến hành thực nghiệm dựa trên corpus tương tự như đã mô tả trong phần 3.1 và thực nghiệm cho kết quả như bảng 3.3: Độ chính xác trung bình trên 5 fold lên khá cao 89.22%.

3.4 Đánh giá kết quả gán nhãn từ loại

Từ kết quả thực nghiệm ta nhận thấy rằng đặc trưng tỏ ra hữu ích với tiếng Anh thì lại không thực sự hữu ích đối với tiếng Việt bởi sự khác biệt về ngôn ngữ (tiếng Anh là ngôn

Bảng 3.3: Kết quả gán nhãn POS dựa vào thông tin âm tiết

Fold	**Precision**
1	88.63
2	89.64
3	89.26
4	89.36
5	89.63
Averg.	**89.22**

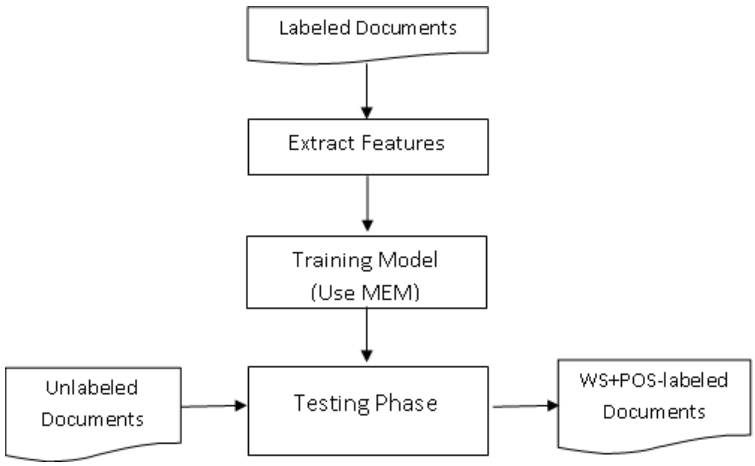
ngữ biến hình trong khi đó tiếng Việt là ngôn ngữ đơn lập, không biến hình). Kết quả thực nghiệm cũng chỉ ra rằng cách trích các đặc trưng dựa trên âm tiết cho kết quả cao hơn đáng kể (89.22%) so với cách trích các đặc trưng dựa trên thông tin về từ (85.57%). Như vậy, đối với tiếng Việt thì hướng tiếp cận dựa trên âm tiết tỏ ra hữu ích hơn hẳn hướng tiếp cận dựa trên từ.

Chương 4

Mô hình tích hợp tách từ và gán nhãn từ loại tiếng Việt

4.1 Lựa chọn mô hình tích hợp cho tiếng Việt.

Tới đây, chúng tôi đã xây dựng thành công bộ tách từ và gán nhãn POS cho tiếng Việt. Trong đó, bộ tách từ đạt state-of-the-art và công trình đã được công bố trong [17]. Với bộ POS chúng tôi đã tìm các đặc trưng hữu ích cho tiếng Việt và kết quả đạt được là rất khả quan. Từ sự khích lệ đó cộng với sự thành công của cách tiếp cận gán nhãn POS dựa vào âm tiết, chúng tôi lựa chọn phương pháp tích hợp giống như của [14] (đã trình bày khái quát phương pháp trong chương 1, phần 1.3).



Hình 4.1: Kiến trúc tích hợp tách từ và gán nhãn từ loại tiếng Việt.

Kiến trúc tích hợp được mô tả trong hình 4.1. Với một văn bản đầu vào, qua bước tiền xử lý để tách câu thì đầu vào đối với hệ thống là câu. Với mỗi câu đầu vào, chúng tôi sẽ cho qua bộ phân lớp tích hợp, và đầu ra sẽ gán cho mỗi âm tiết trong câu một thẻ bao gồm hai thông tin: Thông tin về từ (word boundary) và thông tin về thẻ từ loại (POS). Số lớp của mô hình sẽ bằng tích của số lớp thông tin từ nhân với số lớp thông tin về thẻ từ loại (các lớp này giống như đã trình bày ở phần 2 và phần 3 ở trên). Ví dụ đầu ra cho câu *"Công ty đang mở chiến dịch quảng cáo ."* như trong bảng 4.1:

Bảng 4.1: Một ví dụ ouput của mô hình tích hợp.

Công	ty	đang	mở	chiến	dịch	quảng	cáo	.
B_NN	I_NN	B_AD	B_VB	B_NN	I_NN	B_VB	I_VB	B_.

4.2 Xây dựng mô hình và tiến hành thực nghiệm

4.2.1 Features

Các đặc trưng được tổng hợp từ các đặc trưng của mô hình tách từ và các đặc trưng của mô hình gán nhãn từ loại. Trong đó, đặc trưng của mô hình gán nhãn từ loại sẽ lấy dựa vào hướng tiếp cận dựa trên âm tiết. Chú ý rằng khi đó đặc trưng về thông tin thẻ POS của âm tiết được thay bằng:

- $B(S_{-1W_0})POS(S_{-1W_0})$
- $B(S_{-2W_0})POS(S_{-2W_0})B(S_{-1W_0})POS(S_{-1W_0})$

B là thông tin về từ hoặc là B(Begin_Of_Word) hoặc là I(Inner_Of_Word), còn POS là thông tin về từ loại của âm tiết đang xét đó. Như vậy, so với hướng tiếp cận gán nhãn từ loại theo kiểu pipeline thì thông tin thẻ không chỉ gồm thông tin từ loại (POS) mà còn bao gồm cả thông tin về từ (word boundary).

4.2.2 Giải mã

Trong giải mã, chúng tôi cũng sử dụng giải mã bằng thuật toán BEAM SEARCH như đã trình bày ở trên với N = 3. Trong đó chú ý là khi chọn tập thẻ tốt nhất cho âm tiết hiện tại thì chỉ xét các thẻ hợp lệ tức là thẻ thỏa mãn rằng các âm tiết trong cùng một từ thì phải có cùng thẻ từ loại.

4.2.3 Kết quả

Kết quả thực nghiệm 5-fold test trên corpus xây dựng được trình bày trong bảng 4.2: Nhìn vào bảng kết quả thực nghiệm, chúng ta nhận thấy rằng hướng tiếp cận tích hợp

Bảng 4.2: Kết quả thực nghiệm tích hợp WS và POS tagging.

Fold	**Word Segmentation**			**POS Tagging**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
1	91.75	94.41	93.06	84.97	87.45	86.2
2	92.1	94.53	93.32	84.3	88.12	86.21
3	91.76	95	93.38	84.65	89.01	86.83
4	92.53	95.11	93.82	83.71	88.15	85.93
5	91.87	95.2	93.54	84.76	88.92	86.84
Average	**92**	**94.85**	**93.42**	**84.48**	**88.33**	**86.40**

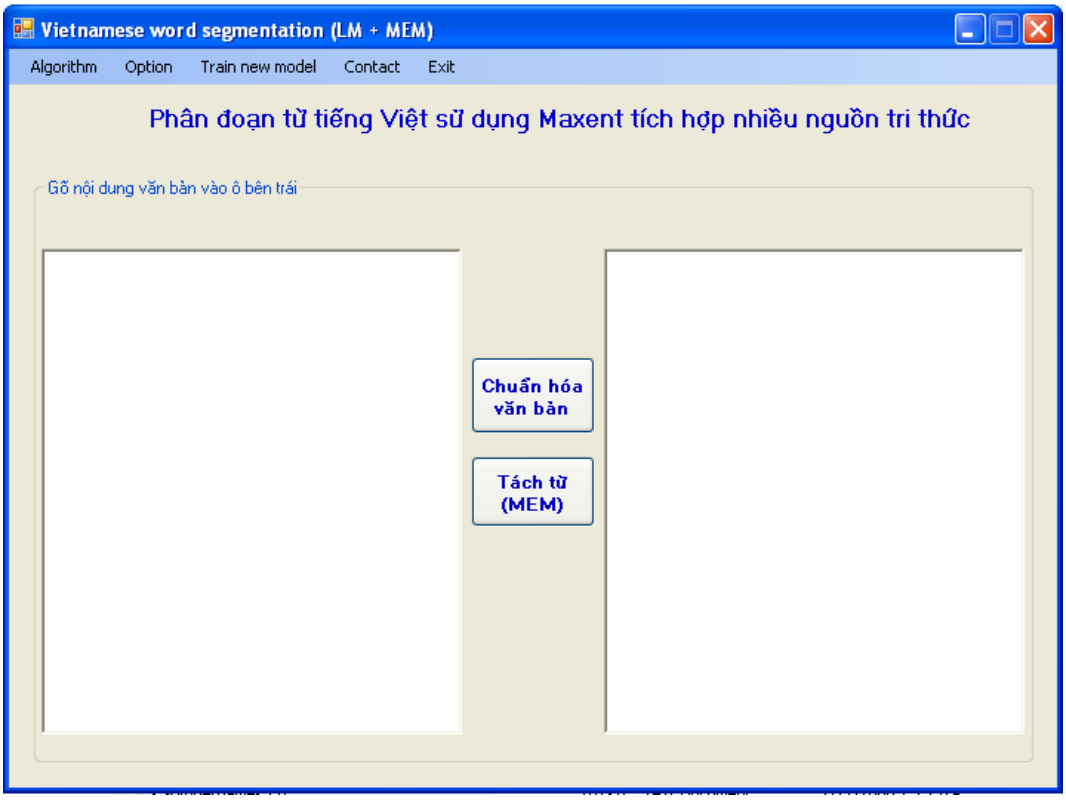
cho kết quả đều thấp hơn so với hướng tiếp cận pipeline trên cả hai bài toán đó.

4.3 Thảo luận

Kết quả thực nghiệm cho thấy tiếp cận tích hợp áp dụng cho tiếng Việt không làm tăng hiệu quả của hai bộ riêng rẽ (khác so với các nghiên cứu của tiếng Trung khi tích hợp thì cho kết quả nhìn chung là cao hơn đối với cả hai bài toán). Lý do có thể là do đặc điểm về sự khác biệt ngôn ngữ hoặc có thể do trong corpus này không hỗ trợ nhiều trường hợp để ứng dụng được phương pháp tích hợp đó.

Kết luận

Luận văn đã quan tâm tới hai bài toán nền tảng đóng vai trò quan trọng trong xử lý ngôn ngữ nói chung và xử lý tiếng Việt nói riêng đó là bài toán tách từ và bài toán gán nhãn từ loại tiếng Việt và trình bày một mô hình tích hợp hai bài toán trên áp dụng cho tiếng Việt. Đối với bài toán tách từ, luận văn đã tiến hành xây dựng mô hình state-of-the-art và đạt được kết quả là sản phẩm 1 bài báo khoa học và công cụ thực thi tách từ đi kèm. Giao diện của công cụ tách từ được mô tả trong hình 4.2



Hình 4.2: Giao diện công cụ tách từ tiếng Việt

Đối với bài toán gán nhãn từ loại, đóng góp của luận văn là xây dựng một corpus gán từ loại tiếng Việt khá lớn (8000 câu), dựa trên corpus đó chúng tôi nghiên cứu các cách

trích chọn đặc trưng và đề xuất các đặc trưng hữu ích đối với tiếng Việt. Từ đó, chúng tôi đã xây dựng một mô hình gán nhãn từ loại đạt kết quả khả quan (90%). Giao diện của mô hình được mô tả trong hình 4.3

Hình 4.3: Giao diện công cụ tách từ tiếng Việt

Khi xem xét vấn đề tích hợp, từ thực nghiệm chúng tôi kết luận rằng hướng tiếp cận hai bài toán này theo kiểu pipeline cho kết quả tốt hơn so với hướng tích hợp hai bài toán lại.

Tài liệu tham khảo

[1] Ratnaparkhi A. A simple introduction to maximum entropy models for natural language processing. In *Technical Report 97-08*. Institute for Research in Cognitive Science, University of Pennsylvania, 1997.

[2] Steven J. Benson and Jorge J. More. A limited-memory variable-metric method for bound-constrained minimization. In *Preprint ANL/MCS*, pages 909–0901, 2001.

[3] Xuan-Hieu Phan Le-Minh Nguyen Cam-Tu Nguyen, Trung-Kien Nguyen and Quang-Thuy Ha. Vietnamese word segmentation with crfs and svms: An investigation. In *Proceeding of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC20)*, pages 215–222. Wuhan, China, 2005.

[4] James H.Martin Daniel Jurafsky. *Speech and Language Processing*. Prentice Hall, Englewood Cliffs, New Jersey 07632, 1999.

[5] H.Kiem D.Dien and N.V.Toan. Vietnamese word segmentation. In *Proceedings of NLPRS'01 (The 6th Natural Language Processing Pacific Rim Symposium)*, pages 749–756. Tokyo, Japan, 2001.

[6] Kiem Hoang Dien Dinh. Pos-tagger for english-vietnamese bilingual corpus. In *Workshop On Building And Using Parallel Texts: Data Driven Machine Translation And Beyond*, 2003.

[7] Hoang Dan Diep Quang Ban. *Ngu phap tieng Viet*. NXB Giao Duc, Ha Noi, 2000.

[8] Vu Thuy Dinh Dien. A maximum entropy approach for vietnamese word segmentation. In *In Proceedings of 4th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future*, pages 12–16. HoChiMinh City, Vietnam, 2006.

[9] Li M. Wu A. Huang-C.N Gao, J.F. Chinese word segmentation and named entity recognition: A pragmatic approach. In *Computational Linguistics*. MIT Press, 2005.

[10] Le An Ha. A method for word segmentation in vietnamese. In *Proceedings of Corpus Linguistics*. Lancaster, UK, 2003.

[11] Hwee Tou Ng Jin Kiat Low and Wenyuan Guo. A maximum entropy approach to chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 161–164, 2005.

[12] Dayang Shen Maosong Sun and Benjamin K. Tsou. Chinese word segmentation without using lexicon and hand-crafted training data. In *In Proceeding. of COLING-ACL*, pages 1265–1271, 1998.

[13] Charenpornsawat P. Mekanavin, S. and B. Kijsirikul. Feature-based thai words segmentation. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 41–48. Phuket, Thailand, 1997.

[14] Hwee Tou Ng and Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *In Proceedings of EMNLP*, 2004.

[15] Cao Hoang Tru Nguyen Quang Chau, Phan Thi Tuoi. Gan nhan tu loai cho tieng viet dua tren van phong va tinh toan xac suat. In *Tap chi phat trien KHCN tap 9*, page 11, So 2, nam 2006.

[16] Le Hong Phuong Nguyen Thi Minh Huyen, Vu Xuan Luong. Su dung bo gan nhan tu loai xac suat qtag cho van ban tieng viet. In *Ky yeu hoi thao ICT.rda'03*. Ha Noi, 2003.

[17] Thuy Ha Oanh Tran, Cuong Le. Improving vietnamese word segmentation by using multiple knowledge resourses. In *Proceeding of workshop on EMALP, PRICAI08*, pages 1–12. Hanoi, Vietnam, 2008.

[18] C. Chan P. Wong. Chinese word segmentation based on maximum matching and word binding force. In *Proceedings of Coling 96*, pages 200–203, 1996.

[19] Yanxin Shi and Mengqiu Wang. A dual-layer crf based joint decoding method for cascade segmentation and labelling tasks. In *In Proceedings of the IJCAI Conference, Hyderabad, India*, 2007.

[20] Tsou B.K Sun M.S., Xu D.L. Integrated word segmentation and part-of-speech tagging based on the divide and conquer strategy. In *In Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering*, 2003.

[21] Hung Ngo Q.-Dien Dinh Tri Tran Q., Thao Pham T. X.. and Nigel Collier. Named entity recognition in vietnamese documents. In *Progress in Informatics*, pages No.4, pp. 5–13, 2007.

[22] Stephen Clark Yue Zhang. Joint word segmentation and pos tagging using a single perceptron. In *In proceedings of ACL*, 2008.

Phụ lục A

Một số thuật ngữ tiếng Anh tương ứng

Bảng A.1: Bảng thuật ngữ Anh - Việt

STT	**Thuật ngữ tiếng Anh**	**Viết tắt**	**Thuật ngữ tiếng Việt**
1	Natural Language Processing	NLP	Xử lý ngôn ngữ tự nhiên
2	Word Segmentation	WS	Tách từ
3	Part of speech tagging	POS tagging	Gán nhãn từ loại
4	Maximum Entropy Markov Model	MEM	Mô hình Markov cực đại entropy
5	Named Entity Recognition	NER	Nhận dạng thực thể
6	Hidden Markov Model	HMM	Mô hình Markov ẩn
7	Maximal Matching	MM	Phù hợp tốt nhất
8	Longest Matching	LM	Phù hợp dài nhất
9			

Phụ lục B

Chú giải tập từ loại vnPOS

AD - Phó từ (Phụ từ)
Khái niệm: Phó từ là hư từ thường dùng kèm với thực từ (động từ, tính từ). Chúng biểu thị ý nghĩa về quan hệ giữa quá trình và đặc trưng với thực tại, đồng thời cũng biểu hiện ý nghĩa về cách thức nhận thức và phản ánh các quá trình và đặc trưng trong hiện thực. Phó từ không có khả năng làm trung tâm ngữ nghĩa – ngữ pháp trong kết hợp thực từ, và rất ít có khả năng làm thành phần chính trong câu. Phó từ thường xuất hiện phổ biến ở vị trí thành tố phụ trong kết hợp thực từ, và trong cấu tạo thành phần câu.
Phân loại : Phó từ bao gồm các loại sau:

- Phó từ thời gian: đã, từng, mới, sẽ, sắp ...
- Phó từ so sánh và tiếp diễn: cũng, đều, vẫn, cứ, còn, nữa, cùng.
- Phó từ trình độ: rất, lắm, quá, cực kỳ, hơi, khí, khá.
- Phó từ phủ định, khẳng định: không, chẳng, chưa, có.
- Phó từ sai khiến: hãy, đừng, chớ
- Phó từ chỉ kết quả: mất, được, ra, đi.
- Phó từ chỉ tần số: thường, năng, ít, hiếm, luôn luôn, thường thường
- Phó từ tác động: cho
- Phó từ chỉ ý nghĩa tình thái chủ quan hoặc khách quan: vụt, thốt, chợt, bỗng, bỗng dưng, thình lình, đột nhiên, ... ắt, ắt là, hẳn là, chắc hẳn, ...

Ví dụ:
Chúng em [đã/Ad] rải bao nhiêu đá mà đường còn ra thế.
Nó [đang/Ad] di chuyển về phía chúng ta.
Con [mới/Ad] về.
Anh [vừa/Ad] trên đó xuống.
Tôi [lại/Ad] [sắp/Ad] đi xa một chuyến nữa đây.
Nhưng anh [lại/Ad] yêu tôi, tôi phải nói, vì tôi [cũng/Ad] yêu anh.
Mọi người [đều/Ad] nhảy, trừ chị Lộc.
Keng [vẫn/Ad] chạy suốt ngày.

CC - Liên từ (Kết từ đẳng lập)
Khái niệm: Chỉ ý nghĩa quan hệ đẳng lập, dùng để nối kết các từ, các kết hợp từ (ở bậc cụm từ hay ở bậc câu, đoạn văn). Kết từ đẳng lập không gắn bó với bát cứ thành tố nào trong một kết hợp có quan

hệ đẳng lập. Kết từ đẳng lập có thể là một từ đơn hay một cặp hô ứng. Kết từ đơn thường đứng giữa hai thành tố có quan hệ đẳng lập. Cặp kết từ hô ứng (hai từ) thường phân phối đứng trước mỗi thành tố trong kết hợp *Phân loại*:
Và, với, cùng, hay, hoặc, rời, là, rằng, hình như, còn, thì, cũng như, chứ càng... càng... *Ví dụ*
Kính mong ông [cùng/CC] các vị giáo viên phổ biến [và/CC] giải thích...
Vậy mà ba [với/CC] con tưởng má đến mai mới về.
Sáu năm [hay/CC] bao nhiêu năm thì Đông vẫn là bạn của anh.
Anh vít cần uống thêm [rồi/CC] tiếp.

textbfD - Số từ và định từ
Xét trong văn phạm tiếng Việt, vị trí của số từ và định từ trong luật sinh của văn phạm gần như nhau. Do đó ta có thể gộp số từ và định từ vào chung một loại tag. *Số từ*
Số từ gồm những từ biểu thị ý nghĩa số. Xét theo đối tượng phản ánh trong nhận thức và tư duy, ý nghĩa số vừa có tính chất thực, vừa có tính chất hư. Khả năng kết hợp với số từ phổ biến là được dùng kèm danh từ để biểu thị số lượng sự vật được nêu ở danh từ. Số từ có thể đảm nhiệm một số chức năng cú pháp (làm chủ ngữ, làm vị ngữ), nhưng bị hạn chế trong những điều kiện nhất định của kết cấu câu trong văn bản.

Bao gồm: *Một, hai, ... Vài, dăm ba, ... Định từ* Là những từ biểu thị quan hệ về số lượng với sự vật được nêu ở danh từ, chuyên dùng kèm với danh từ, với chức năng làm thành tố phụ trong kết hợp từ có trung tâm ngữ nghĩa – ngữ pháp là danh từ. Số lượng định từ tuy không nhiều, nhưng chúng có tác dụng dạng thức hóa một số ý nghĩa ngữ pháp quan trọng của từ loại danh từ. Bao gồm: *Những, các, một, ... Mỗi, từng, mọi, ... Cái, mấy, ...*
Ví dụ: *Trâu đứng ăn [năm/D]*
Đi cách đây [hai/D] cây số.

IN - Giới từ (Kết từ chính phụ)
Khái niệm: Kết từ chính phụ chỉ ý nghĩa quan hệ chính phụ. Kết từ chính phụ dùng để nối kết thành tố phụ vào thành tố chính (nối kết từ phụ với từ chính, thành phần phụ với thành phần chính của câu...).
Phân loại
Bao gồm: của, cho, bằng, do, vì, tại, bởi, để, mà, ở, ở tại, đối với, với, cùng, cùng với, về, đến, tới, từ, trong, ngoài, trên, dưới, giữa... tuy, dù, mặc dù, nhưng... nếu /giả /hễ /miễn /giả thử/... thì /là / thì là... *Ví dụ*
Quần áo [của/IN] tôi để đấy, tôi tự giặt lấy.
Tôi yêu anh [vì/IN] những nguyên nhân sâu xa hơn.
Tôi định [để/IN] mặc, đến lúc nào hay lúc ấy.
Cả hai chúng em phải dựa vào nhau [mà/IN] sống.

JJ - Tính từ
Khái niệm: Là lớp từ chỉ ý nghĩa đặc trưng. Ý nghĩa đặc trưng được biểu hiện trong tính từ thường có tính chất đối lập phân cực hoặc có tính chất mức độ. Tính từ có khả năng kết hợp được với phụ từ, nhưng không kết hợp được với "hãy", "đừng", "chớ". Tính từ cũng có thể kết hợp được với thực từ đi kèm để bổ nghĩa cho tính từ. Làm vị ngữ trong câu được coi là chức năng chính của tính từ, nhưng tính từ cũng được dùng kèm danh từ hoặc động từ để bổ nghĩa cho danh từ hay động từ.

Phân loại Bao gồm:
Tốt, đẹp, xấu, khéo, vụng, ...
Nhiều, ít, rậm, thưa, ngắn, dài, ...
Mạnh, yếu, nóng, lạnh, sáng, tối, lạnh lẽo, ...
Vuông, tròn, thẳng, gãy, ...
Xanh, đỏ, vàng, nâu, ...

Ồn, im, vắng, ồn ào, lặng lẽ, ...
Thối, đắng, cay, ngọt, bùi, ...
Riêng, chung, công, tư, ..
Đỏ lòm, trắng phau, đen sì, xanh xanh, ...
Ồn ào, đùng đùng, lè tè, lênh khênh, ... *Ví dụ*
Tôi nghe tiếng máy tàu [hu hu/JJ] mỗi lúc một gần.
Ý nghĩ nó [nhoang nhoáng/JJ] qua đầu như trời chớp vậy

NC - Danh từ chỉ loại

Khái niệm: Danh từ chỉ loại là tất cả những từ có tính chất từ loại của danh từ và có nội dung ý nghĩa chỉ thứ, loại, hạng của sự vật, kể cả những danh từ có kiểu ý nghĩa từ vựng trực tiếp chỉ loại là các từ như thứ, loại, hạng, kiểu, ... Chúng mang đầy đủ đặc tính của danh từ, dùng rời được như một từ đơn, có thể kết hợp phía sau với những từ này, nọ. Chúng vừa có tác dụng sắp xếp các sự vật vào cùng loại khái quát, đồng thời lại có khả năng làm cho sự vật tách bạch ra thành đơn vị rời, thành vật lẻ, đếm được.

Phân loại: Các danh từ chỉ loại thường gặp với vai trò thành tố chính cụm danh từ và trực tiếp đứng sau số từ số đếm là:
Chỉ loại, chỉ đơn vị tập hợp: bọn, lũ, tốp, đám, đoàn, đội, ...
Chỉ loại, chỉ đơn vị riêng lẻ: con, cái, đứa, bức, mét, kg, giờ, ...
Cục, hòn, miếng, mẩu, vụn, hạt, thanh, tấm, ...
Làn, cơn, trận, ...
Tên, tay, đầu, gốc, chân, ...
Thứ, loại, hạng, kiểu, cách, ...
Nước, khu, tỉnh, huyện, xã, làng, ...
Chỗ, nơi, chốn, xứ, miền, khu, khoảnh, vùng, ...
Màu, sắc, mùi, hương, vị, tiếng, giọng, ...

Ví dụ
Một con//NC gà, ra bờ ao.
Hai cái//NC bàn ở trong nhà đều mới.
Đồ cục//NC đất.
Cho tôi xem bức//NC ảnh này với.
Bọn//NC cướp thật độc ác.

NN - Danh từ thường

Khái niệm Là danh từ chỉ người, đồ đạc, động thực vật, khái niệm trừu tượng, ... Là danh từ không đếm được, thường đứng sau danh từ chỉ loại, kết hợp với danh từ chỉ loại làm thành tố chính của cụm danh từ.

Phân loại
Ví dụ: Hai đứa sinh viên nghèo ấy. Ở ví dụ trên, "đứa" là danh từ chỉ loại, "sinh viên" là danh từ thường. "đứa sinh viên" là thành tố chính của cụm danh từ "hai đứa sinh viên nghèo ấy".

Một số danh từ thường:
chị em, bàn ghế, nhà cửa...
chó, mèo, gà...

NP - Danh từ riêng

Khái niệm: Là tên riêng của từng người, từng sự vật cụ thể... *Ví dụ:*

- Danh từ chỉ tên riêng: Nguyễn Tất Thành, Hồ Chí Minh...
- Danh từ chỉ tên tỉnh, thành phố: Hà Nội, Hải Phòng, Sài Gòn...
...

PP - Đại từ

Khái niệm Đại từ là lớp từ dùng để thay thế và chỉ trỏ. Đại từ không trực tiếp biểu thị thực thể, quá trình hoặc đặc trưng như danh từ, động từ, tính từ. Đại từ nói chung, có thể đảm nhiệm các chức năng c ú pháp của thực từ được thay thế. Ngoài ra đại từ còn được dùng để thay thế và chỉ trỏ vào người và vật tham gia quá trình giao tiếp. *Phân loại*

Bao gồm:

Tôi, tao, mày, nó ,họ , ... , Ta, chúng ta, mình, chúng mình,
Nhau, ... Ai, ai ai,
Bây giờ, giờ, rày, nay, bấy nãy, bấy giờ
Này, đây, đó, kia, kìa, nọ, ấy, nọ
Bao lâu, ... , Đâu, nào,
Tất cả, cả thảy, hết thảy, bấy nhiêu, ..., Gì, sao, nào

Ví dụ

$$Ti/PP$$

muốn mua hai con mèo.

$$Chngta/PP$$

cùng đi mua nhé.

RB - Trợ từ

Khái niệm Dùng trong câu biểu thị ý nghĩa tình thái, bằng cách nhấn mạnh vào từ, kết hợp từ. . . có nội dung phản ánh liên quan với thực tại mà người nói muốn lưu ý người nghe. Vị trí của trợ từ thường tương ứng với chỗ ngừng hay chỗ ngắt đoạn khi phát ngôn câu. Do đó trợ từ có thể có tác dụng phân tách các thành phần câu.

Phân loại

Bao gồm:

Thì, ngay, ngay cả, đúng, đúng là, những, mà, là, chính, đích, nhất là, chỉ, chỉ là, thật, thật ra, thực ra, đến, đến cả, đến nỗi, tự.

Ví dụ

Tôi [thì/RB] tôi quay lại phía biển. [Ngay/RB] lúc chập tối, đồng chí Quỳnh đã đi.

[Đúng là/RB] tụi giặc đuổi theo rồi.

UH - Tình thái từ (Thán từ)

Khái niệm Là tiểu từ chuyên dùng biểu thị ý nghĩa tình thái trong quan hệ của chủ thể phát ngôn với người nghe hay với nội dung phản ánh; hoặc ý nghĩa tình thái gắn với mục đích phát ngôn.

Tình thái từ có vị trí trong câu rất linh hoạt. Chúng có thể đặt đầu câu, hoặc cuối câu hay ở trong câu. Khi đứng trong câu, tình thái từ thường có tác dụng phân tách ranh giới các thành phần câu, tạo dạng thức các kiểu câu theo mục đích phát ngôn. Tình thái từ có thể đứng riêng biệt, làm thành câu đặc biệt.

Phân loại

Bao gồm:

à, ư, chăng, hử, hả, không, phỏng,
đi, với, nhé, mà, nào, thôi,
à, á, vậy, kia, mà, cơ, cơ mà, hứ, hé, thật,
ôi, ối, ái, ồ, ái chà, ...
ơi, hỡi, à, này, vâng, dạ, đây ...

Ví dụ

VB - Động từ

Khái niệm Động từ là những từ biểu thị ý nghĩa khái quát về quá trình. Ý nghĩa quá trình thể hiện trực tiếp đặc trưng vận động của thực thể. Đó là ý nghĩa hành động. Ý nghĩa trạng thái được khái quát hóa trong mối liên hệ với vận động của thực thể trong thời gian và không gian. Động từ thường có các phụ từ đi kèm, để biểu thị các ý nghĩa quan hệ có tính tình thái giữa quá trình với cách thức và với các đặc trưng vận động của quá trình trong không gian, trong thời gian và trong hiện thực. Động từ có thể kết hợp được với thực từ nhằm phản ánh các quan hệ trong nội dung vận động của quá trình. Khả năng kết hợp với "hãy", "đừng", "chớ" có tác dụng quy loại động từ. Động từ có khả năng đảm nhiệm nhiều chức năng Cú pháp khác nhau, nhưng chức năng phổ biến và quan trọng nhất là làm vị ngữ trong cấu tạo câu.

Ví dụ

Chị đừng hỏi/VB.

Chị hãy bình tĩnh/VB lại.

Ai buôn/VB thì mặc/VB người ta, con

Thành ngữ *Khái niệm*

Phân loại

Ví dụ