

XÂY DỰNG HỆ THỐNG PHÂN TÍCH CÚ PHÁP TIẾNG VIỆT SỬ DỤNG VĂN PHẠM HPSG

Implementing a Vietnamese syntactic parser using HPSG

Đỗ Bá Lâm, Lê Thanh Hương
Khoa Công nghệ Thông tin, trường Đại học Bách khoa Hà Nội

Tóm tắt

Bài này giới thiệu một cách tiếp cận phân tích cú pháp tiếng Việt sử dụng văn phạm cấu trúc đoạn hướng trung tâm (Head-Driven Phrase Structure Grammar - HPSG). Cách tiếp cận này cho phép xử lý các vấn đề bùng nổ tổ hợp, nhập nhằng cấu trúc, và các câu đặc biệt bằng cách sử dụng các luật cấu tạo cú pháp và ràng buộc ngữ nghĩa. Chúng tôi đề xuất cách biểu diễn và quản lý luật HPSG cho tiếng Việt dựa trên các đặc điểm riêng của ngôn ngữ này. Đồng thời, chúng tôi đề xuất các cải tiến với giải thuật Earley cho HPSG. Kết quả thử nghiệm cho thấy hệ thống này có kết quả chính xác hơn so với các hệ thống phân tích cú pháp tiếng Việt hiện có.

Từ khóa: phân tích cú pháp, HPSG, tiếng Việt

Abstract

This paper presents an approach to Vietnamese syntactic parsing using Head-Driven Phrase Structure Grammar (HPSG). This approach permits us handle structural ambiguities, combination explosion, and ill-formed sentences by using syntactic and shallow semantic constraints. A presentation of rule set in HPSG is proposed, basing on characteristics of Vietnamese grammar. An improvement of the Earley parsing algorithm for HPSG is presented. Experimental results show that our system provides more accurate results comparing to other existing Vietnamese syntactic parsers.

Keywords: Vietnamese, syntactic parsing, HPSG

1. Giới thiệu

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

Việc phân tích cú pháp câu có thể chia làm hai mức chính. Mức thứ nhất là tách từ và xác định thông tin từ loại. Mức thứ hai là sinh cấu trúc cú pháp cho câu dựa trên các từ và từ loại do bước trước cung cấp. Do tiếng Việt là ngôn ngữ đơn âm tiết nên chúng ta thường gặp phải vấn đề nhập nhằng ở cả hai mức. Chúng ta đã có một số bộ tách từ với độ chính xác tương đối cao [8]. Vì vậy chúng tôi chỉ tập trung giải quyết mức sinh cấu trúc cú pháp

câu. Các khả năng nhập nhằng ở bước này có thể do nguyên nhân sau:

1. Một từ có thể có nhiều ý nghĩa khác nhau và nhiều chức năng ngữ pháp trong các ngữ cảnh khác nhau. Ví dụ từ “đá” đầu tiên trong câu “con ngựa đá con ngựa đá” là một động từ, trong khi từ “đá” thứ hai là một tính từ.
2. Một câu có thể có nhiều cây cú pháp khác nhau, trong đó chỉ có một cây đúng. Lý do là có nhiều luật cú pháp có thể áp dụng để phân tích câu mà không cần quan tâm đến ngữ nghĩa của câu đó.
3. Một câu có thể hiểu theo nhiều cách khác nhau. Vì lý do này, một câu cũng có thể có nhiều cây cú pháp đúng.

Một vấn đề khác trong phân tích cú pháp tiếng Việt là các hiện tượng ngữ pháp đặc biệt. Ví dụ, hiện tượng thiếu giới từ trong các

danh ngữ. Các danh ngữ với cấu trúc cú pháp này đúng trong một số trường hợp nhưng lại không đúng trong các trường hợp khác. Chúng ta có thể nói “*bạn tôi*”, “*con tôi*”, nhưng lại không thể nói “*sách tôi*”, “*ghế tôi*”. Thay vì thế, ta phải nói “*sách của tôi*”, “*ghế của tôi*”. Phần lớn các hệ thống phân tích cú pháp coi trường hợp “*sách tôi*”, “*bút tôi*” là đúng ngữ pháp.

Để giải quyết vấn đề này, chúng ta cần đưa thông tin cú pháp và ngữ nghĩa vào tập luật văn phạm. Chúng tôi thêm thông tin vào các luật cú pháp bằng cách sử dụng văn phạm cấu trúc đoạn hướng trung tâm (Head-Driven Phrase Structure Grammar - HPSG). Văn phạm này cho phép biểu diễn các mối quan hệ giữa các từ, và làm tăng ràng buộc kết hợp. Thuật toán Earley cải tiến tích hợp cấu trúc thuộc tính của HPSG cho phép chúng tôi thực hiện xử lý nhập nhằng về cú pháp và các câu không đúng ngữ pháp trong tiếng Việt.

Phần tiếp theo của bài này được tổ chức như sau. Cách tổ chức biểu diễn văn phạm HPSG cho tiếng Việt được giới thiệu ở phần 2. Phần 3 trình bày sự cải tiến đối với thuật toán Earley cho văn phạm HPSG. Các kết quả thử nghiệm được trình bày trong phần 4. Phần 5 kết luận và đề xuất hướng phát triển cho cách tiếp cận này.

2. Văn phạm HPSG

HPSG [9] tạm dịch là văn phạm cấu trúc đoạn hướng trung tâm, do Carl Pollard và Ivan Sag đưa ra với mục đích xây dựng một học thuyết khoa học về khả năng hiểu ngôn ngữ nói. HPSG có thể được nhìn nhận như một sự mở rộng của văn phạm phi ngữ cảnh (context free grammar – CFG) bằng việc thêm vào các thuộc tính trong cấu trúc mô tả từ và các ràng buộc trong các luật cú pháp. Khi đó quá trình phân tích cú pháp sẽ là sự kết hợp giữa luật cú pháp và những ràng buộc ngữ nghĩa. HPSG có hai đặc điểm chính:

1. HPSG sử dụng cấu trúc thuộc tính để biểu diễn các thông tin về từ. Cấu trúc này thường được mô tả dưới dạng một ma trận giá trị thuộc tính (attribute-value-matrix (AVM)), nhằm mô tả các đặc tính cụ thể của từ như các thông tin cú pháp và ngữ nghĩa.

2. HPSG tích hợp các ràng buộc về cú pháp và ngữ nghĩa vào tập luật. Các ràng buộc này được dùng để kiểm soát các quan hệ cú pháp và ngữ nghĩa giữa các từ/ngữ trong câu.

2.1. Mô hình biểu diễn từ và ngữ tiếng Việt

Một AVM biểu diễn từ/ngữ trong HPSG có thể rất phức tạp như đã được giới thiệu trong [10]. Tuy nhiên trong biểu diễn từ và ngữ cho tiếng Việt, chúng tôi sử dụng một AVM đơn giản hơn. Cấu trúc này chú trọng vào các quy tắc kết hợp ngữ pháp của động từ. Lý do là, với các ngôn ngữ, động từ là thành phần quan trọng nhất, có tác dụng gắn kết các thành phần khác trong câu. AVM của từ được biểu diễn như sau:

$$\left[\begin{array}{l} \textit{Phon} \\ \textit{Head} \left[\begin{array}{l} \langle \textit{Category} \rangle \\ \langle \textit{SubCategory} \rangle \\ \langle \textit{Category Meaning} \rangle \end{array} \right] \\ \textit{Spr} \left[\begin{array}{l} \langle \textit{SubCategory} \rangle \\ \langle \textit{Category Meaning} \rangle \end{array} \right] \\ \textit{Comp} \left[\begin{array}{l} \langle \textit{SubCategory} \rangle \\ \langle \textit{Category Meaning} \rangle \end{array} \right] \end{array} \right]$$

trong đó

- Phon: thể hiện từ
- Head: cho biết thông tin về bản thân từ/cụm từ. Head gồm 3 thuộc tính là từ loại (Category), tiểu từ loại (SubCategory), và nghĩa loại (CategoryMeaning) của từ. Các nhãn nghĩa loại (CategoryMeaning) được quản lý bởi một cây ngữ nghĩa thiết lập sẵn. Cây ngữ nghĩa này do Trung tâm từ điển học xây dựng [14].
- Spr và Comp gồm 2 thuộc tính là: SubCategory và CategoryMeaning. Spr (Specifier) thể hiện những ràng buộc của từ về tiểu từ loại và nghĩa loại với từ/ngữ đứng trước, còn Comp (Complement) thể hiện những ràng buộc về tiểu từ loại và nghĩa loại của từ với từ/ngữ đứng sau.

Ví dụ: từ “ăn” trong câu “anh ăn bánh”

Phon	ăn
Head	$\left[\begin{array}{c} V \\ Vt \\ Action \end{array} \right]$
Spr	$\left[\begin{array}{c} N \\ LivingThing \end{array} \right]$
Comp	$\left[\begin{array}{c} N \\ Food \end{array} \right]$

Từ “ăn” có mẫu động từ là Sub+V+Dob, với chủ ngữ (Sub) phải là danh từ (N) và có nghĩa loại (CategoryMeaning) là vật thể sống (LivingThing), bổ ngữ trực tiếp (Dob) phải là danh từ (N) và có nghĩa loại (CategoryMeaning) là thức ăn (Food). Những ràng buộc này được đưa vào hai cấu trúc Spr và Comp. Vì vậy ta có ma trận AVM của từ “ăn” như trên. Trong trường hợp từ không có thông tin về Spr và Comp, các giá trị của hai thuộc tính này sẽ được bỏ trống.

Ma trận AVM mà chúng tôi đề xuất cũng biểu diễn được những ràng buộc ngữ nghĩa cho các từ loại khác. Do từ điển mà chúng tôi sử dụng hiện mới chỉ có các ràng buộc liên quan đến động từ nên các ràng buộc đối với các từ loại khác sẽ thể hiện qua tập luật cú pháp.

2.2. Xây dựng tập luật cú pháp HPSG cho tiếng Việt

Như trên đã nói, có thể coi HPSG là mở rộng của văn phạm phi ngữ cảnh bằng cách tích hợp các ràng buộc thuộc tính của từ/ngữ vào tập luật cú pháp. Với các luật cú pháp HPSG, ngoài các ràng buộc tường minh thể hiện qui tắc kết hợp các thành phần ngữ pháp (ví dụ, $VP \rightarrow V N$) còn có các ràng buộc tiềm ẩn trong cấu trúc thuộc tính của từ. Khi kiểm tra khả năng áp dụng của một luật cú pháp đối với một ngữ cụ thể, ta cần kiểm tra cả hai loại ràng buộc này. Việc kiểm tra các ràng buộc tiềm ẩn có thỏa mãn hay không được thực hiện qua phép hợp nhất thuộc tính. Phép hợp nhất thuộc tính này còn nhằm xác định thuộc tính của ngữ trên cơ sở thuộc tính của các thành phần cấu tạo nên nó. Sau đây chúng tôi

sẽ giới thiệu chi tiết cách biểu diễn luật và quy tắc hợp nhất thuộc tính do chúng tôi đề xuất.

2.2.1 Luật cú pháp HPSG và quy tắc hợp nhất thuộc tính

Tập luật mà chúng tôi đề xuất là một tập luật có tích hợp cấu trúc thuộc tính, do vậy phải đưa ra một quy tắc hợp nhất để xác định giá trị các thuộc tính của ngữ thu được. Trong mỗi kết hợp đều phải xác định một thành phần trung tâm (Head). Quy tắc xác định cấu trúc AVM của ngữ như sau:

- Giá trị Phon sẽ là sự kết hợp giá trị Phon từ các thành phần trong về phải luật.
- Giá trị Head.Category là ngữ loại của về trái luật
- Giá trị Head.SubCategory được nhận từ giá trị SubCategory của thành phần trung tâm
- Giá trị Head.CategoryMeaning được nhận từ giá trị CategoryMeaning của thành phần trung tâm.
- Nếu thành phần trung tâm đã thực hiện quá trình hợp nhất dựa trên ràng buộc về Spr hay Comp thì giá trị các thuộc tính trong Spr hay Comp của ngữ thu được sẽ được bỏ trống. Ngược lại chúng nhận các giá trị từ Spr và Comp của thành phần trung tâm.

Chúng tôi minh họa với việc phân tích động ngữ: “ăn bánh” với luật cú pháp HPSG biểu diễn tường minh các ràng buộc tiềm ẩn:

$$\begin{array}{ll}
 1. VP & \rightarrow V + N \\
 V.Comp.SubC & \supset N.Head.SubC \\
 V.Comp.CatM & \supset N.Head.CatM \\
 Head & = 1
 \end{array}$$

Ở đây cần phân biệt Head trong ma trận AVM biểu diễn của từ/ngữ (ví dụ, N.Head) và Head trong luật (ví dụ, Head = 1). Trong các luật, giá trị Head cho biết số thứ tự của thành phần trung tâm, với việc đánh số bắt đầu từ 0. Ví dụ trong luật trên, VP, V, N có số thứ tự lần lượt là 0, 1, 2. Head = 1 có nghĩa thành phần trung tâm của VP là V.

$$\begin{array}{c} \left[\begin{array}{l} \text{Phon } \textit{ăn} \\ \text{Head } \left[\begin{array}{l} V \\ Vt \\ \textit{Action} \end{array} \right] \\ \text{Spr } \left[\begin{array}{l} N \\ \textit{LivingThing} \end{array} \right] \\ \text{Comp } \left[\begin{array}{l} N \\ \textit{Food} \end{array} \right] \end{array} \right] + \begin{array}{c} \left[\begin{array}{l} \text{Phon } \textit{bánh} \\ \text{Head } \left[\begin{array}{l} N \\ Nc \\ \textit{Dish} \end{array} \right] \\ \text{Spr } \left[\begin{array}{l} \\ \end{array} \right] \\ \text{Comp } \left[\begin{array}{l} \\ \end{array} \right] \end{array} \right] = \begin{array}{c} \left[\begin{array}{l} \text{Phon } \textit{ăn bánh} \\ \text{Head } \left[\begin{array}{l} VP \\ Vt \\ \textit{Action} \end{array} \right] \\ \text{Spr } \left[\begin{array}{l} N \\ \textit{LivingThing} \end{array} \right] \\ \text{Comp } \left[\begin{array}{l} \\ \end{array} \right] \end{array} \right] \end{array}$$

Trong luật cú pháp trên, phép “ \subset ” biểu diễn quan hệ “thành phần con”. Phép “ \subset ” được sử dụng thay vì phép bằng “ $=$ ” trong quá trình hợp nhất là vì

- Về CategoryMeaning: giá trị ràng buộc đối với CategoryMeaning trong động từ luôn mang nghĩa khái quát nhất. Ví dụ đứng trước từ “*ăn*” phải là từ có nghĩa là LivingThing (vật thể sống). LivingThing lại chứa trong nó nhiều nghĩa loại nhỏ hơn như People (con người), Animal (động vật)... và trong People, Animal lại có thể chia nhỏ hơn như Person (cá nhân), Organization (tổ chức), Mammal (thú)... do vậy với CategoryMeaning phải sử dụng phép toán chứa “ \subset ”. Khi đó các chủ ngữ như: “*anh*” trong “*anh ăn bánh*”, “*con mèo*” trong “*con mèo ăn bánh*”... đều thỏa mãn ràng buộc vì chủ ngữ của chúng có CategoryMeaning thuộc về lớp LivingThing.
- Về SubCategory: tuy từ “*ăn*” ràng buộc đứng trước có Category là N, nhưng chúng tôi vẫn đưa vào thuộc tính SubCategory của Spr. Từ đó sử dụng phép toán “ \subset ” để kiểm tra quan hệ với SubCategory của từ “*bánh*” là Nc. Việc đưa ràng buộc Category vào SubCategory có thể gây một chút nghi ngờ về sự không rõ ràng. Nhưng nếu chúng ta xét đứng trước là danh ngữ như “*anh tôi*” hay “*anh của tôi*”, chủ ngữ sẽ là NP chứ không phải là N nữa. Điều đó cho thấy phải xử lý linh hoạt ràng buộc về từ loại đứng trước. Đối với các danh ngữ này, NP sẽ có SubCategory là Nc (là subCategory của từ “*anh*” – từ trung tâm), do vậy việc kiểm tra ràng buộc sẽ không bị thay đổi. Thuộc tính SubCategory còn được sử dụng để gia tăng ràng buộc giữa các

thành phần trong luật. Điều này sẽ khiến việc biểu diễn các luật cú pháp có thể chi tiết đến mức tiểu từ loại (SubCategory) thay vì chỉ đến mức từ loại (Category).

Những quy tắc trong xây dựng tập luật sẽ được trình bày cụ thể ở phần sau.

2.2.2 Các loại luật trong tập luật

Trong từ điển hiện chỉ có các động từ mới có giá trị ở hai thành phần Spr và Comp. Đối với các nhãn từ loại khác, các giá trị trong Spr và Comp đều để trống. Điều này sẽ làm hạn chế ràng buộc về ngữ nghĩa trong kết hợp các nhãn từ loại khác động từ với nhau. Do vậy chúng tôi đưa ra hai loại luật như sau.

- Loại thứ nhất: các luật thông thường. Loại luật này giống như các luật CFG, nhưng có bổ sung thêm thành phần Head để xác định thành phần trung tâm trong kết hợp. Loại luật này chủ yếu biểu diễn các quy tắc tạo ra động ngữ. Bởi vì bản thân động từ đã chứa các ràng buộc tiềm ẩn.
Ví dụ: $VP \rightarrow V + N$ Head = 1
- Loại thứ hai: các luật ràng buộc về tiểu từ loại và nghĩa loại đối với một thành phần nào đó trong luật. Các luật loại này cho phép bổ sung thêm thông tin ràng buộc đối với các từ loại khác ngoài động từ. Trong loại này có thể chia ra thành 3 loại con nhỏ hơn.
 - Ràng buộc có:

$$NP \rightarrow N@Nc\text{-Person,PartOfAnimal}$$

$$N@Nc\text{-Person} \text{ Head} = 1$$
 Các luật loại này quy định tiểu từ loại và nghĩa loại của một hay nhiều thành phần trong luật. Đối với luật trên danh từ thứ nhất phải có tiểu từ loại là Nc (danh từ đơn thể), và có nghĩa loại là Person (người) hay bộ phận của cơ thể

(PartOfAnimal), danh từ thứ hai phải có tiểu từ loại là Nc, và nghĩa loại là Person. Luật này được áp dụng cho các danh ngữ như “con anh”, “chân anh”... các danh ngữ như “bút anh”, “sách anh”... sẽ bị lỗi khi hợp nhất thuộc tính. Với luật này, chúng tôi đã xử lý được hiện tượng ngữ pháp đặc biệt như đã nêu trên

o Ràng buộc không:

NP → N@!Ns-!Concept P@Pd

Head = 1

Các luật loại này quy định một hay nhiều thành phần trong luật không được có tiểu từ loại là gì, và nghĩa loại là gì.

o Kết hợp:

Sub → N@Nc,Ng,Np-!Concept

Head = 1

Đây là các luật kết hợp cả hai điều kiện có và không. Một hay nhiều thành phần trong luật phải có tiểu từ loại là gì, không có nghĩa loại là gì, hoặc ngược lại.

Trong các biểu diễn luật, chúng tôi sử dụng kí hiệu “@” sau nhãn từ loại để xác định ràng buộc; dấu “-” để ngăn cách hai thuộc tính tiểu từ loại (SubCategory) và nghĩa loại (CategoryMeaning); dấu “,” với ý nghĩa là hoặc; dấu “!” với ý nghĩa là phủ định.

Với hai loại luật này, tập luật do chúng tôi đề xuất đã cho phép biểu diễn luật cú pháp chi tiết đến mức tiểu từ loại và nghĩa loại. Nó có khả năng bao phủ được những loại ràng buộc khi phân tích cú pháp dựa trên ngữ nghĩa. Chúng tạo ra nền tảng cho việc xây dựng tập luật có ràng buộc chặt chẽ hơn.

3. Thuật toán phân tích cú pháp cho văn phạm HPSG

Chúng tôi sử dụng giải thuật Earley [5] trong phân tích cú pháp. Khác với Earley áp dụng cho văn phạm phi ngữ cảnh truyền thống, chúng tôi phải tích hợp cấu trúc thuộc tính vào giải thuật Earley để đảm bảo các ràng buộc của luật..

Xét luật phân tích cú pháp biểu diễn tường minh ràng buộc tiềm ẩn:

VP → V + N

V.Comp.SubC ⊃ N.Head.SubC

V.Comp.CatM ⊃ N.Head.CatM

Head = 1

Chúng tôi nhận thấy những ràng buộc trong luật xuất phát từ những ràng buộc của từ. Do vậy chúng tôi kết hợp giữa cấu trúc biểu diễn từ và luật CFG để thực hiện biểu diễn luật mở rộng. Do đó luật sẽ gồm hai thành phần, thành phần thứ nhất là luật CFG: VP → V+N Head =1, thành phần thứ hai là ma trận AVM biểu diễn từ/ngữ mà chúng tôi đề xuất ở trên.

Dành một chút xem xét lại giải thuật Earley. Earley là một giải thuật sử dụng chiến lược top-down, và sử dụng bảng trong phân tích. Tại mỗi cột trong bảng, Earley thực hiện 3 bước

- Bước quét (*Scanning*): đọc từ trong câu, xác định luật phù hợp để phân tích từ này.
- Bước hoàn thiện (*Completion*): tìm kiếm một/nhiều luật trong cột trước đó phù hợp với luật đang được xem xét để tạo ra một/nhiều luật mới. Bước này thực hiện ghép các từ/ngữ đã phân tích lại với nhau và xác định chức năng cú pháp của ngữ này trong câu..
- Bước dự đoán (*Prediction*): khai triển các kí hiệu không kết thúc, dự đoán các khả năng của nhãn từ loại của từ được đọc tiếp theo.

Với việc bổ sung thêm ma trận AVM vào luật, chúng tôi thực hiện giải thuật Earley như sau.

- Bước quét: đọc ma trận AVM của từ, và gán cho ma trận AVM của luật.
- Bước hoàn thiện: bước này tương đương với phép toán hợp nhất thuộc tính. Ở bước hoàn thiện mở rộng này, ngoài việc tìm từng luật phù hợp như trong giải thuật ban đầu, chúng tôi kiểm tra sự hợp nhất về thuộc tính được biểu diễn trong các ma trận AVM. Nếu sự hợp nhất này là thành công, khi đó luật được tạo ra mới được đưa vào trong cột.

- Bước dự đoán: ma trận AVM của luật được khởi tạo mặc định gồm các giá trị rỗng vì chưa đọc được từ nào.

Ví dụ:

Xem xét quá trình phân tích của danh ngữ “*ăn bánh*”

Giả sử chúng ta đã phân tích được từ “*ăn*”. Khi đó AVM của luật này là AVM của từ “*ăn*”

$$VP \rightarrow V \bullet N, AVM_1$$

$$AVM_1 = \begin{bmatrix} Phon & \text{ăn} \\ Head & \begin{bmatrix} V \\ Vt \\ Action \end{bmatrix} \\ Spr & \begin{bmatrix} N \\ LivingThing \end{bmatrix} \\ Comp & \begin{bmatrix} N \\ Food \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} Phon & \text{ăn} \\ Head & \begin{bmatrix} V \\ Vt \\ Action \end{bmatrix} \\ Spr & \begin{bmatrix} N \\ LivingThing \end{bmatrix} \\ Comp & \begin{bmatrix} N \\ Food \end{bmatrix} \end{bmatrix} + \begin{bmatrix} Phon & \text{bánh} \\ Head & \begin{bmatrix} N \\ Nc \\ Dish \end{bmatrix} \\ Spr & \begin{bmatrix} \\ \end{bmatrix} \\ Comp & \begin{bmatrix} \\ \end{bmatrix} \end{bmatrix} = \begin{bmatrix} Phon & \text{ăn bánh} \\ Head & \begin{bmatrix} VP \\ Vt \\ Action \end{bmatrix} \\ Spr & \begin{bmatrix} N \\ LivingThing \end{bmatrix} \\ Comp & \begin{bmatrix} \\ \end{bmatrix} \end{bmatrix}$$

AVM_1
 AVM_2
 AVM

Sau khi bộ phân tích tiến hành đọc từ “*bánh*” trong bước quét, chúng ta có luật như sau

$$N \rightarrow \text{bánh} \bullet, AVM_2$$

$$AVM_2 = \begin{bmatrix} Phon & \text{bánh} \\ Head & \begin{bmatrix} N \\ Nc \\ Dish \end{bmatrix} \\ Spr & \begin{bmatrix} \\ \end{bmatrix} \\ Comp & \begin{bmatrix} \\ \end{bmatrix} \end{bmatrix}$$

Ở bước hoàn thiện, tiến hành hợp nhất thuộc tính trong hai ma trận AVM_1 và AVM_2 . Nếu quá trình hợp nhất thành công, một luật mới được đưa vào cột trong bảng phân tích Earley với AVM là sự hợp nhất thuộc tính của hai ma trận AVM trên.

$$VP \rightarrow VN \bullet, AVM$$

4. Các thử nghiệm

Để có một đánh giá khách quan về hệ thống, chúng tôi tiến hành thử nghiệm hệ thống trong 2 trường hợp.

- Trường hợp thứ nhất là 12 câu đơn giản trong đó có chứa câu sai do thiếu giới từ trong danh ngữ. Trường hợp thử nghiệm này cho ra kết quả mà mọi người đều có thể kiểm chứng về mặt nội dung vì cấu trúc cú pháp đơn giản.
- Trường hợp thứ hai là 9 câu phức tạp đã được các chuyên gia ngôn ngữ phân tích từ trước để so sánh kết quả.

Trường hợp thứ nhất: 12 câu đơn giản

1. Tôi sẽ mua một quyển sách.
2. Tôi mua tất cả những quyển sách.
3. Tôi mua quyển sách màu xanh.
4. Cái máy tính mà tôi mua đang đọc dữ liệu.
5. Cô ấy rất xinh.
6. Cô ấy hơi xinh
7. Tôi sẽ ăn cơm.
8. Quả bóng màu xanh
9. Con chó của tôi đang ăn cơm.
10. Con của tôi đang ăn cơm.
11. Con chó đang ăn cơm.
12. Con chó anh đang ăn cơm.

Hệ thống BKParser do chúng tôi xây dựng đã đưa ra được cấu trúc cú pháp chính xác của 11 câu đầu tiên. Câu thứ 12 hệ thống đã nhận

biết được sai về mặt ngữ pháp. Trong 11 câu phân tích được, chỉ có câu số 8 bị nhập nhằng ra 2 cây cú pháp. Kết quả này có được là nhờ hệ thống của chúng tôi đã xây dựng được một tập luật có ràng buộc chặt chẽ. Bên cạnh đó hệ thống sử dụng một từ điển được thiết kế mới (chứa các thông tin ngữ nghĩa của từ) do Trung tâm từ điển học xây dựng. Từ điển này có độ chính xác cao nên đã góp phần hạn chế sự nhập nhằng trong phân tích.

Trường hợp thứ hai: 9 câu phức tạp

1. Gió chướng thổi mạnh, chiếc ghe cào như muốn rung lên.
2. Hàm răng tôi cũng đánh lập cập.
3. Chiếc ghe trong bờ to vậy mà ra tới cửa Hàm Luông sao bé tẹo.
4. Ba người con của ông Tám Hòa là Tư Lý, Năm Long, Út Tòng, tuổi ngoài đôi mươi, miệng ngậm ống hơi thả ngựa mình tự do xuống sông.
5. Tôi cũng ngậm ống hơi, đeo băng chì rồi lần dây mồi xuống theo.
6. Càng xuống sâu nước càng lạnh, ép tai, nghe lũng bùng.
7. Năm Long bắt đầu vác neo khum người đi theo dòng nước.
8. Tôi lọ mọ theo sau, thấy hơi rờn rợn người.
9. Vừa qua khỏi đụn cát, chân tôi trơn tuột như giẫm phải mỡ.

Trong 9 câu trên, hệ thống BKParser phân tích chính xác 6 câu, không phân tích được câu 1, câu số 3 và 5 bị nhập nhằng ra 2 cây. Nguyên nhân sai ở câu 1 là do trong phân tích cú pháp mẫu của các chuyên gia ngôn ngữ đối với câu này, “*ghè cào*” không được coi là có trong từ điển mà là sự kết hợp giữa danh từ “*ghè*” và động từ “*cào*” để tạo một danh ngữ. Tuy vậy chúng tôi nhận thấy từ “*ghè cào*” cũng giống như các danh từ “*cây trồng*”, “*áo khoác*”, “*khăn quàng*” đều chỉ vật thể. Do vậy việc đưa từ “*ghè cào*” vào từ điển giống như các từ này là điều hợp lý. Đối với các danh ngữ khác như “*cuộc chiến đấu*”, “*phong trào đấu tranh*”... chúng tôi khởi tạo luật giữa danh từ và động từ nội động để tạo ra danh ngữ.

5. Kết luận

Trong nghiên cứu này, chúng tôi đã thực hiện được các nội dung sau:

- Đưa ra mô hình biểu diễn từ theo văn phạm HPSG. Mô hình này tập trung vào việc mô tả cấu trúc động từ - thành phần quan trọng nhất trong câu. Đồng thời mô hình này cũng cho phép mô tả mối quan hệ ràng buộc giữa các từ loại khác.
- Xây dựng mô hình biểu diễn luật chứa các ràng buộc cú pháp và ngữ nghĩa. Mô hình này dựa trên sự mở rộng của luật trong CFG, bổ sung thêm thành phần Head xác định thành phần trung tâm trong ngữ. Với việc đưa ra hai loại luật, tập luật của chúng tôi cho phép bao phủ ràng buộc giữa các thành phần dựa trên thông tin ngữ nghĩa.
- Xây dựng giải thuật phân tích cho mô hình biểu diễn từ và luật đề xuất. Trong mô hình này, luật bao gồm hai thành phần. Một thành phần biểu diễn biểu thức luật. Thành phần còn lại là cấu trúc biểu diễn từ hoặc ngữ.

Hệ thống phân tích cú pháp tiếng Việt sử dụng văn phạm HPSG đã được cài đặt. Do hạn chế về thời gian nên hiện tại chúng tôi mới xây dựng được một tập luật HPSG nhỏ với 95 luật. Tập luật này đã cho phép phân tích được các câu đơn và câu ghép trong loại câu trần thuật. Kết quả phân tích cho thấy sự nhập nhằng đã được hạn chế đáng kể. Bộ phân tích cho kết quả tương đối khả quan.

Trong thời gian tới, chúng tôi sẽ phát triển tập luật để nâng cao khả năng phân tích và độ chính xác hệ thống. Tập luật mới cần phân tích được các loại câu đa dạng hơn như câu trần thuật, câu cảm thán, câu cầu khiến và câu hỏi. Đồng thời, tập luật cần cho phép giảm thiểu các hiện tượng nhập nhằng có thể xảy ra với tiếng Việt.

Lời cảm ơn

Nghiên cứu này được thực hiện trong khuôn khổ Đề tài Nhà nước “Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt” mã số KC01.01/06-10.

Tài liệu tham khảo

- [1]Jame Allen. Natural language understanding. Addison Wesley. 1995
- [2]Bộ giáo dục và đào tạo. Ngữ pháp tiếng Việt. Giáo trình trường Cao đẳng Sư phạm. NXB Giáo dục. 2000.
- [3]Diệp Quang Ban. Ngữ pháp tiếng Việt, NXB Giáo Dục. 1998
- [4]Daniel Jurafsky, James H. Martin. Speech and language processing, Prentice Hall. 2000.
- [5]J. Earley. An efficient context-free parsing algorithm. 1970.
- [6]Lê Thanh Hương. Phân tích cú pháp tiếng Việt. Luận văn cao học. ĐHBK Hà Nội. 2000
- [7]Nguyễn Hữu Quỳnh. Ngữ pháp tiếng Việt, NXB Từ điển Bách Khoa Hà Nội. 2001
- [8]Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phương. Sử dụng bộ gán nhãn từ loại xác suất Qtag cho văn bản tiếng Việt. Hội thảo khoa học quốc gia lần thứ nhất về Nghiên cứu phát triển và ứng dụng công nghệ thông tin và truyền thông, ICT.rda. 2003
- [9]Pollard, C.J., Sag, I. Head-Driven Phrase Structure Grammar, CSLI Publications/Cambridge University Press. 1994.
- [10]Susanne Riehemann. *The HPSG Formalism*. Unpublished manuscript: Stanford University. 1995. <http://www-csli.stanford.edu/~sag/L221a/hand2-formal.pdf>
- [11] A basic overview of HPSG. http://www.emsah.uq.edu.au/linguistics/Working%20Papers/ananda_ling/HPSG_Summary.htm
- [12] Head-driven phrase structure grammar. http://en.wikipedia.org/wiki/Head-driven_phrase_structure_grammarHPSG
- [13] Linguistic approach, formal foundations, computational realization. <http://www.ling.ohio-state.edu/~dm/papers/ell2-hpsg.pdf>
- [14] Vietlex Semantic Tree. 2008. <http://www.vietlex.com/resources/semanticTree.html>