Phân tích cú pháp tiếng Việt sử dụng văn phạm phi ngữ cảnh từ vựng hoá kết hợp xác suất

Nguyễn Quốc Thể, Lê Thanh Hương
Khoa Công nghệ Thông tin - Trường Đại học Bách khoa Hà Nội

Tóm tắt
Trong bài này, chúng tôi nghiên cứu phương pháp xử lý hiện tượng nhập nhằng và các hiện tượng cú pháp phụ thuộc từ trong phân tích cú pháp tiếng Việt. Chúng tôi đề xuất việc xây dựng một công cụ phân tích cú pháp dựa trên văn phạm phi ngữ cảnh với luật có chứa thông tin về xác suất và từ vựng. Xác suất luật được tính dựa trên tập ngữ liệu mẫu, sử dụng mô hình bigram, kết hợp với phương pháp làm trơn nội suy tuyến tính để giảm ảnh hưởng của từ cụ thể đối với xác suất. Việc phân tích cú pháp câu được tiến hành dựa trên từ trọng tâm của câu (từ điều khiển trung tâm). Các kết quả đạt được bước đầu cho thấy cách tiếp cận này khả thi.

Từ khoá: xử lý ngôn ngữ tự nhiên, phân tích cú pháp, xác suất, văn phạm phi ngữ cảnh

1. Giới thiệu

Phân tích cú pháp là một vấn đề cơ bản và quan trọng trong xử lý ngôn ngữ tự nhiên. Với một công cụ phân tích cú pháp tốt, chúng ta có thể tích hợp vào nhiều ứng dụng trong xử lý ngôn ngữ tự nhiên như dịch máy, tóm tắt văn bản, các hệ thống hỏi đáp, … để tăng tính chính xác của các ứng dụng đó. Hiện nay, các công cụ phân tích cú pháp tiếng Việt đã đạt được một số kết quả nhất định. Tuy nhiên, phần lớn các kết quả đạt được mới dừng ở một số trường hợp câu cơ bản như câu đơn và các câu ghép đơn giản. Hiện tượng nhập nhằng và những trường hợp đặc biệt trong phân tích câu vẫn chưa được giải quyết thoả đáng. Trong bài này, chúng tôi sẽ đề xuất cách giải quyết các vấn đề đó thông qua văn phạm phi ngữ cảnh có bổ sung thông tin về từ vựng và xác suất vào luật cú pháp. Việc phân tích cú pháp câu được tiến hành dựa trên từ trọng tâm của câu, sử dụng một phương pháp cải tiến của mô hình xác suất thống kê Collins [5].

Trong phần sau, chúng tôi sẽ trình bày một số vấn đề còn tồn tại trong phân tích cú pháp tiếng Việt và đề xuất cách giải quyết cho các vấn đề đó. Phần 3 giới thiệu một số nét chính trong việc sử dụng văn phạm phi ngữ cảnh từ vựng hoá kết hợp xác suất (Lexicalized Probability Context Free Grammar – LPCFG) vào phân tích cú pháp tiếng Việt. Cách tính xác suất luật dùng trong LPCFG được thảo luận phần 4. Tiếp theo, chúng tôi sẽ mô tả thuật toán phân tích cú pháp sử dụng LPCFG. Phần 6 giới thiệu một số kết quả đạt được. Cuối cùng là kết luận và hướng phát triển của nghiên cứu này.

2. Một số vấn đề trong phân tích cú pháp tiếng Việt

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

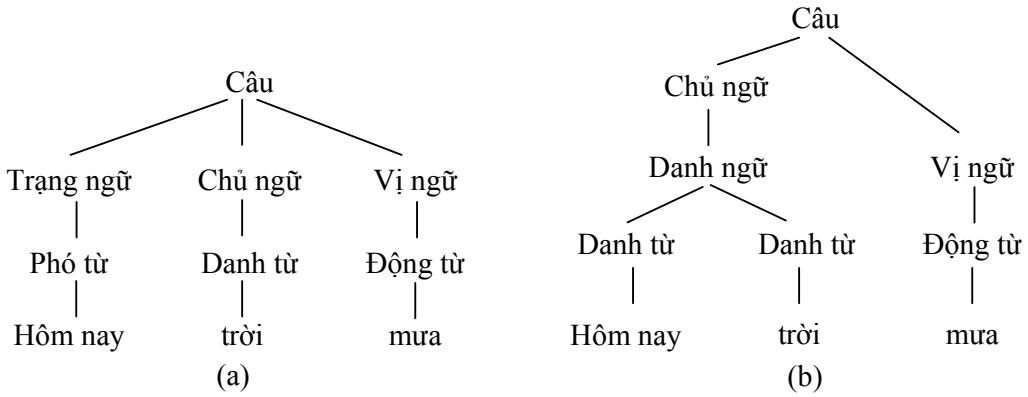
Trong phân tích cú pháp tiếng Việt, hiện tượng nhập nhằng xảy ra ở nhiều mức, từ mức từ, từ loại đến mức cú pháp câu. Điều này dẫn đến một câu có thể được phân tích theo nhiều cách khác nhau, trong khi chỉ có một vài cách phân tích trong số đó đúng. Trong bài này, chúng tôi chú trọng giải quyết vấn đề ở mức cú pháp. Qua khảo sát việc phân tích cú pháp các câu tiếng Việt, chúng tôi thấy có hai loại nhập nhằng. Một loại nhập nhằng do câu có thể hiểu theo nhiều nghĩa khác nhau dẫn đến các cây cú pháp khác nhau. Trong trường hợp này, mỗi cách hiểu sẽ ứng với một cây cú pháp và các cây cú pháp đó đều được chấp nhận. Ví

dụ câu "*Tôi nhìn thấy anh Hải ở tầng hai*" có thể hiểu theo hai cách. Cách thứ nhất, khi tôi nhìn thấy anh Hải thì anh ấy đang ở tầng hai. Trong trường hợp này, *ở tầng hai* bổ nghĩa cho danh ngữ *anh Hải*. Cách hiểu thứ hai, khi tôi đứng ở tầng hai thì tôi nhìn thấy anh Hải. Trong trường hợp này, *ở tầng hai* là bổ ngữ của *tôi nhìn thấy anh Hải*.

Với loại nhập nhằng thứ hai, câu chỉ có một nghĩa nhưng bộ phân tích cú pháp vẫn tạo ra nhiều cây cú pháp, trong đó chỉ có một cây đúng. Lý do của sự nhập nhằng này là quá trình phân tích cú pháp đã lược bỏ ngữ nghĩa từ/ngữ mà chỉ quan tâm đến nhãn cú pháp của chúng, dẫn đến nhiều luật cú pháp có thể áp dụng để phân tích câu. Ví dụ, với câu "*Hôm nay trời mưa*", tập luật cú pháp thuộc văn phạm phi ngữ cảnh (Context Free Grammar – CFG)[1] cần để phân tích câu này là:

1. <Câu>→<Chủ ngữ><Vị ngữ>
2. <Câu>→<Trạng ngữ><Chủ ngữ><Vị ngữ>
3. <Trạng ngữ>→<Phó từ>
4. <Danh ngữ>→<Danh từ><Danh từ>
5. <Chủ ngữ>→<Danh ngữ>
6. <Chủ ngữ>→<Danh từ>
7. <Vị ngữ>→<Động ngữ>
8. <Vị ngữ>→<Động từ>

Trong từ điển từ, *hôm nay* là danh từ hoặc phó từ, *trời* là danh từ, còn *mưa* là động từ. Với tập luật cú pháp trên, các cây cú pháp có thể sinh ra cho câu này được biểu diễn ở hình 1.



Hình 1 – Các cây cú pháp dựa trên tập luật phi ngữ cảnh của câu "*Hôm nay trời mưa*"

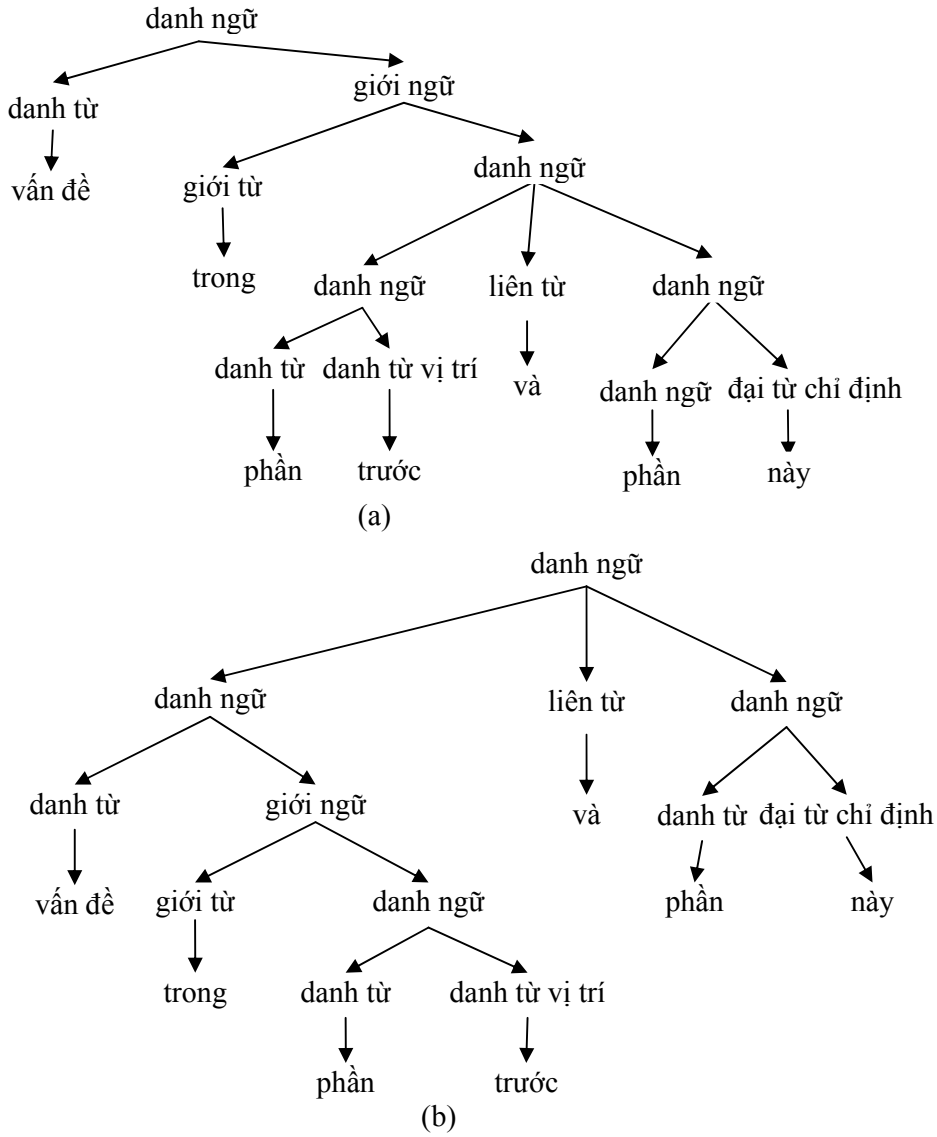
Trong hai cây cú pháp ở hình 1 chỉ có cây (a) đúng, cây (b) cần được loại bỏ hoặc không sinh ra trong quá trình phân tích cú pháp. Một trong những cách giải quyết vấn đề trên là phân loại chi tiết hơn các nhãn từ loại/ngữ loại và kiểm soát khả năng kết hợp giữa chúng. Thay vì luật <Danh ngữ>→<Danh từ><Danh từ>, ta đưa ra luật <Danh ngữ>→<Danh từ loại A><Danh từ loại B>. Nhược điểm của phương pháp này là:

- Hiện nay, việc đặt tên cho các từ loại/ngữ loại vẫn còn nhiều điểm chưa thống nhất. Vì vậy việc phân chia từ loại ở mức chi tiết hơn nữa sẽ càng có nhiều bất đồng quan điểm. Nếu việc này không phải do các nhà ngôn ngữ chuyên về tiếng Việt làm thì khả năng sai sót là rất lớn.
- Khi phân loại chi tiết các từ loại/ngữ loại, kích thước tập luật cú pháp tăng lên đáng kể.
- Với phương pháp này, muốn có một bộ phân tích cú pháp chính xác, chúng ta sẽ phải xây dựng một cách thủ công tập luật cú pháp ứng với tập nhãn từ loại mới. Việc xây dựng một cách đầy đủ tập tất cả các luật ở mức chi tiết như vậy là một giải pháp khó thực hiện do khó kiểm soát được tính chính xác và chặt chẽ của chúng. Hoàng Vĩnh Sơn [7] sử dụng cách tiếp cận này. Do việc đặt tên từ loại ở mức chi tiết có nhiều chỗ không hợp lý (từ loại đặt tên sai hoặc một từ có thêm nhiều từ loại chi tiết) và tập luật cú pháp chưa chuẩn, bộ phân tích cú pháp trong [7] đưa ra quá nhiều cây cú pháp tương tự nhau (chỉ khác tên từ loại) và nhiều cây cú pháp sai cho một câu đầu vào. Thay vì xây dựng

[1] Theo [8], văn phạm phi ngữ cảnh là lựa chọn thích hợp để phân tích cú pháp tiếng Việt.

tập luật cú pháp một cách thủ công, chúng ta cũng có thể tự động học luật cú pháp từ tập ngữ liệu mẫu. Tuy nhiên, giải pháp này yêu cầu tập ngữ liệu phải được phân tích chính xác và phải đủ lớn để bao phủ tất cả các khả năng kết hợp ngữ pháp có thể. Hiện nay chúng ta chưa có tập luật cú pháp hoặc tập ngữ liệu mẫu nào như vậy nên việc xây dựng một trong hai tập này đều khó ngang nhau.

Cách giải quyết thứ hai không cần phân loại chi tiết từ loại/ngữ loại là đưa xác suất vào tập luật cú pháp CFG. Văn phạm thuộc loại này gọi là văn phạm phi ngữ cảnh kết hợp xác suất (Probability Context Free Grammar – PCFG). Trong trường hợp bộ phân tích cú pháp sinh ra nhiều cây cú pháp cho một câu đầu vào, các cây cú pháp sẽ được xếp hạng dựa trên giá trị xác suất của cây đó. Công thức tính xác suất của một cây cú pháp T là: $P(T,S) = \Pi_{i=1...n} P(r_i)$ với S là ký hiệu không kết thúc, biểu diễn đỉnh xuất phát của cây cú pháp, r_i là luật áp dụng để sinh ra cây cú pháp, $P(r_i)$ là xác suất xuất hiện của luật r_i. Nhược điểm của cách tiếp cận này là chỉ dựa trên xác suất kết hợp giữa các từ loại/ngữ loại, chưa giải quyết được sự nhập nhằng liên quan đến tính chất của các từ cụ thể. Ví dụ, áp dụng PCFG vào danh ngữ "*vấn đề trong phần trước và phần này*", ta được hai cây cú pháp vẽ ở hình 2.



Hình 2 – Các cây cú pháp của danh ngữ "*vấn đề trong phần trước và phần này*"

Hai cây (2.a), (2.b) cùng áp dụng một tập luật phân tích cú pháp như nhau nhưng với thứ tự khác nhau. Theo cách tính P(cây)=Tích(P(các luật áp dụng))), hai cây cú pháp có giá trị xác suất ngang nhau nhưng chỉ có cây (2.a) đúng. Điểm mấu chốt trong phân tích danh ngữ này là: từ "*và*" thường dùng để kết nối hai phần có nội dung tương đương nhau, "*phần trước*" và "*phần này*" tương đương nhau hơn là "*vấn đề trong phần trước*" và "*phần này*". Vì vậy cây cú pháp (2.a) được chọn. Với những trường hợp như vậy, ngoài việc sử dụng xác suất, việc đưa thông tin của từ vào trong tập luật cú pháp là cần thiết.

2.2 Việc xác định câu đúng cú pháp đôi khi phụ thuộc vào các từ cụ thể cấu tạo nên câu

Như đã nói ở phần trên, để giải quyết nhập nhằng trong phân tích cú pháp, đôi khi chúng ta cần đến thông tin về từ cụ thể. Chúng ta còn gặp nhiều trường hợp khác trong tiếng Việt mà việc xác định câu đúng cú pháp hay không phụ thuộc vào từ cụ thể cấu tạo nên câu. Ví dụ "*Tôi ăn*" ít khi được chấp nhận là một câu hoàn chỉnh trong một ngữ cảnh chung. Tính hoàn chỉnh ở đây nhìn từ phía cảm nhận của người nghe, anh ta có cảm thấy thỏa mãn một lượng thông tin hay không. Trong ngữ cảnh chung "*Tôi ăn*" mang một giá trị thông tin nhỏ. Với câu này, nếu ta chỉ dựa trên các từ loại của câu và luật cú pháp câu có thể được hình thành từ một danh từ đứng trước một động từ thì câu trên hoàn toàn đúng ngữ pháp. "*Tôi đang ăn*" dễ được chấp nhận là câu hoàn chỉnh hơn vì trong một ngữ cảnh chung mệnh đề trên mang một giá trị thông tin khá lớn. Với những trường hợp nói trên, chúng ta phải dựa trên tính chất cụ thể của từ giữ vai trò chính trong câu hoặc ngữ để xác định xem câu/ngữ đó có đúng cú pháp hay không.

Trong phân tích cú pháp tiếng Việt, chúng ta còn thấy hiện tượng nhập nhằng do lược bỏ quan hệ từ. Chúng ta có thể nói *bạn tôi, con tôi* mà không nói *con chó tôi, con mèo tôi*. Trong trường hợp này, *bạn tôi, con tôi* cần được coi là các danh ngữ, trong khi *con chó tôi, con mèo tôi* cần được coi là các cụm từ sai ngữ pháp.

Qua các vấn đề đã phân tích ở trên, chúng tôi thấy rằng bản thân từ cũng có vai trò quan trọng trong quá trình phân tích cú pháp. Vì vậy, chúng tôi đề xuất việc xây dựng một công cụ phân tích cú pháp cho phép phân tích sâu hơn văn phạm phi ngữ cảnh kết hợp xác suất bằng cách đưa thông tin từ vựng vào văn phạm. Văn phạm này sẽ được trình bày kỹ hơn ở phần sau.

3. Phân tích cú pháp sử dụng Văn phạm phi ngữ cảnh từ vựng hoá kết hợp xác suất

Văn phạm phi ngữ cảnh từ vựng hoá kết hợp xác suất (Lexicalized Probability Context Free Grammar – LPCFG) là một biến thể của văn phạm phi ngữ cảnh[2] bằng cách đưa thêm xác suất luật và thông tin từ vựng vào các luật cú pháp. Trong văn phạm này, từ vựng đóng vai trò quan trọng trong việc xác định các từ/ngữ nào có thể kết hợp với nó.

Thành phần chính trong CFG là tập luật cú pháp. Với mô hình PCFG, mỗi luật cú pháp được gắn với xác suất sử dụng của nó. Nếu ta lưu các luật LPCFG theo cách lưu của mô hình CFG/PCFG, với mỗi luật đi kèm với các từ cụ thể thì không khả thi vì lúc đó số lượng các luật cần đưa vào bộ phân tích cú pháp quá lớn. Để giải quyết vấn đề này, ta sử dụng cách lưu trữ luật khác: ta chỉ ghi lại các thành phần chính của luật thay vì cả luật. Ví dụ:

(1) Câu → Động ngữ (3) Câu → Tính ngữ

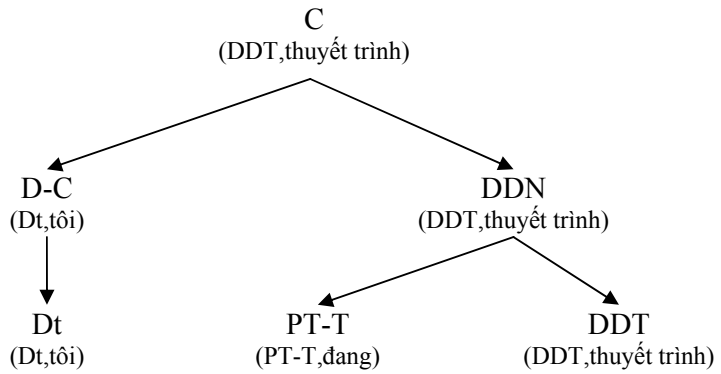
(2) Động ngữ → Động từ (4) Tính ngữ → Tính từ

Luật (1) có thể hiểu là một cụm từ có thành phần trung tâm là động ngữ thì có thể phát triển lên thành câu. Luật (2) có thể hiểu là một cụm từ có thành phần trung tâm là động từ thì có thể phát triển lên thành động ngữ.

[2] Trong bài này, chúng tôi giả thiết người đọc đã có kiến thức về văn phạm phi ngữ cảnh [9] và văn phạm phi ngữ cảnh dùng trong phân tích cú pháp tiếng Việt.

Một thành phần phụ có kết hợp được với thành phần trung tâm hay không phụ thuộc vào xác suất kết hợp với trung tâm, tính được dựa trên tập mẫu. Với mỗi tập mẫu cho trước, trước tiên chúng ta tiến hành trích rút luật dựa trên tần số xuất hiện của nó trong tập mẫu, sau đó rút ra các xác suất phụ thuộc giữa từ loại và ký hiệu không kết thúc. Cách biểu diễn luật và cách tính xác suất luật cú pháp trong LPCFG được mô tả trong phần 4.

Khi phân tích câu theo LPCFG, chúng ta dựa trên từ chính trong các thành phần câu. Thuật toán phân tích cú pháp LPCFG được trình bày ở phần 5. Mỗi thành phần câu trong cách biểu diễn phi ngữ cảnh được gắn với từ chính và từ loại tương ứng của nó. Ví dụ, cây cú pháp của câu "*Tôi đang thuyết trình*" trong LPCFG được biểu diễn như sau:



Hình 3 – Cây cú pháp của câu "*Tôi đang thuyết trình*"

trong đó C là câu, DDT là động từ, DDN là động ngữ, PT-T là phụ từ chỉ thời gian, DT là danh từ, D-C là đối tượng làm đối thể hay chủ thể của câu.

4. Tính xác suất luật cú pháp trong văn phạm LPCFG

Luật cú pháp trong LPCFG chú trọng đến từ chính trong đoạn mà nó phân tích. Dựa trên từ chính, bộ phân tích cú pháp mở rộng sang trái và phải để xây dựng ngữ và câu. Một luật cú pháp LPCFG được biểu diễn như sau:

$PP(h) \rightarrow L_n(l_n) \ldots L_1(l_1)\ H(h)\ R_1(r_1) \ldots R_m(r_m)$,

với PP, L_i, R_i tương ứng với ngữ loại/từ loại của ngữ với từ chính h, từ l_i, từ r_i.

Ta gọi phía bên trái của luật là LHS (Left Hand Side), phía phải của luật là RHS (Right Hand Side). H là thành phần trung tâm (Head) của luật PP, được thừa hưởng từ chính h của luật PP. Ngữ chứa từ chính H được mở rộng sang hai bên bằng các cặp nhãn/từ $L_i(l_i)$ và $R_i(r_i)$. Đây là các thành phần phụ cho trung tâm H(h) để tạo thành PP. Trong trường hợp n=0 và m=0, ta có thể hiểu là con chính H không thể mở rộng sang hai bên được nữa. Trong cách viết mở rộng, ta bổ sung vào phía phải của luật PP hai thành phần L_{n+1} = Start, R_{m+1} = Stop, ứng với các ký hiệu bắt đầu và kết thúc ngữ. Ví dụ, xét luật:

C(DDT,thuyết trình)→D-C(Dt,tôi)DDN(DDT,thuyết trình)

với C là câu, DDT là động từ, DDN là động ngữ, D-C là đối tượng làm đối thể hay chủ thể của câu, Dt là đại từ.

Thành phần chính trong vế phải của luật là động ngữ. Các tham số của luật PP trong trường hợp này là:

n = 1	m = 0	PP = C
H = DDN	L_1 = D-C	L_2 = Start
h = (DDT,thuyết trình)	l_1 = (Dt,tôi)	R_1 = Stop

Xác suất luật được tính theo công thức P(RHS/LHS)=đếm(RHS)/đếm(LHS). Tuy nhiên, do ta đưa từ vựng vào luật nên xác suất luật sẽ rất nhỏ. Để giải quyết vấn đề này, chúng tôi loại bỏ các thành phần độc lập hoặc phụ thuộc rất ít vào luật. Nói cách khác, chúng tôi chia

nhỏ vế phải của mỗi luật thành các nhóm nhỏ hơn, sau đó sử dụng giả thiết độc lập có điều kiện để giảm số lượng tham số trong mô hình.

Charniak [1] đề xuất công thức tính xác suất luật theo hai bước. Đầu tiên, bộ phân tích cú pháp sinh ra luật phi ngữ cảnh, sau đó thêm vào các yếu tố từ vựng. Xác suất một luật được tính theo công thức:

$P(\text{Luật}) = P(\text{RHS/LHS}) = P(L_n(l_n)\ldots L_1(l_1)H(h)R_1(r_1)\ldots R_m(r_m)|\ PP,h)$

$= P(L_n\ldots L_1HR_1\ldots R_m|PP,h) * \prod_{i=1\ldots n} P_l(l_i \mid L_i, PP, h) * \prod_{j=1\ldots m} P_r(r_j \mid R_j, PP, h)$

Phương pháp này kế thừa được những kết quả trong cách biểu diễn PCFG, đưa thêm thành phần từ vào luật và coi xác suất mỗi thành phần con là độc lập với nhau. Phương pháp này có tính đến mức độ kết hợp giữa hai phía trong vế phải của luật. Một nhược điểm của mô hình Charniak [1] là đối với những luật không có trong tập mẫu thì $P(L_n\ldots L_1\ H\ R_1\ldots R_m|\ P,h) = 0$. Nói cách khác, mô hình này không cho phép sinh ra các luật cú pháp hay cách kết hợp các thành phần ngữ pháp mới không cho sẵn trong tập mẫu.

Mô hình của Collins [5] giải quyết được vấn đề này. Collins lấy xác suất của thành phần trung tâm làm trọng tâm để phân tích các thành phần sau.

$P(L_{n+1}(l_{n+1})\ldots L_1(l_1)H(h)R_1(r_1)\ldots R_{m+1}(r_{m+1})|\ PP,h) =$

$$P_h(H \mid PP,h) * \prod_{i=1\ldots n+1} P_l(L_i(l_i) \mid L_1(l_1)\ldots L_{i-1}(l_{i-1}), PP, h, H) *$$

$$\prod_{j=1\ldots m+1} P_r(R_j(r_j) \mid L_1(l_1)\ldots L_{n+1}(l_{n+1}), R_1(r_1)\ldots R_{j-1}(r_{j-1}), PP, h, H)$$

Trong tiếng Việt, các thành phần biên của các ngữ (như danh ngữ) thường phụ thuộc vào thành phần liền kề với nó nhiều hơn là phụ thuộc vào thành phần trung tâm. Vì vậy, đối với các phần biên của các ngữ, chúng tôi tính xác suất luật theo công thức.

$P_l(L_i(l_i)|\ H,PP,h,L_1(l_1)\ldots L_{i-1}(l_{i-1})) = P_l(L_i(l_i)|\ PP,\ L_{i-1}(l_{i-1}))$

$P_r(R_i(r_i)|\ H,PP,h,R_1(r_1)\ldots R_{i-1}(r_{i-1})) = P_r(R_i(r_i)|\ PP,R_{i-1}(r_{i-1}))$

Đa số các ngữ trong tiếng Việt không có mối liên hệ mạnh giữa hai phía thành phần chính (H) của ngữ. Mô hình của Collins [5] phù hợp với những trường hợp này. Tuy nhiên, danh ngữ trong tiếng Việt có mối liên hệ đó. Ví dụ, với danh ngữ "*Cái con mèo đen*", định tố *cái* trở thành điều kiện rất mạnh làm cho xác suất kết thúc của danh ngữ có trung tâm là *mèo* tại sau từ *mèo* rất nhỏ (nghĩa là *cái con mèo* chưa thể là một danh ngữ hoàn chỉnh mà phải có từ nào đó đi sau nó). Để giải quyết trường hợp đó, chúng tôi đề xuất dùng mô hình của Collins nhưng đưa thêm xác suất kết nối của các từ trong hai bên thành phần chính (H) trong RHS, được biểu diễn bởi P(RS|LS,H,PP), với LS và RS ứng với từ bên trái nhất và phải nhất của từ chính. Trong ví dụ trên, nếu ta lấy LS là "*cái*", RS là phần tử rỗng thì P(RS|LS,H,PP) = P(""|"*cái*", Danh từ, Danh ngữ).

Như vậy, công thức tính xác suất một luật của chúng tôi là:

$P(L_{n+1}(l_{n+1})\ldots L_1(l_1)H(h)R_1(r_1)\ldots R_{m+1}(r_{m+1})|\ PP,h) =$

$$P_h(H \mid PP,h) * \prod_{i=0\ldots n} P_l(L_{i+1}(l_{i+1}) \mid PP, L_i(l_i)) * \prod_{j=0\ldots m} P_r(R_{j+1}(r_{j+1}) \mid PP, R_j(r_j)) * P(RS|LS,H,PP) \quad (1)$$

trong đó

$P_l(L_{i+1}(l_{i+1})|\ PP,L_i(l_i)) = P(l_{i+1}t\ l_{i+1}w \mid PP,l_it\ l_iw) = P(l_{i+1}t \mid PP,l_it\ l_iw) * P(l_{i+1}w|\ PP,\ l_{i+1}t\ l_it\ l_iw)$ (2)

với t là nhãn từ loại, w là từ của một thành phần nào đó.

$P_r(R_{j+1}(r_{j+1})|\ PP,\ R_j(r_j))$ được tính theo cách tương tự.

Trong công thức (2), các giá trị xác suất $P(l_{i+1}t \mid PP,\ l_it\ l_iw)$ và $P(l_{i+1}w \mid PP,\ l_{i+1}t\ l_it\ l_iw)$ phụ thuộc vào từ cụ thể w. Trong trường hợp tập ngữ liệu dùng để huấn luyện không xuất hiện từ này, xác suất này sẽ tiến tới 0. Để giảm ảnh hưởng của từ cụ thể đối với xác suất, chúng tôi sử dụng phương pháp làm trơn nội suy tuyến tính để ước lượng xác suất bằng cách đưa vào hệ số làm trơn λ, λ_1, λ_2 ($0 \le \lambda, \lambda_1, \lambda_2 \le 1$) như sau:

$$P(l_{i+1}t \mid PP,\ l_it\ l_iw) = \lambda * P(l_{i+1}t|\ PP,\ l_it\ l_iw) + (1-\lambda) * P(l_{i+1}t \mid PP,\ l_it) \quad (3)$$

$P(l_{i+1}w \mid PP, l_{i+1}t\ l_it\ l_iw) = \lambda_1 * P(l_{i+1}w \mid PP, l_{i+1}t\ l_it\ l_iw)+(1- \lambda_1)(\lambda_2 * P(l_{i+1}w \mid PP, l_{i+1}t\ l_it)+(1- \lambda_2)* P(l_{i+1}w \mid PP, l_{i+1}t)$. (4)

Trong công thức (3), nếu $\lambda = 0.5$, từ và từ loại có vai trò ngang nhau trong việc ước lượng xác suất $P(l_{i+1}t \mid PP, l_it\ l_iw)$. Nếu $\lambda = 0$, xác suất $P(l_{i+1}t \mid PP, l_it\ l_iw)$ chỉ phụ thuộc từ loại mà không phụ thuộc từ. Tương tự với công thức (4).

Cách ước lượng này đem lại tính linh động cho mô hình. Nó cho phép biểu diễn luật phụ thuộc từ mức các ký hiệu không kết thúc đến mức từ. Khi điều kiện ở mức chi tiết hơn không đáp ứng được, hệ thống có thể điều chỉnh hệ số làm trơn để quy về mức thô hơn. Nếu một câu đầu vào nào đó sử dụng các từ khác xa so với các câu trong tập mẫu, bộ phân tích vẫn có thể đưa ra được cây phân tích cho câu đó nếu trong tập mẫu có các luật cú pháp phù hợp. Trường hợp này tương tự như trường hợp không đưa từ vựng vào luật phân tích cú pháp vì các xác suất có kết hợp từ vựng được giảm nhẹ về trường hợp không kết hợp từ vựng.

Điều chỉnh hệ số làm trơn λ

Vấn đề quan trọng ở đây là tìm giá trị λ thích hợp. Để xác định giá trị của λ , chúng tôi dùng công thức là biến thể của công thức đề xuất bởi Witten and Bell [10]:

$\lambda=f/(Hs*u+f)$

với Hs là giá trị trọng số nhằm thay đổi ảnh hưởng của u trong công thức tính. Trong phân tích cú pháp tiếng Anh dùng Wall Street Journal của Penn Treebank, hệ số này nhận giá trị trong khoảng 2→5. Chúng tôi chọn giá trị Hs = 4.

Để hiểu ý nghĩa của λ**,** chúng tôi xin lấy một ví dụ sau. Giả sử cần tìm P(A|BC). Công thức ước lượng xác suất dùng phương pháp làm trơn nội suy tuyến tính là

$\hat{P}(A|BC)= \lambda *P(A|BC)+(1- \lambda)*P(A|B)$.

Khi đó f = số lần suất hiện của các bộ 3 XBC thống kê được trong tập mẫu với mọi X.

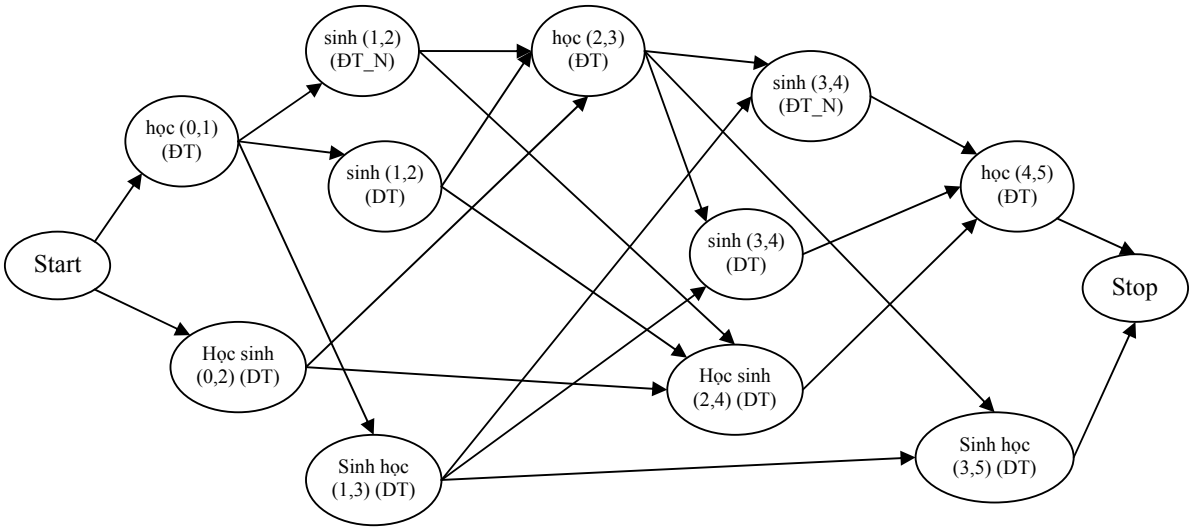
u= số lần xuất hiện bộ ba XBC *phân biệt* trong tập mẫu.

Nếu ta thống kê được f=10, u=1 và giả sử chọn Hs=1. Khi đó ta có $\lambda= 10/(1*10+1)=10/11$.

$\hat{P}(A|BC)= 10/11 *P(A|BC)+1/11*P(A|B)$.

5. Thuật toán phân tích cú pháp

Sau bước tách từ ta có các cụm âm tiết khác nhau, trong đó mỗi cụm có thể có hơn một nhãn từ loại. Ví dụ câu *"Học sinh học sinh học"* sau khi thực hiện bước tách từ ta có các từ và nhãn từ loại tương ứng được biểu diễn trong đồ thị ở Hình 4.



Hình 4 - Từ và từ loại sau khi thực hiện bước tách từ với câu "*Học sinh học sinh học*"

Trong quá trình phân tích cú pháp, câu được biểu diễn thành các trung tâm với trung tâm nhỏ nhất là từ có được sau bước tách từ. Mỗi trung tâm lưu các thông tin về vị trí của cụm từ, nhãn từ loại/ngữ loại tương ứng và nhãn ngữ loại mà nhãn cụm từ có thể phát triển lên. Ví dụ, từ "*học*" đầu tiên trong câu "*Học sinh học sinh học*" được biểu diễn thành các trung tâm sau là ((0,1), động từ, động từ) → ((0,1), động từ, động ngữ) → ((0,1), động từ, câu).

Chúng tôi sử dụng thuật toán phân tích cú pháp theo kiểu tìm kiếm sâu và duyệt theo trình tự trái-phải dưới-lên. Thuật toán được mô tả như sau:

```
Duyệt lần lượt qua các trung tâm (trái sang phải). Tại mỗi trung tâm:
I. Duyệt qua các điểm làm việc từ trái qua phải, tìm xem có thành phần nào
kết hợp được với nó không. Nếu có thành phần như vậy:
      1. Trường hợp trung tâm PH tìm được bổ sung cho trung tâm đang xét TT
            a) Kết hợp hai trung tâm PH vào TT, đánh dấu kết thúc cho các
               điểm làm việc đó và các điểm cấp trên.
               (ví dụ TT là Động ngữ, PH là Phụ tố khẳng định/phủ định, sau
               khi kết hợp, đánh thêm dấu kết thúc cho Trung tâm TT-Động
               ngữ, và Trung tâm cấp trên là TT'-Câu phát triển từ TT-Động
               ngữ và cấp trên của TT'-Câu đó)
            b) Nhảy tới các điểm làm việc mới vừa tạo ra.(TT làm trung tâm,
               Điểm phụ PH, và TT'(cấp trên của TT)làm trung tâm, điểm phụ
               là điểm vừa bị ảnh hưởng do PH kết nạp vào TT).
            c) Gọi  đệ quy thủ tục (Quay lại bước I, với mỗi điểm làm việc
               mới).
      2. Trường hợp trung tâm đang xét TT là bổ sung cho trung tâm mới PH
         (ví dụ TT là Phụ tố khẳng định/phủ định, và Trung tâm PH là Động
         ngữ)
            a) Kết hợp trung tâm TT vào PH, đánh dấu kết thúc cho các cho
               điểm làm việc đó và các điểm cấp trên.
            b) Nhảy tới vị trí mới vừa tạo ra.
            c) Gọi  đệ quy thủ tục (Quay lại bước I).
II. Nhảy sang trung tâm chưa duyệt.
```

Khác với thuật toán CYK truyền thống, thuật toán này luôn quan tâm đến từ chính trong cụm từ. Vì vậy, trong quá trình phân tích cú pháp câu đầu vào, khi cần tìm mối quan hệ giữa 2 trung tâm Y và Z (Z đứng liền sau Y), bộ phân tích cú pháp sẽ xét hai trường hợp: (i) Y là thành phần trung tâm; (ii) Z là thành phần trung tâm.

Ví dụ, cụm từ "*anh ấy*" gồm Danh từ(*anh*), Danh ngữ(*anh*) và Đại từ chỉ định(*ấy*). Để xác định ngữ loại của "*anh ấy*", thuật toán xét trường hợp:

1. Danh ngữ(Danh từ, *anh*) làm thành phần trung tâm; Đại từ chỉ định(Đại từ chỉ định, *ấy*) làm thành phần phụ. Trường hợp này có thể kết hợp được.
2. Danh ngữ(Danh từ, *anh*) làm thành phần phụ; Đại từ chỉ định(Đại từ chỉ định, *ấy*) làm thành phần trung tâm. Trường hợp này không kết hợp được.
3. Danh từ(Danh từ, *anh*) làm thành phần trung tâm; Đại từ chỉ định(Đại từ chỉ định, *ấy*) làm thành phần phụ. Trường hợp này không kết hợp được.
4. Danh từ(Danh từ, *anh*) làm thành phần phụ; Đại từ chỉ định(Đại từ chỉ định, *ấy*) làm thành phần trung tâm. Trường hợp này không kết hợp được.

6. Một số kết quả thử nghiệm

Chúng tôi đã cài đặt một phiên bản thử nghiệm cho công cụ phân tích cú pháp tiếng Việt đề xuất sử dụng ngôn ngữ lập trình Java. Trong hệ thống này, các cây cú pháp được xếp hạng theo giá trị xác suất của cây đó. Do hiện nay chúng ta chưa có một tập ngữ liệu chuẩn các câu tiếng Việt có chú giải ngữ pháp, đặc biệt là các câu được chú giải theo cách thức của LPCFG, nên chúng tôi đã xây dựng theo cách thủ công tập ngữ liệu mẫu sử dụng trong

chương trình. Nếu có một tập ngữ liệu chú giải theo cách thức của CFG, chúng tôi hoàn toàn có thể chuyển sang LPCFG qua một chương trình chuyển đổi.

Do tập mẫu còn nhỏ nên số lượng mẫu câu chương trình xử lý được còn hạn chế. Tuy nhiên, quá trình thử nghiệm chương trình đã cho thấy ảnh hưởng của từ vựng trong việc phân tích cú pháp và xử lý nhập nhằng mức từ. Đồng thời, kết quả thử nghiệm cũng chứng minh được hệ thống vẫn có thể sinh ra được cây cú pháp nếu câu không có trong tập ngữ liệu.

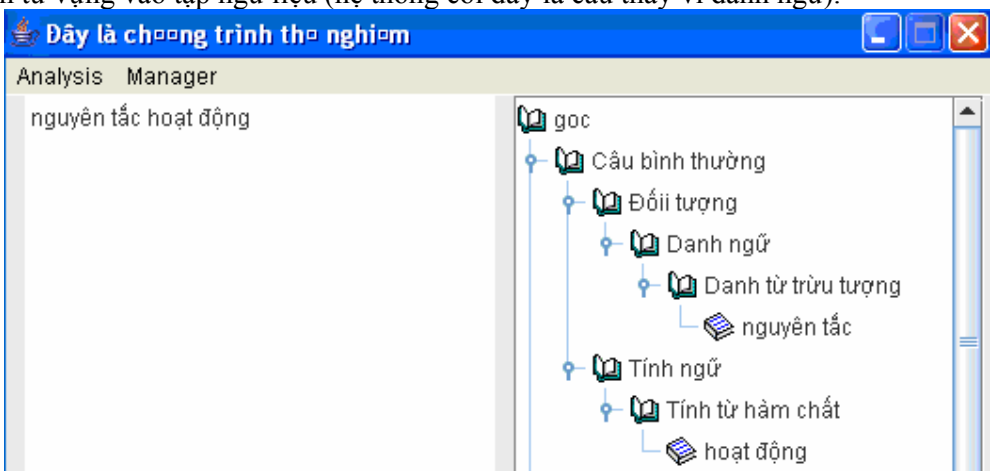
Ví dụ, với câu "*vấn đề này rất khó*" không có trong tập ngữ liệu, cây cú pháp hệ thống sinh ra là:

(a) (b)

Hình 5 – Cây cú pháp của câu "*vấn đề này rất khó*"

Trong trường hợp này, hệ thống tìm những luật có từ loại/ngữ loại tương tự trong tập ngữ liệu. Cây cú pháp (5a) có xác suất cao hơn cây (5b) và cây (5a) là cây phân tích đúng.

Với cụm từ "*nguyên tắc hoạt động*", hệ thống sẽ đưa ra phân tích sai nếu ta không đưa thông tin từ vựng vào tập ngữ liệu (hệ thống coi đây là câu thay vì danh ngữ).



Hình 6 – Cây cú pháp của câu "*nguyên tắc hoạt động*"

Nếu hệ thống sinh ra cây cú pháp như Hình 6 thì có nghĩa là tập ngữ liệu chưa cho phép xác định sự phù hợp giữa đối tượng làm chủ thể và động ngữ. Điều đó dẫn đến sự nhầm

lẫn giữa danh ngữ và câu. Để giải quyết vấn đề này cần bổ sung các ví dụ mẫu vào tập ngữ liệu liên quan đến danh từ "*nguyên tắc*". Trên cơ sở đó, hệ thống có thể tính được xác suất kết hợp giữa danh từ "*nguyên tắc*" với các từ/loại từ khác. Khi đó, xác suất kết hợp "*nguyên tắc"* và "*hoạt động*" là câu sẽ nhỏ hơn là danh ngữ.

7. Kết luận

Trong bài này, chúng tôi đề xuất một mô hình phân tích cú pháp sử dụng văn phạm LPCFG. Mô hình này cho phép xử lý nhập nhằng và xử lý các trường hợp ngữ pháp phụ thuộc từ mà các văn phạm CFG và PCFG không giải quyết được. Thuật toán phân tích cú pháp sử dụng trong hệ thống khá linh động so với dùng thuật toán phân tích cú pháp CYK truyền thống.

Trong thời gian tới, chúng tôi sẽ tiếp tục cải tiến chất lượng của bộ phân tích cú pháp bằng cách nghiên cứu cách xác định các trung tâm bắt đầu và các quy tắc di chuyển giữa các trung tâm, nghiên cứu phương pháp tự động thay đổi giá trị λ để điều chỉnh ảnh hưởng của từ và từ loại/ngữ loại trong tập luật cú pháp. Chúng tôi sẽ cho hệ thống học trên tập ngữ liệu có chú giải cú pháp đầy đủ hơn và đánh giá độ chính xác của hệ thống trên tập ngữ liệu lớn.

Tài liệu tham khảo

1. Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press/MIT Press, Menlo Park.
2. Nguyễn Tài Cẩn. 1999. Ngữ pháp tiếng Việt. NXB Đại Học Quốc Gia Hà Nội.
3. Nguyễn Tài Cẩn. 1975. Từ loại danh từ trong tiếng Việt hiện đại. NXB Khoa học xã hội Hà Nội.
4. Stanley F.Chen and Joshua Goodsman. 1998. Empirical Study of Smoothing Technique for Language Modeling. Center for Research in Computing Technology Harvard University Cambrigde, Massachusetts.
5. Micheal Collins. 2003. Head-Driven Statisticcal Models for Natural Language Parsing. MIT Artificial Intelligence Laboratory.
6. Jason Eisner and Giorgio Satta. 1999. Efficient Parsing for Bilexical Context-Free Grammars and Head Automaton Grammars. In Proceedings of the 37th Annual Meeting of the ACL.
7. Hoàng Vĩnh Sơn, "Phân tích cú pháp tiếng Việt", Đồ án tốt nghiệp đại học. Trường ĐHBK Hà Nội – 2005.
8. Lê Thanh Hương, Phạm Hồng Quang, Nguyễn Thanh Thuỷ. 2000. Một cách tiếp cận trong việc tự động phân tích cú pháp văn bản tiếng Việt. Báo Tin học và Điều khiển học, 15(4).
9. Vũ Lục. 1990. Phân tích cú pháp. Trường Đại học Bách khoa Hà Nội.
10. Ian Witten and Timothy C. Bell. 1991. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. IEEE Transactions on Information Theory, 37(4): pp.1085 – 1094.