CẢI TIẾN GIẢI THUẬT EARLEY TRONG PHÂN TÍCH CÚ PHÁP TIẾNG VIỆT

Improvement of Earley parsing for Vietnamese language[1]

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

Tóm tắt

Giải thuật Earley là một giải thuật cơ bản, được sử dụng tương đối rộng rãi trong các hệ thống phân tích cú pháp. Tuy nhiên, giải thuật này vẫn còn hạn chế như sinh ra quá nhiều luật dư thừa trong quá trình phân tích. Trong bài này, chúng tôi đề xuất các cải tiến với giải thuật Earley nhằm giúp cho quá trình phân tích cú pháp được thực hiện nhanh chóng và hiệu quả hơn. Các cải tiến này được thực hiện tại tất cả các bước của giải thuật Earley. Các đặc điểm của tiếng Việt cũng được xét tới trong việc xây dựng giải thuật này.

Từ khóa: giải thuật Earley , cải tiến, tiếng Việt

Abstract

The Earley algorithm is a fundamental algorithm, which is widely used in syntactic parsing. However, this algorithm still has limitations. For example, it produces a large amount of redundant rules during parsing process. This paper introduces some improvements with the Earley algorithm, aiming at increasing its speed and efficiency. The improvements are taken place in every steps of the original method. The characteristics of Vietnamese language are also considered in constructing this parsing algorithm.

Keywords: Earley algorithm, improve, Vietnamese language

1. Giới thiệu

Giải thuật Earley là một trong những giải thuật được sử dụng phổ biến trong việc xây dựng các hệ thống phân tích cú pháp. Giải thuật này sử dụng chiến lược phân tích kiểu trên xuống (top-down), bắt đầu với một ký hiệu không kết thúc đại diện cho câu và sử dụng các luật khai triển cho đến khi thu được câu vào. Hạn chế của cách tiếp cận này là không chú trọng nhiều đến các từ đầu vào. Vì vậy trong quá trình phân tích, giải thuật Earley sản sinh ra rất nhiều luật dư thừa.

Ngoài ra, giải thuật Earley được xây dựng cho tiếng Anh nên khi áp dụng cho tiếng Việt sẽ có hạn chế. Mỗi câu vào tiếng Anh chỉ có một cách tách từ, trong khi với tiếng Việt, mỗi câu vào có thể có nhiều cách tách từ khác nhau. Với đặc điểm đầu vào của giải thuật Earley chỉ là một câu với một cách tách, bộ phân tích cú pháp sẽ phải thực hiện lặp đi lặp lại giải thuật này cho từng trường hợp tách từ đối với tiếng Việt. Để giải quyết vấn đề này, chúng tôi nhận thấy trong các cách tách từ Việt tồn tại các cặp cách tách giống nhau ở danh sách các từ loại đầu tiên và chỉ khác nhau ở phần đuôi của chúng. Chúng tôi sẽ tận dụng đặc điểm này trong việc cải tiến giải thuật Earley cho phân tích tiếng Việt.

Những cải tiến của chúng tôi được thực hiện ở tất cả các bước của giải thuật Earley, trên cơ sở đặc điểm ngữ pháp tiếng Việt. Phần tiếp theo của bài này được tổ chức như sau. Phần 2 giới thiệu giải thuật Earley cơ bản, giúp người đọc có thể hình dung một cách khái quát về giải thuật này. Những cải tiến đối với giải thuật Earley sẽ được phân tích ở phần 3. Cuối cùng là phần kết luận, tóm tắt lại các kết quả mà chúng tôi đã đạt được.

[1] Nghiên cứu này được thực hiện trong khuôn khổ của 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.

2. Giải thuật Earley cơ bản

Giải thuật Earley cơ bản được phát biểu như sau:
Đầu vào: Văn phạm G = (N, T, S, P), trong đó:

- N: tập kí hiệu không kết thúc
- T: tập kí hiệu kết thúc
- S: kí hiệu không kết thúc bắt đầu
- P: tập luật cú pháp

Xâu vào $w = a_1 a_2 \ldots a_n$
Đầu ra: Phân tích đối với w hoặc "sai"
Kí hiệu

- α, β, γ biểu diễn xâu chứa các kí hiệu kết thúc, không kết thúc hoặc rỗng
- X, Y, Z biểu diễn các kí hiệu không kết thúc đơn
- a biểu diễn kí hiệu kết thúc

Earley sử dụng cách biểu diễn luật thông qua dấu chấm "• "
X→ α • β có nghĩa :

- Trong P có một luật sản xuất X→ α β
- α đã được phân tích
- β đang được chờ phân tích
- Khi dấu chấm "• " được chuyển ra sau β có nghĩa đây là một luật hoàn thiện. Thành phần X đã được phân tích đầy đủ, ngược lại nó là một luật chưa hoàn thiện.

Đối với mỗi từ thứ j của xâu đầu vào, bộ phân tích khởi tạo một bộ có thứ tự các trạng thái S(j). Mỗi bộ tương ứng với một cột trong bảng phân tích. Mỗi trạng thái có dạng (X → α • β, i), thành phần sau dấu phẩy xác định rằng luật này được phát sinh từ cột thứ i.

Khởi tạo

- S(0) được khởi tạo chứa ROOT → • S.
- Nếu tại bộ cuối cùng ta có luật (ROOT → S•, 0) thì có nghĩa xâu vào được phân tích thành công.

Thuật toán

Thuật toán phân tích thực hiện 3 bước: Dự đoán (Predictor), Duyệt (Scanner), và Hoàn thiện (Completer) đối với mỗi bộ S(j).

Dự đoán

Với mọi trạng thái trong S(j): (X → α • Y β, i), ta thêm trạng thái (Y → • γ, j) vào S(j) nếu có luật sản xuất Y → γ trong P.

Duyệt

Nếu a là kí hiệu kết thúc tiếp theo.
Với mọi trạng thái trong S(j): (X → α • a β, i), ta thêm trạng thái (X → α a • β, i) vào S(j+1).

Hoàn thiện

Với mọi trạng thái trong S(j): (X → γ• , i), ta tìm trong S(i) trạng thái (Y → α • X β, k), sau đó thêm (Y → α X • β, k) vào S(j).

Ở mỗi bộ S(j) phải kiểm tra xem trạng thái đã có chưa trước khi thêm vào để tránh trùng lặp.
Để minh họa cho thuật toán trên, chúng ta phân tích câu "*học sinh học sinh học*" với tập luật cú pháp sau:

S → N VP	VP → V N	N → học sinh
S → P VP	VP → V NP	N → sinh học
S → N AP	NP → N N	V → học
S → VP AP	NP → N A	V → sinh
	AP → R A	

trong đó

S	– câu	AP	– cụm tính từ	V	– động từ
VP	– cụm động từ	P	– đại từ	A	– tính từ
NP	– cụm danh từ	N	– danh từ	R	– phụ từ

Do câu trên có nhiều cách tách từ, trong khi đầu vào của giải thuật Earley chỉ là một câu với một cách tách từ nên chúng tôi minh họa giải thuật Earley với cách tách từ trong trường hợp câu được phân tích là: *học sinh*, *học*, *sinh học*.

Bảng phân tích cho cách tách này như sau

Cột 0	1	2	3
ROOT • S, 0	N học sinh•, 0	V học•, 1	N sinh học•, 2
S •N VP, 0	S N •VP, 0	VP V •N, 1	VP V N•, 1
S •P VP, 0	S N •AP, 0	VP V •NP, 1	NP N •N, 2
S •N AP, 0	VP •V N, 1	NP •N N, 2	NP N •A, 2
S •VP AP, 0	VP •V NP, 1	NP •N A, 2	S N VP•, 0
VP •V N, 0	AP •R A, 1	N •học sinh, 2	ROOT S•, 0
VP •V NP, 0	V •học, 1	N •sinh học, 2	
N •học sinh, 0			
N •sinh học, 0			
V •học, 0			

Bảng 1. Bảng minh họa giải thuật Earley

3. Những cải tiến đối với giải thuật Earley

Trong phần này chúng tôi sẽ trình bày về ý tưởng và cách cài đặt các cải tiến. Sau đó sẽ kết hợp các ý tưởng này lại với nhau một cách hợp lý để các cải tiến có thể phát huy tối đa hiệu quả.

3.1 Cải tiến bước Dự đoán

Khi nghiên cứu giải thuật Earley, cụ thể qua ví dụ ở phần 2, chúng tôi có những nhận xét ở bước Dự đoán như sau:

Nhận xét 1

Việc đưa các luật mà vế phải là từ vào trong tập luật như N → học sinh, N → sinh học, V → học sẽ khiến gia tăng nhiều luật dư thừa, đồng thời khiến bộ phân tích phải xử lý nhiều hơn. Bởi trong bước Dự đoán chúng ta sẽ phải đưa một vài luật trong số chúng vào trong cột, tiếp đó bước Quét chúng ta lại phải duyệt tất cả các luật trong cột để tìm ra một trong các luật này mà vế phải là từ được đọc tiếp theo. Do vậy *chúng tôi sẽ không đưa các luật suy ra từ vào trong tập luật phân tích câu*. Khi đọc đến một từ trong câu, căn cứ vào nhãn từ loại của từ này, bước Quét sẽ tạo ra một luật phân tích như N→học sinh• và đưa vào cột. Điều này sẽ khiến bước Quét sẽ trở nên đơn giản rất nhiều, bước Dự đoán cũng sẽ thực hiện ít hơn vì số luật phải duyệt giảm đi. Đây là một điều phù hợp với cài đặt thực tế bởi chúng ta không thể đưa luật suy ra từ vào tập luật được. Những cải tiến tiếp theo mà chúng tôi đưa ra sẽ lấy điểm xuất phát là sự cải tiến này.

Với việc bỏ đi các luật suy ra từ khỏi các dòng cuối cùng của cột (nếu có), chúng ta thu được bảng phân tích như sau:

Cột 0	1	2	3
ROOT • S, 0	N học sinh•, 0	V học•, 1	N sinh học•, 2
S •N VP, 0	S N• VP, 0	VP V •N, 1	VP V N•, 1
S •P VP, 0	S N• AP, 0	VP V •NP, 1	NP N• N, 2
S •N AP, 0	VP •V N, 1	NP •N N, 2	NP N• A, 2
S •VP AP, 0	VP •V NP, 1	NP •N A, 2	S N VP•, 0
VP •V N, 0	AP •R A, 1		ROOT S•, 0
VP •V NP, 0			

Bảng 2. Bảng phân tích Earley đã sửa đổi

Nhận xét 2

Xét cột 0, chúng ta nhận thấy do không xác định từ loại của từ "*học sinh*" trong cột 1 là N nên khi thực hiện bước Dự đoán đối với luật ROOT → • S, Earley đã tạo ra thừa hai luật là S → •P NP và S → •VP AP. Dẫn đến tiếp tục dư thừa 2 luật VP → •V N và VP → •V NP.khi thực hiện bước Dự đoán với luật S → •VP AP. Những luật dư thừa do từ loại đi sau dấu chấm "•" không phù hợp với từ loại của từ trong cột tiếp theo được gạch chân ở trong bảng 2.

Như vậy ở trong bước Dự đoán do không biết được từ loại của từ được đọc tiếp theo, Earley đã đưa tất cả những luật có thể có để bao trùm lên nhãn của từ này. Một luật dư thừa loại này lại kéo theo sự dư thừa của nhiều luật khác được tạo ra khi thực hiện bước Dự đoán đối với nó. Do vậy cần thiết phải bổ sung thêm thông tin về nhãn từ loại của từ được đọc tiếp theo vào trong bước Dự đoán. Điều này có thể thực hiện được dễ dàng bởi chúng ta đã có danh sách các từ đầu vào và các nhãn từ loại của chúng. Chúng tôi đã cải tiến như sau:

Trong bước Dự đoán tại cột i, i = 0÷n-1 với n là độ dài xâu vào, chúng tôi đưa thêm nhãn từ loại của từ thứ i. *Trước khi quyết định có thêm một luật vào trong cột phân tích hay không, chúng tôi kiểm tra xem nhãn từ loại của từ này có phù hợp với nhãn của thành phần đi sau dấu chấm hay không.* Sự phù hợp thể hiện ở điểm hai nhãn từ loại này giống nhau, hoặc thành phần sau dấu chấm là một cụm từ, có thể phân tích ra nhãn từ loại của từ này. Ví dụ, luật VP → V• NP được đưa vào cột 2 là vì thành phần NP có thể phân tích ra N (nhãn từ loại của từ "*sinh học"* ở cột tiếp theo). Cách cải tiến này giống như việc nhìn trước một bước trong các thuật toán chơi cờ của lĩnh vực trí tuệ nhân tạo. Với cải tiến này chúng ta thu được bảng số 3.

Liệu chúng ta có thể dành sự quan tâm nhiều hơn đến các từ loại xa hơn nữa không?. Câu trả lời là hoàn toàn có thể. Chúng ta có thể đưa ra những cải tiến nhìn trước k từ, k > 1. Tất nhiên sự phức tạp trong cài đặt sẽ lớn hơn.

Cột 0	1	2	3
ROOT • S, 0	N học sinh•, 0	V học•, 1	N sinh học•, 2
S •N VP, 0	S N• VP, 0	VP V •N, 1	VP V N•, 1
S •N AP, 0	VP •V N, 1	VP V •NP, 1	NP N• N, 2
	VP •V NP, 1	NP •N N, 2	NP N• A, 2
		NP •N A, 2	S N VP•, 0
			ROOT S•, 0

Bảng 3. Bảng phân tích Earley khi cải tiến việc nhìn trước một từ

Nhận xét 3

Chúng tôi nhận thấy rằng trong bảng 3 vẫn còn có những luật dư thừa. *Các luật dư thừa ở đây thể hiện ở việc chúng không có khả năng được hoàn thiện do không đủ số từ cần thiết để thực hiện các luật này*. Ví dụ như luật VP → •V NP ở trong cột 1, để luật này được hoàn thiện thì cần có 3 từ theo sau từ "*học sinh",* 1 từ có từ loại V và 2 từ để tạo ra danh ngữ NP. Trong khi đó sau từ "*học sinh"* chỉ có 2 từ. Tương tự như vậy ở cột 3, hai luật NP → N• N và NP → N• A đều dư thừa do đây đã là cột cuối cùng trong bảng nên chúng không thể được hoàn thiện. Những luật dư thừa loại này được chúng tôi gạch chân ở trên bảng 3.

Loại bỏ các luật dư thừa trên, chúng ta thu được bảng sau

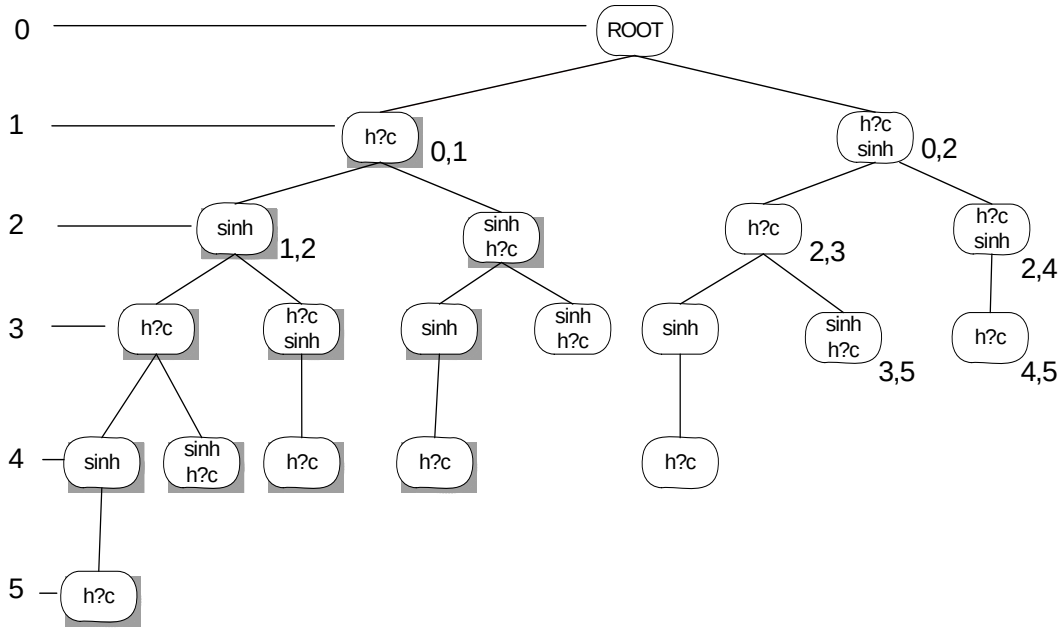
Cột 0	1	2	3
ROOT • S, 0	N học sinh•, 0	V học•, 1	N sinh học•, 2
S •N VP, 0	S N• VP, 0	VP V •N, 1	VP V N•, 1
S •N AP, 0	VP •V N, 1		S N VP•, 0
			ROOT S•, 0

Bảng 4. Bảng phân tích Earley khi cải tiến bước Dự đoán

3.2. Cải tiến dựa trên đặc điểm ngữ pháp tiếng Việt

Tiếp đến, chúng tôi nghiên cứu việc sử dụng Earley một cách hiệu quả khi phân tích câu tiếng Việt. Sự cải tiến này dựa trên hai đặc điểm của tiếng Việt: (i) một câu có thể có nhiều cách tách từ; và (ii) một từ có thể có nhiều nhãn từ loại khác nhau.

Ví dụ: câu "*học sinh học sinh học*" có đến 8 cách tách từ. Các cách tách được thể hiện thông qua cây dưới đây.



Hình 1. Các cách phân tách câu "học sinh học sinh học"

Chú thích: Khi phân tách từ trong câu, chúng tôi bổ sung thêm thông tin vào mỗi từ. Thông tin này cho biết vị trí bắt đầu và vị trí kết thúc của từ đó trong câu. Ví dụ với từ "*học sinh*" đứng đầu câu có hai cách tách là ("*học*", 0,1), ("*sinh*",1,2) và ("*học sinh*", 0,2). Trong hình vẽ trên chúng tôi hiển thị

thông tin này ở một số nút, đồng thời chúng tôi cũng hiển thị thông tin về mức của các nút trên cây với nút gốc ROOT có mức là 0.

Nhận xét 4

Vì Earley sử dụng chiến lược phân tích kiểu trên xuống (top-down) nên cây phân tích theo giải thuật Earley được mở rộng dần mỗi khi đọc một từ của câu đầu vào. Nếu tại một vị trí nào đó trong câu xảy ra sự nhập nhằng về tách từ hoặc từ đó có nhiều nhãn từ loại thì các cách tách thu được vẫn giống nhau ở các nhãn từ loại trước đó. Chính vì vậy *nếu tại một nút có nhiều nút con thì quá trình phân tích của các cách tách đi qua nút này chỉ khác nhau từ sau vị trí này*. Do đó chúng tôi đưa ra cách sử dụng hiệu quả giải thuật Earley trong phân tích cú pháp tiếng Việt đó là sau khi phân tích xong đối với một cách tách đi qua một nút rẽ nhánh và chuyển sang phân tích cách tách kế tiếp đi qua nút này, chúng tôi sẽ loại bỏ đi các cột trong bộ phân tích mà thực hiện phân tích các nút sau nút này.

Ví dụ đối với việc phân tích cho hai cách tách cuối cùng, từ ("*học sinh*",0,2) ở mức 1 có hai nút con mức 2 như trên hình vẽ. Khi khi phân tích xong cách tách thứ 7 chứa nút con ("*học*",2,3*)*, để phân tích cách tách thứ 8 tiếp theo chúng tôi sẽ loại bỏ đi các cột trong bảng sau cột phân tích từ ("*học sinh*",0,2*)* này, và chỉ cần thực hiện tiếp việc phân tích đối với từ ("*học sinh*"2,4) và ("*học*",4,5). Điều này tương ứng với việc giữ lại cột 0, cột 1 và xóa bỏ đi cột 2, cột 3 và thực hiện việc xây dựng bảng phân tích bắt đầu từ cột 2 cho cách tách thứ 8 này.
Bảng phân tích Earley của hai cách tách như sau:

Cột 0	1	2	3
ROOT • S, 0	N học sinh•, 0	V học•, 1	N sinh học•, 2
S •N VP, 0	S N• VP, 0	VP V •N, 1	VP V N•, 1
S •P VP, 0	S N• AP, 0	VP V •NP, 1	NP N• N, 2
S •N AP, 0	VP •V N, 1	NP •N N, 2	NP N• A, 2
S •VP AP, 0	VP •V NP, 1	NP •N A, 2	S N VP•, 0
VP •V N, 0	AP •R A, 1		ROOT S•, 0
VP •V NP, 0			

Cột 0	1	2	3
ROOT • S, 0	N học sinh•, 0	N học sinh•, 1	V học•, 2
S •N VP, 0	S N• VP, 0		
S •P VP, 0	S N• AP, 0		
S •N AP, 0	VP •V N, 1		
S •VP AP, 0	VP •V NP, 1		
VP •V N, 0	AP •R A, 1		
VP •V NP, 0			

Bảng 5. Bảng phân tích Earley cho cách tách thứ 7 và cách tách thứ 8

Chúng tôi xin nói rõ hơn về việc quay lui trong phân tích các cách tách và việc loại bỏ các cột tại vị trí một nút rẽ nhánh như sau.

- Việc quay lui trong phân tích thực hiện rất đơn giản bởi đi kèm với mỗi từ có thêm thông tin cho biết vị trí bắt đầu và vị trí kết thúc của từ này trong câu. Trong một cách tách từ, với hai từ kế tiếp thì từ đứng trước có vị trí kết thúc bằng vị trí bắt đầu của từ đứng liền sau. Vị trí bắt đầu của từ đầu tiên là 0, vị trí kết thúc của từ cuối cùng là độ dài của câu đầu vào. Vì vậy chúng tôi sử dụng các vòng lặp for và đệ quy để duyệt quá trình tách từ, dựng lên cây phân tách và thực hiện quay lui.
- Việc loại bỏ các cột tại vị trí nút rẽ nhánh cũng rất đơn giản bởi giải thuật Earley sử dụng bảng trong phân tích. Mỗi từ tương ứng với một cột trong bảng (trừ cột đầu tiên là khởi tạo). Đồng thời chỉ số của cột cũng chính là mức của nút thể hiện từ này trên cây. Ví dụ như nút ROOT tương ứng với cột 0 - cột khởi tạo. Nút ("*học*",0,1) và ("*học sinh*",0,2) có mức trên cây là 1 và trong phân tích cũng sẽ là cột 1 trong bảng. Do vậy tại vị trí một nút rẽ nhánh, chúng tôi chỉ cần thực hiện đơn giản là loại bỏ các cột trong bảng phân tích có chỉ số lớn hơn mức của nút này trên cây. Vì vậy tại nút rẽ nhánh ("*học sinh*",0,2) có mức là 1, chúng tôi thực hiện loại bỏ các cột với chỉ số là 2, 3 để chuyển sang xây dựng bảng phân tích cho cách tách thứ 8.

Như vậy, với việc xây dựng cây để quản lý các cách tách từ cho một câu, chúng tôi đã cài đặt được một cách đơn giản bước cải tiến giải thuật Earley xuất phát từ các đặc điểm: (i) tiếng Việt có sự nhập nhằng trong tách từ và (ii) mỗi từ có nhiều nhãn từ loại.

3.3. Cải tiến bước Hoàn thiện

Câu đầu vào ở trên có số cách tách từ khá lớn. Tuy vậy chỉ có một cách tách có chứa tập nhãn từ loại có thể phân tích được thành câu, đó là cách tách thứ 7. Do vậy một câu hỏi đặt ra là liệu có thể đánh giá một nút là không có hy vọng cho việc phân tích để khi đó chúng ta có thể dừng việc phân tích cho cho tất cả các cách phân tách chứa nút này được hay không?.

Câu hỏi này có thể trả lời được thông qua việc xem xét kĩ lại giải thuật Earley. Có thể nêu ra ý nghĩa của mỗi bước được thực hiện lần lượt tại mỗi cột như sau (ngoại trừ cột khởi tạo).

- Bước Quét: đọc từ trong câu, xác định luật phù hợp để phân tích từ này.
- Bước Hoàn thiện: 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: khai triển các kí hiệu không kết thúc, dự đoán các khả năng của nhãn từ loại của từ được đọc tiếp theo

Sau khi thực hiện tạo ra một luật để đọc từ ở bước Quét, nếu ở trong bước Hoàn thiện tiếp đó giải thuật Earley không đưa được một luật chưa hoàn thiện vào trong cột thì bước Dự đoán sẽ không bổ sung được thêm bất cứ luật nào. Do đó ở các cột tiếp theo chúng ta sẽ chỉ bổ sung thêm được các luật hoàn thiện ở bước Quét. Quá trình hợp nhất các luật được thực hiện ở bước Hoàn thiện vì vậy sẽ không thể thực hiện được, và không thể phân tích ra câu. Chúng tôi đánh giá một nút gây ra tình trạng như vậy là một nút không có hy vọng. Do vậy sau khi thực hiện bước Hoàn thiện đối với mỗi cột, chúng tôi thực hiện một hàm đánh giá hy vọng. Hàm này sẽ kiểm tra xem trong cột có luật nào chưa hoàn thiện hay không. *Nếu tất cả các luật trong cột đều đã hoàn thiện thì chứng tỏ nút này là một nút không hy vọng và việc phân tích sẽ dừng lại đối với cách tách này*. *Đồng thời bộ phân tách cũng trả về mức của nút này trên cây, để tất cả các cách tách đi qua nút này sẽ không cần thực hiện phân tích.*

Ví dụ đối với bảng 5 ở trên trong cột 2 chúng ta chỉ có luật hoàn thiện do vậy có thể dừng việc phân tích ngay từ đây mà không cần thiết phải xây dựng tiếp cột 3. Do đây đã là cách tách cuối cùng nên chúng ta chưa nhìn thấy việc không cần phân tích các cách tách đi qua nút không hy vọng này. Để minh họa điều này chúng ta bắt đầu từ việc xây dựng bảng phân tích cho cách tách đầu tiên. Bảng phân tích này như sau:

Cột 0	1	2
ROOT • S, 0	V học •, 0	V sinh•, 1
S •N VP, 0	S V• N, 0	
S •P VP, 0	S V• NP, 0	
S •N AP, 0	NP •N N, 1	
S •VP AP, 0	NP •N A, 1	
VP •V N, 0		
VP •V NP, 0		

Bảng 6. Bảng phân tích Earley cho cách tách đầu tiên

Do ở cột 2 chúng ta chỉ thu được luật hoàn thiện nên việc phân tích được dừng lại, bộ phân tích trả về chỉ số cột là 2, hay cũng là mức của từ ("*sinh*",1,2). Các cách phân tích đi qua nút này sẽ không cần thực hiện phân tích. Do vậy cách tách thứ 2 và 3 sẽ được dừng lại không cần phân tích.

Cải tiến bộ phân tích Earley trả về mức của một nút (hay chỉ số cột trong phân tích) để thực hiện cắt bỏ cách phân tích có một sự tương đồng với giải thuật cắt tỉa Anpha-Beta được sử dụng trong việc giảm không gian tìm kiếm trong các giải thuật chơi cờ.

3.4. Cải tiến bước Quét

Sự cải tiến này xuất phát từ giải thuật Earley mở rộng mà chúng tôi trình bày ở [5]. Trong giải thuật mở rộng này, luật là sự kết hợp giữa biểu thức luật CFG (Context-free grammar) và cấu trúc biểu diễn từ thông qua ma trận giá trị thuộc tính (attribute-value-matrix, AVM) trong văn phạm cấu trúc đoạn hướng trung tâm (Head-Driven Phrase Structure Grammar – HPSG).

Từ điển mà chúng tôi sử dụng trong phân tích là một từ điển dựa trên ý nghĩa về từ.Ví dụ từ "*ăn*" có 13 ý nghĩa khác nhau. Đi kèm với mỗi ý nghĩa là một tập các thuộc tính mô tả về từ. Chúng tôi sử dụng ma trận AVM để lưu trữ những thuộc tính cần thiết về từ được sử dụng trong phân tích.

Một số ý nghĩa của từ "*ăn*" như sau

STT	Ý nghĩa	Ví dụ
1	tự cho vào cơ thể thức nuôi sống	ăn cơm, lợn ăn cám...
2	ăn uống nhân dịp gì	ăn cưới, về quê ăn Tết...
3	máy móc phương tiện vận tải tiếp nhận nhiên liệu để hoạt động	Loại xe này rất ăn xăng
4	đơn vị tiền tệ, có thể đổi ngang giá	Một đô la ăn 16000 đồng Việt Nam

Bảng 7. Một số nghĩa của từ ăn

Nhận xét 5

Trong giải thuật Earley không cho phép sự trùng lặp giữa các luật trong một cột phân tích. Tuy nhiên *khi mở rộng Earley gồm hai thành phần như trên, các luật có thể giống nhau về biểu thức luật, nhưng khác nhau về ma trận AVM hoặc ngược lại thì luật mở rộng này vẫn là khác nhau, do vậy chúng vẫn thỏa mãn quy định trong giải thuật Earley*. Tại bước Quét chúng tôi không chỉ tạo ra một luật như trong giải thuật Earley truyền thống, mà có thể tạo ra rất nhiều luật và có biểu thức luật có thể hoàn toàn giống nhau. Sự giống nhau về biểu thức luật sẽ tạo ra những sự giống nhau tiếp theo khi thực hiện bước Hoàn thiện, tất nhiên chúng vẫn khác nhau về ma trận AVM đi kèm. Và hiệu quả sẽ thu được ở bước Dự đoán. Ở bước Dự đoán, ma trận AVM của các luật đều có giá trị mặc định do bước Dự đoán chưa phân tích được ra từ nào. Do vậy tất cả các luật giống nhau về biểu thức luật giờ chỉ có thể tạo ra chung một tập các luật. Sự vận dụng linh hoạt nguyên tắc này đối với giải thuật Earley trong hai bước Quét và Dự đoán đã thu gọn số lượng luật đi đáng kể.

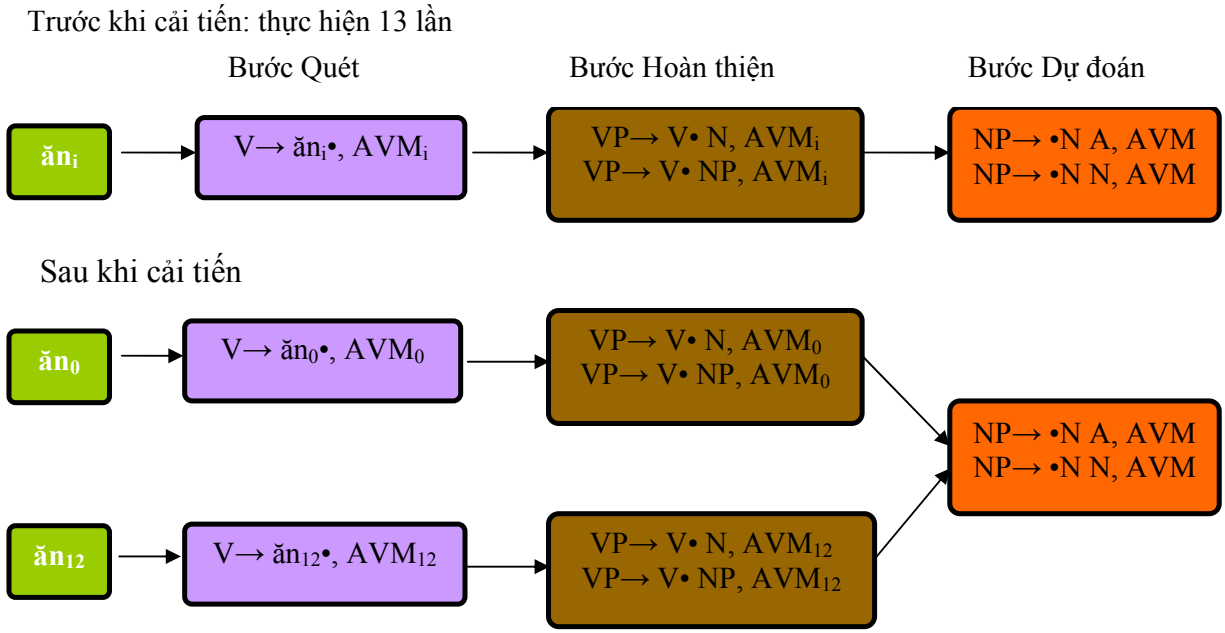
Ví dụ: Giả sử tập luật của chúng ta là

VP → V NP
NP → N A
NP → N N

Chúng tôi đang phân tích từ "*ăn*", từ "*ăn*" trong tiếng Việt có 13 nghĩa ứng với 13 ma trận AVM khác nhau nhưng đều có chung nhãn từ loại là V.

Trong bước Quét, chúng ta có 13 luật với cùng biểu diễn V → ăn• nhưng đi kèm với ma trận AVM khác nhau. Giả sử trong bước Hoàn thiện, chúng ta chỉ thu được 13 luật VP → V • NP. Khi thực hiện bước Dự đoán, mỗi luật VP → V • NP sẽ tạo ra thêm 2 luật là NP → •N N và NP → •N A. Tuy nhiên do ma trận AVM của hai luật này đều là rỗng, do vậy 13 luật VP → V • NP sẽ chỉ tạo ra được 2 luật thay vì 26 luật (13*2).

Hình minh họa cho cải tiến này



Hình 2. Minh họa cải tiến Earley ở bước Quét

3.5. Kết hợp các cải tiến trong giải thuật Earley

Trong nội dung ở trên chúng tôi đã trình bày các cải tiến. Việc kết hợp các cải tiến lại với nhau không chỉ đơn giản là việc gộp rời rạc các cải tiến lại mà cần tìm hiểu xem liệu các cải tiến này có mối liên hệ nào lẫn nhau hay không, và liệu khi áp dụng đồng thời như vậy việc phân tích có bị sai hay không.

Chúng tôi nhận thấy có một sự liên hệ giữa cải tiến ở bước Dự đoán và cải tiến sử dụng lại dựa trên đặc điểm ngữ pháp tiếng Việt. Sự liên hệ này nếu không được giải quyết sẽ dẫn đến việc bộ phân tích không phân tích ra câu. Ví dụ tại nút con đầu tiên của nút ROOT, do "*học*" là một động từ nên trong cột 0, tại bước Dự đoán chúng ta chỉ đưa vào các luật mà có thể phân tích ra động từ. Như vậy khi quay lui để chuyển sang việc phân tích cho nút con thứ hai là "*học sinh*", nếu chỉ tiến hành loại bỏ các cột có chỉ số lớn hơn 0 (giữ lại duy nhất cột khởi tạo), bộ phân tích sẽ thực hiện sai do từ "*học sinh*" là một danh từ. Do đó mỗi khi quay lui tại một nút rẽ nhánh để chuyển sang một cách phân tách khác, chúng ta phải thực hiện lại bước Dự đoán để bổ sung vào cột các luật phù hợp với nhãn từ loại của từ được đọc tiếp theo.

4. Kết luận

Như vậy chúng tôi đã đưa ra những cải tiến cho giải thuật Earley ở cả 3 bước Dự đoán, Quét và Hoàn thiện. Sự cải tiến ở bước Dự đoán giúp nhìn nhận trước khả năng phân tích, hạn chế được các luật dư thừa. Sự cải tiến ở bước Quét đem lại hiệu quả tích cực đối với các bộ phân tích sử dụng tập luật có chứa ngữ nghĩa của từ. Còn sự cải tiến ở bước Hoàn thiện có tác dụng cắt tỉa không gian tìm kiếm trong các trường hợp phân tích sai do cách tách từ không tạo ra câu hay câu đầu vào sai. Đồng thời chúng tôi cũng đưa ra được cải tiến sử dụng lại dựa trên đặc điểm sự nhập nhằng trong tách từ và từ có thể có nhiều nhãn từ loại trong ngữ pháp tiếng Việt. Sự kết hợp hiệu quả các cải tiến trên đã cho phép chúng tôi thực thi chương trình với tốc độ nhanh hơn..

Tài liệu tham khảo

[1] Diệp Quang Ban. Ngữ pháp tiếng Việt, NXB Giáo Dục. 1998

[2] J. Earley, "An efficient context-free parsing algorithm", Communications of the ACM, 1970.

[3] Nguyễn Gia Định, Trần Thanh Lương, Lê Việt Mẫn. Một số cải tiến giải thuật Earley cho việc phân tích cú pháp trong xử lý ngôn ngữ tự nhiên, Tạp chí Khoa học Đại học Huế 6/2004

[4] Daniel Jurafsky, James H. Martin. Speech and language processing, Prentice Hall. 2000.

[5] Đỗ Bá Lâm, Lê Thanh Hương, Xây dựng hệ thống phân tích cú pháp tiếng Việt sử dụng văn phạm HPSG, ICT RDA ’08.

[6] Pollard, C.J., Sag, I. Head-Driven Phrase Structure Grammar, CSLI Publications/Cambridge University Press. 1994.