CÁI TIẾN MỘT SỐ GIẢI THUẬT PHÂN TÍCH CỦA PHÁP TRONG XỬ LÝ NGÔN NGỮ TỰ NHẤN

PHAN THỊ TƯƠI

Abstract. Earley and Chart algorithms are often used to parse natural language. However, these algorithms are limited when they are used to work with large grammar. This paper presents some improvements for Earley and top-down chart algorithms in natural language processing.

Tóm tắt. Để phân tích cú pháp cho ngôn ngữ tự nhiên, người ta thường dùng các giải thuật như Earley và biểu đồ. Tuy nhiên khi xử lý các văn phán lên thì các giải thuật trên đều bị hạn chế. Bài báo này sẽ trình bày một số cải tiến cho giải thuật Earley và biểu đồ từ trên xuống trong xử lý ngôn ngữ tự nhiên.

1. MỞ DÀU

Vài trào của phần tích cú pháp trong xử lý ngôn ngữ tự nhiên là vở cùng quan trọng. Tuy nhiên không phải tất cả các giải thuật phân tích cú pháp cho ngôn ngữ lập trình đều có thể áp dụng cho ngôn ngữ tự nhiên, bởi vì văn phạm của ngôn ngữ tự nhiên là không tương minh. Ngày càng ta dùng các giải thuật như Tomita [10], Earley [1] và Chart [3], là những giải thuật phân tích cú pháp cho văn phạm không tương minh thì cũng gặp nhiều khó khăn khi áp dụng chúng cho xử lý ngôn ngữ tự nhiên, bởi vì muốn phân tích một chuỗi nhập vào là câu hoặc đoạn câu của ngôn ngữ tự nhiên thì bỏ phân tích bước phải kiểm tra từ vựng cho hàng chúng, hàng từ vựng từ loại khác nhau (từ loại được hiểu như các token trong ngôn ngữ lập trình), đều do sẽ dẫn đến sự bùng nổ của. Trong bài báo này chúng tôi trình bày việc cải tiến giải thuật Earley và giải thuật biểu đồ từ trên xuống (top-down chart parsing) cho phân tích cú pháp của ngôn ngữ tự nhiên.

2. MỘT SỐ GIẢI THUẬT EARLEY VÀ LR(k) CÁI TIẾN


3. SƠ LƯỢC VĂN PHÁM TIẾNG VIỆT

Trước khi trình bày việc cải tiến giải thuật Earley và biểu đồ từ trên xuống tôi sẽ nêu sơ lược về văn phạm tiếng Việt. Khi nghiên cứu tiếng Việt chúng ta phải nghiên cứu hai thành phần là từ pháp và cú pháp. Từ pháp chuyển nghiên cứu các biểu hiện của từ và đặc tính ngữ pháp của các loại từ cũng như sự cấu tạo của từ, còn cú pháp nghiên cứu cách cấu thành các từ và cấu từ các nhóm từ theo một quy tắc nhất định.

3.1. Từ loại

Trong tiếng Việt được chia thành các loại: danh từ, động từ, thời vở từ, số từ, tính từ, đại
từ, phó tử, giới tử, liên tử, trợ tử và thân tử. Vời mỗi loại tử trên lại có thể được chia nhở hơn, mang nghĩa khác nhau tùy thuộc vào ngữ cảnh. Ví dụ: danh từ có 27 loại, được ký hiệu bắt đầu bằng chữ N, như: N20, N51,... Đồng tử có 22 loại bắt đầu bằng V: V45, V20,... Tính từ có 16 loại được bắt đầu bằng A. Chỉ tiết về chia tử loại được trình bày ở [5, 6].

3.2. Từ tổ
Hai hay nhiều hơn hai thực thể ở trong câu có quan hệ với nhau về ý nghĩa và cấu pháp thì được gọi là tổ tổ. Trong mỗi tổ tổ bao giờ cũng có từ trọng tâm. Thử thuốc từ trọng tâm mà ta phân tử tổ tổ thành tổ tổ gọn tổ, từ tổ danh từ, từ tổ số từ, từ tổ thời tổ vị từ. Mỗi loại tổ tổ lại có từ một đến nhiều dạng khác nhau. Từ câu trực từ tổ ta có thể xây dựng văn phạm đề kiến.trả câu từ cấu trúc của chúng.

4. CẢI THIỂN PHƯƠNG PHÁP EARLEY CHO XỬ LÝ NGỒN NGỮ TỰ NHIÊN

Khi bộ phân tích cú pháp muốn xác định xem chuỗi nhập là câu hoặc đoạn câu của ngôn ngữ tự nhiên có đúng cú pháp không thì nó phải phân tích các chuỗi từ loại tương ứng của câu, đoạn câu. Trong trường hợp xảy ra nhất bộ phân tích phải kiểm tra tất cả các chuỗi mới có thể kết luận về cú pháp của câu nhập.

Ví dụ 1. Câu cần phân tích: Cánh bô công chức quy định tại pháp lệnh này là công dân Việt Nam.
Câu trên được bộ phân tích tự vụ quy định và xuất ra các chuỗi từ loại tương ứng:
N20 N20 N61 L22 N51 D71 H10 N20 N13
N20 N20 V48 L22 N51 D71 H10 N20 N13
Ở đây chỉ có từ quy định là có hai từ loại: N61 và V48. Do đó việc phân tích cú pháp cho câu trên không có gì khó khăn.

Ví dụ 2. Đoạn câu cần phân tích: trong biên chế và hướng lượng từ ngân sách.
Có 12 chuỗi từ loại được xuất ra:
A11 N22 L10 V43 N23 F10 N23 N50
A11 N22 L10 V43 N23 F11 N23 N50
A11 V40 L10 V43 N23 F10 N23 N50
A11 V40 L10 V43 N23 F11 N23 N50
F10 N22 L10 V43 N23 F10 N23 N50
F10 N22 L10 V43 N23 F11 N23 N50
F10 V40 L10 V43 N23 F10 N23 N50
F10 V40 L10 V43 N23 F11 N23 N50
F11 N22 L10 V43 N23 F10 N23 N50
F11 N22 L10 V43 N23 F11 N23 N50
F10 V40 L10 V43 N23 F10 N23 N50
F10 V40 L10 V43 N23 F11 N23 N50


Hình 1. Kiểm tra cú pháp cho câu có nhiều chuỗi từ loại bằng giải thuật Earley.
Trong thực tế các chuỗi từ loại đều có các chuỗi con giống nhau, chúng hạn trong các ví dụ 1 và 2. Lời dụng dịch điểm này, chúng tôi đã cải thiện giải thuật Earley như sau: nếu bộ phân tích thất bại khi dùng kiểm tra một chuỗi, thì nó sẽ so trỏ các chuỗi con lại với đoạn vừa kiểm tra thành công và sẽ tiếp tục quy trình phân tích ở vị trí của một chuỗi khác có chuỗi con dài nhất trong vòng đó phân tích. Quá trình này được lập lại cho tới khi bộ phân tích “đi” hết một chuỗi nào đó. Lức đó cấu nhập được xác nhận là đúng cấu pháp. Ngược lại khi đi đến chuỗi cuối cùng mà vẫn không phân tích thành công thì bộ phân tích sẽ kết luận rằng cấu nhập vào không đúng cấu pháp. Quá trình phân tích cấu pháp của giải thuật Earley cái thiện được minh họa ở hình 2.

Hình 2. Quá trình phân tích của giải thuật Earley cải thiện

Giải thuật Earley cải thiện:

Nhập: Vận phán phi ngữ cảnh $G = (N, \Sigma, P, S)$ và các chuỗi từ loại của cấu can phân tích cấu pháp:

$W_1, W_2, ..., W_m \in \Sigma^*$. Kích thước các chuỗi bằng kích thước cấu nhập.

Xuất: Đánh sách các tập thực thể $I_0, I_1, ..., I_n$.

Phương pháp: Thực hiện phân tích trên $W_1$:

Trước tiên ta tạo $I_0$:

1. Nếu $S \rightarrow \alpha$ là luật sinh thuộc $P$, thì thêm $[S \rightarrow \bullet \alpha, 0]$ vào $I_0$. Tiếp tục bước 2 và 3 cho tới khi không có thêm thực thể mới chưa cho vào $I_0$.

2. Nếu $[B \rightarrow \gamma, j] \in I_0$ thì thêm vào $I_0$ thực thể $[A \rightarrow \alpha B \bullet \beta, 0]$ cho tất cả các thực thể có dạng $[A \rightarrow \alpha \bullet \beta, 0]$ ở trong $I_0$.

3. Giả sử $[A \rightarrow \alpha \bullet \beta, 0]$ là thực thể ở trong $I_0$, thêm vào $I_0$ tất cả các thực thể $[B \rightarrow \gamma, 0]$, nếu tồn tại các luật sinh có dạng $B \rightarrow \gamma$ ở trong $P$ và các thực thể đó chưa có ở trong $I_0$.

Khởi động các trừ: $m_1 := 1$; $m_2 := 1$; $t := 1$;

Xây dựng $I_1$, khi đã có các $I_0, I_1, ..., I_{j-1}$.

4. Nếu một $[B \rightarrow \alpha \bullet \alpha, j]$ trong $I_{j-1}$ mà $a = a_j$ thì thêm $[B \rightarrow \alpha \bullet \beta, i]$ vào $I_j$ và thực hiện các bước 5, 6 cho tới khi không còn thêm thực thể mới nào chưa có trong $I_j$.

Ngược lại, nếu $[B \rightarrow \alpha \bullet \beta, i]$ trong $I_{j-1}$ mà $a \neq a_j$ thì thực hiện các bước sau:

a. Gán $m_2 := j - 1$;

b. So trùng chuỗi $W_{lm} = a_{m1}a_{m1+1}...a_{m2}$ với các chuỗi $W_1, W_2, ..., W_m$ để tìm một chuỗi con đa nhất trùng với chuỗi $W_m$, gọi là $W_{lm}$ thuộc chuỗi $W_h$.

c. Gopi $a_j$ là ký hiệu dừng Sağ bên phải ký hiệu cuối $W_{lm}$ của chuỗi $W_h$.

d. Gán $m_2 := j$; $m_1 := m_2$; $l := h$.

e. Quay về bước (4) để tạo $I_j$, với ký hiệu dừng được độc là $a_j$ ở vị trí $j$ của chuỗi $W_h$.

5. Giả sử $[A \rightarrow \alpha \bullet \beta, i]$ là thực thể ở trong $I_j$, kiểm tra $I_j$ cho tất cả các thực thể có dạng $[B \rightarrow \alpha \bullet \beta, k]$ với mỗi thực thể tìm thấy ta thêm $[B \rightarrow \alpha A \bullet \beta, k]$ cho $I_j$.

6. Giả sử $[A \rightarrow \alpha \bullet \beta, i]$ là thực thể trong $I_j$, với tất cả các luật sinh có dạng $B \rightarrow \gamma$ trong $P$, ta thêm $[B \rightarrow \alpha \bullet \gamma, j]$ vào $I_j$. Chung ta nhận thấy rằng thực thể nào có ký hiệu kết thúc ở bên
phái của đầu • sẽ không phải là thực thể mới trong bước 2, 3, 5, 6. Giải thuật được mô tả ở hình 2, giải thuật này được thực hiện bằng chương trình trong [5].

Lưu ý: các động in nghiện là phần cải thiện giải thuật Earley của chúng tôi.

5. CẢI THIỂN PHƯƠNG PHÁP BIẾU ĐO TRONG XỬ LÝ NGÓN NGỮ TỰ NHIÊN

Phần này trình bày việc cải thiện giải thuật phân tích cú pháp biểu đồ từ trên xuống. Giải thuật này hiện quả hơn so với các giải thuật hiện nay cho các văn phạm hợp lý. Nó kết hợp ưu điểm của hai giải thuật phân tích cú pháp biểu đồ từ dưới lên (bottom up chart parsing) và phân tích từ trên xuống (top down parsing). Phương pháp này không bị quay lui và không xét các từ loại mà chúng không thể dùng để tạo ra câu đúng.

Để thuận lợi cho việc giải thích ý tưởng cải thiện giải thuật biểu đồ, chúng tôi tóm tắt giải thuật biểu đồ [9] từ trên xuống như sau.

5.1. Giải thuật trình bày cung tự trên xuống

Để thêm một cung $S \rightarrow C_1 \cdots C_n$ vào cuối của vị trí $j$ thì với mỗi lượt sinh của văn phạm có dạng $C_i \rightarrow X_1 \cdots X_k$ ta thêm vào một cung mới $C_i \rightarrow \bullet X_1 \cdots X_k$ một cách dễ quy từ vị trí $j$ đến $j$.

5.2. Giải thuật phân tích số độ từ trên xuống

Bắt đầu với mỗi lượt sinh của văn phạm có dạng $S \rightarrow X_1 \cdots X_k$, thêm vào một cung có tên $S \rightarrow \bullet X_1 X_2 \cdots X_k$ bằng giải thuật trình bày cung. Thực hiện các bước phân tích sau đây cho mỗi phép cung ký hiệu hiện tại:
   a. Nếu bằng rỗng, hãy tìm các từ loại của ký hiệu nhập kết tiếp và thêm chúng vào bảng.
   b. Chọn một thành phần từ bảng (gọi nó là $C$).
   c. Đừng giải thuật mới rỗng, kết hợp $C$ với từng cung hoạt động trên biểu đồ. Một thành phần mới được thêm vào bảng.
   d. Với bất kỳ một cung mới nào được tạo ra ở bước c, hãy thêm chúng vào biểu đồ bằng giải thuật trình bày cung.

5.3. Giải thuật mới rỗng đề thêm thành phần $C$ vào vị trí từ $p_1$ đến $p_2$

Thực hiện các bước sau:
   a. Thêm $C$ vào số độ từ $p_1$ đến $p_2$.
   b. Với bất kỳ cung hoạt động có dạng $X \rightarrow X_1 \cdots X_n$ từ vị trí $p_0$ đến $p_1$, thêm cung mới $X_1 \rightarrow X_1 \cdots X_n$ từ $p_0$ đến $p_2$.
   c. Với bất kỳ cung hoạt động nào có dạng $X \rightarrow X_1 \cdots X_n \bullet C$ từ vị trí $p_0$ đến $p_1$, thêm một thành phần mới của $X$ từ $p_0$ đến $p_2$ trong bảng.

Mặc dù ưu điểm như vậy nhưng khi áp dụng phương pháp cú pháp cho văn phạm lớn (có từ với ngàn đến vài chục nghìn từ) thì giải thuật này cũng bị hạn chế về tốc độ. Trong xử lý ngôn ngữ tự nhiên chúng ta thường gặp các văn phạm lớn, vì vậy chúng tôi đã thêm vào một số xử lý về tốc độ để giúp cho thực hiện giải thuật nhằm khắc phục những điểm trên.

1) Sơ hóa toàn bộ các ký hiệu kết thúc và không kết thúc của văn phạm (để tiết kiệm bộ nhớ và số trình nhanh).
2) Xây dựng giải thuật tìm kiếm theo phương pháp “bấm” cho từng tập thực thể $I_j$.
3) Tính trước các thể do các ký hiệu không kết thúc tạo nên.
4) Chia nhỏ địa chỉ của các thể do $I_0, I_1, \ldots, I_n$ (được tạo từ các hành vi chuyển dịch đầu • về bên phải, qua các ký hiệu văn phạm). Nhờ các tác vụ trên, nền bộ phân tích không cần lưu trữ một lần nữa dữ liệu đã được tính toán trước đó, mà chỉ cần lưu địa chỉ của tập thực thể tương ứng, do đó tiết kiệm được bộ nhớ đáng kể và bảo đảm tốc độ phân tích. Ví dụ 3 sẽ minh họa quá trình phân tích một câu bằng giải thuật biểu đồ từ trên xuống.

Vì dự 3. Thực hiện phân tích cú pháp cho câu: the large can can hold the water
Cho vấn đề với tập luật sinh:

(1) \( S \rightarrow NP \ VP \)
(2) \( NP \rightarrow \text{art adj n} \)
(3) \( NP \rightarrow \text{art n} \)
(4) \( NP \rightarrow \text{adj n} \)
(5) \( VP \rightarrow \text{aux VP} \)
(6) \( VP \rightarrow v \ NP \)

Hình 3 là các bước phân tích với các \( I_1, I_2, \ldots, I_8 \) ở bên cạnh các cung để biểu hiện cho địa chỉ của các tập thực thể.

<table>
<thead>
<tr>
<th>NP (rule 3)</th>
<th>VP (rule 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>v</td>
<td>N</td>
</tr>
<tr>
<td>aux</td>
<td>V</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S (rule 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP VP (rule 1) (accept)</td>
</tr>
</tbody>
</table>

Hình 5. Qua trình phân tích câu: the large can can hold the water

bằng giải thuật biểu đồ xuong cải tiến

Thực hiện theo bước b, ta có:
- Tập \( S \rightarrow \bullet \ NP \ VP \)
- Tập \( NP \rightarrow \bullet \ \text{adj n} \)
- Tập \( NP \rightarrow \bullet \ \text{art n} \)
- Tập \( VP \rightarrow \bullet \ \text{aux VP} \)
- Tập \( VP \rightarrow \bullet \ v \ NP \)

Theo giải thuật chúng ta tính được các tập thực thể:

\( I_0 (S) \)
\( I_1 (\text{art}) \)
\( I_2 (\text{adj}) \)
\( I_3 (n) \)
\( I_6 (\text{aux}) \)
\( I_7 (\text{NP}) \)
\( I_8 (VP) \)

Lưu ý: Ví dụ ở trên thực hiện với tập thứ 7, triển khai cung NP thì máy tính không tính lại các cung \( NP \rightarrow \text{nadj n} \) và \( NP \rightarrow \text{art adj n} \) mà nó chỉ cần chỉ đến địa chỉ của \( I_1 \). Trong thực, như vậy cho các trường hợp khác.

6. KẾT LUẬN

Qua thực tế khi thực hiện các đề tài nghiên cứu trong lĩnh vực xử lý ngôn ngữ tự nhiên, chúng tôi thấy có thể áp dụng các giải thuật đã có, song phải cải thiện để các giải thuật này đáp ứng được các yêu cầu để ra khả đi sâu với thực nghiệm, những cải thiện này góp phần làm cho các giải thuật
khá nói tiếng như Earley, LR, biểu đồ càng hoàn thiện và hiệu quả hơn hơn trong xử lý ngôn ngữ tự nhiên.

TÀI LIỆU THAM KHẢO

[8] Philippe McLean Q., R. Nigel Horspool, A faster Earley parser, email: pmclean@cs.unic.ca, nigelh@cs.unic.ca.

Trường Đại học Bách khoa - ĐHQG TPHCM

Nhận bài ngày 5 - 1 - 2002