Thuật toán CKY – CKY Parsing Algorithm simulation

1. Giới thiệu thuật toán CKY

CKY (Coke- Kasami – Younger) là một thuật toán cải tiến của thuật toán phân tích cú pháp Bottom-Up (Button-Up Parsing là một chiến lượt phân tích tích cú pháp bắt đầu từ các từ trong các chuỗi đầu vào và xây dựng các thành tố cú pháp).

CKY có thể tránh được những cách phân tích cú pháp không hợp lý so với thuật toán Buttom-Up thông thường. Do CKY sử dụng một hình thức văn phạm đặc biệt được gọi là Chomsky Normal Form (CNF). Các giải pháp trung gian của thuật được lưu trữ và chỉ triển khai những giải pháp trung gian nào có khả năng đóng góp vào việc phân tích đầy đủ cấu trúc cú pháp câu.

2. Mã giả thuật toán CKY

– CKY Algorithm (Recognition)

```
function CKY-Parse (words, grammar) returns table

    for j ← 1 to length(words) do: (loop over columns)

        table[j-1,j] ← {A|A → words[j] ∈ grammar} (add POS)

        for i ← j-2 downto 0 do: (loop over rows, backwards)

            for k ← i+1 to j-1 do: (loop over contents of cell)

            table[i,j] ← table[i,j] ∪

                    {A|A → B  C ∈ grammar,

                    B ∈ table[i,k]

                    C ∈ table[k,j] }
```

– CKY Parsing:

```
function CKY-Parse (words, grammar) returns parses

    for j ← 1 to length(words) do: (loop over columns)

        table[j-1,j] ← for all {A|A → words[j] ∈ grammar} (add all POS)

        for i ← j-2 downto 0 do: (loop over rows, backwards)

            for k ← i+1 to j-1 do: (loop over contents of cell)

                for all {A|A → B C}: (all productions)

                back[i,j,A] ← { k,B,C } (add back pointer)

    return buildtree(back[1, length(words,S]), table[1,LENGTH(words),S]
    (follow back pointer)
```

3. Các vấn đề của CKY

Tính hiệu quả:

– CKY có thể được thực hiện với thời gian: $O(n^3)$, trong đó n=số từ của câu.

– Sự phức tạp của các vòng lặp trong.

– Nhiều quy tắc hơn, thì ít hiệu quả hơn, nhưng điều này làm tăng một tỉ lệ hằng $L = r^2$ với r là số lượng của các biến (non-terminals).

Ngữ pháp đòi hỏi:

– Các thuật toán cơ bản đòi hỏi một ngữ pháp nhị phân: ngữ pháp CNF (Chomsky Normal Form).

– Thuật toán cơ bản có thể được mở rộng để phân tích cho CFGs tùy ý.

– Tuy nhiên, việc chuyển đổi thành ngữ pháp CNF dễ dàng và hiệu quả hơn so với phân tích với ngữ pháp tùy ý.

– Thuật toán Earley cho phép phân tích CFGs tùy ý.

4. Ngữ pháp Chomsky Normal Form

Một ngữ pháp phi ngữ cảnh mà RHS của mỗi quy tắc đưa ra là: 2 non-terminals hoặc 1 terminal. Chúng có thể là:

– Không quy tắc lẫn lộn (NP -> the NN).

– Không có dạng NP -> NNP, ngoại trừ dạng NN -> dog.

– Vế phải không có nhiều hơn 2 non-terminals như: VP -> VBZ NP PP.

Bất kỳ CNG nào cũng có thể biến đổi thành một ngữ pháp tương đương yếu trong CNF tức là chúng tạo ra cũng một bộ chuỗi (các câu).

Quy ước đặt tên, ký hiệu trong văn phạm CNF:

- Sử dụng ký hiệu mới (nhị phân):
  - X1, X2, . . . , Y3
  - S →NP VP PUNC trở thành :
    - S →NP X1,
    - X1→VP PUNC
- Xóa một ký hiệu:
  - SBAR → S , S → NP VP trở thành
  - SBAR → NP VP

Thuật toán CNF cải tiến:

1. Loại bỏ các unit-productions:

   while there is a unit-production A → B,

   Remove A → B.

   For each B → u, add A → u.

2. Loại bỏ các terminals trong quy tắc lẫn lộn

   For each production A → B1 B2 … Bk, containing a terminal x

   Add new non-terminal/production X1 → x (unless it has already been added)

   Replace every Bi = x with X1

3. Loại bỏ các quy tăc với nhiều hơn 2 nonterminals trên RHS (nhị phân)

   For each rule p of form A → B1 B2…Bk

   replace p with

   A → B1 X1,

   X1 → B2 X2,

   X2 → B3 X3, …,

   X(k − 2) → Bk−1 Bk (Xi là các biến mới)

5. Phân tích một câu đơn giản theo văn phạm CNF

Câu: **•0 the •1 chef •2 eats •3 fish •4 with •5 the •6 chopsticks •7**

Phân tích theo văn phạm CNF là:

S → NP VBZ

S → NP VP

VP → VP PP

VP → VBZ NP

VP → VBZ PP

VP → VBZ NNS

VP → VBZ VP

VP → VBP PP

NP → DT NN

NP → DT NNS

PP → IN NP

DT → the

NN → chef

NNS → fish

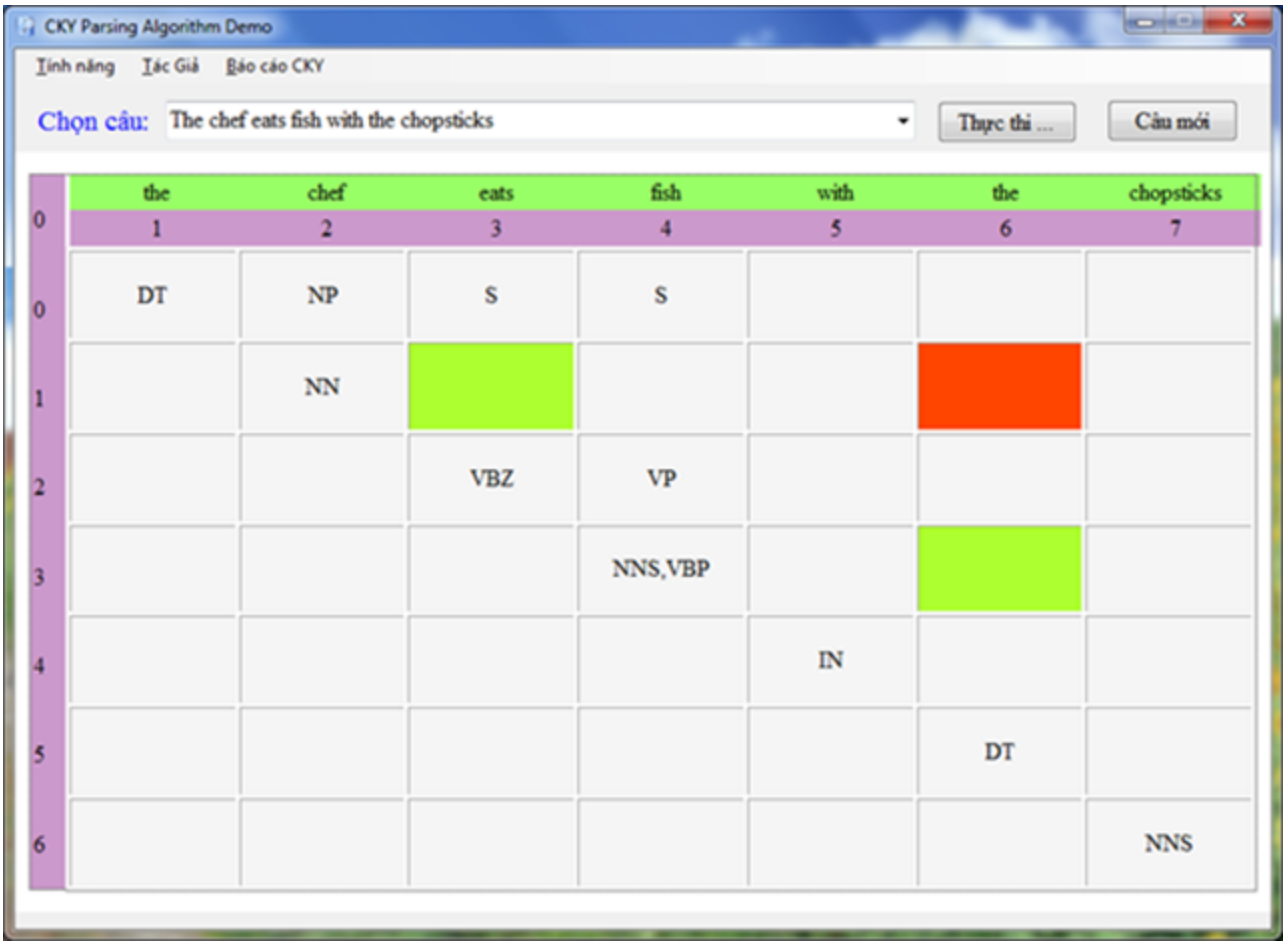
NNS → chopsticks

VBP → fish

VBZ → eats

IN → with

6. Demo CKY algorithm


CKY Parsing Algorithm Demo
Tính năng
Tác Giả
Báo cáo CKY
Chọn câu:
The chef eats fish with the chopsticks
Thực thi ...
Câu mới
the
chef
eats
fish
with
the
chopsticks
1
2
3
4
5
6
7
0
0
DT
NP
S
S
1
NN
2
VBZ
VP
3
NNS,VBP
4
IN
5
DT
6
NNS

BÀI TẬP

1. Cho văn phạm G:

   S → AB | XB

   T → AB | XB X → AT

   A → a

   B → b

   Chỉ ra quá trình thực hiện thuật toán CYK với w = **aaabbb**

2. Cho văn phạm G:

   S → AA | AS | b

   A → SA | AS | a

   Chỉ ra quá trình thực hiện thuật toán CYK với w = **abaab**

3. Sử dụng thuật toán CYK để chỉ ra cây phân tích cho chuỗi **(5+7)*3** thuộc văn phạm G

   E → E + T | T

   T → T * F | F

   F → (E) | số

4. Chỉ ra cây phân tích của chuỗi **true and not false** sinh bởi thuật toán CYK với tập luật văn phạm G E → E and T | T

   T → T or F | F

   F → not F | (E) | true | false

Lưu ý: Cài đặt thuật toán CKY và kiểm tra kết quả thực hiện các câu trên