The CYK Parsing Algorithm

Lecture 19
Section 6.3

Robb T. Koether

Hampden-Sydney College

Fri, Oct 7, 2016

Outline

The Membership Problem for CFGs

Definition (The Membership Problem for CFGs)

Given a context-free language L and a word w, is w in $L(G)$?

- We may assume that G is in Chomsky Normal Form.
- Let n be the length of the word w.

The Membership Problem for CFGs

- An unsophisticated, brute-force algorithm will examine every possible derivation of length $2n - 1$ (actually $n - 1$).
- If none of these produces w, then $w \notin L(G)$.
- There problem is, the number of such derivations may grow exponentially in n. (Why?)
- We need a better algorithm.

Parsing

Definition (To parse)

Given a grammar G and a string W, to parse w is to find a derivation of w in G.

- There are many different parsing algorithms for context-free grammars.
- As usual, the more powerful parsing algorithms are more complex.
- We will examine a relatively simple algorithm developed by Cocke, Younger, and Kasami, known as the CYK algorithm, which runs in polynomial time.

Outline

The CYK Parsing Algorithm

- The strategy of the CYK algorithm is to begin with all one-symbol substrings of w that are derivable from G.
- These come directly from the rules of the form $A \to a$.
- Then proceed inductively.

The CYK Parsing Algorithm

- Let $w = a_1 a_2 a_3 \cdots a_n$.
- Let $w_{ij} = a_i \cdots a_j$, for all $1 \leq i \leq j \leq n$.
- Define sets
$$V_{ij} = \{A \in V \mid A \stackrel{*}{\Rightarrow} w_{ij}\}.$$
- For example,
  - $V_{11} = \{A \in V \mid A \stackrel{*}{\Rightarrow} a_1\}$.
  - $V_{35} = \{A \in V \mid A \stackrel{*}{\Rightarrow} a_3 a_4 a_5\}$.
  - $V_{1n} = \{A \in V \mid A \stackrel{*}{\Rightarrow} w\}$.

The CYK Parsing Algorithm

The Basic Step

- Initialize each V_{ii} to the set of all variables A for which there is a rule $A \to a_i$.

The CYK Parsing Algorithm

The Inductive Step

- For $i < j$, compute

$$V_{ij} = \bigcup_{i \le k < j} \{A \in V \mid A \to BC \text{ and } B \in V_{ik} \text{ and } C \in V_{k+1,j}\}.$$

- That is, $B \overset{*}{\Rightarrow} a_i \cdots a_k$ and $C \overset{*}{\Rightarrow} a_{k+1} \cdots a_j$, so

$$A \Rightarrow BC \overset{*}{\Rightarrow} a_i \cdots a_j.$$

Example

Example (The CYK Parsing Algorithm)

- Let the grammer G be

$$S \to SAS \mid \mathbf{b}A\mathbf{a}$$
$$A \to \mathbf{a}S \mid S\mathbf{b} \mid \mathbf{ab} \mid \lambda$$

- Let $w = \mathbf{ababaa}$.
- Is $w \in L(G)$?

Example

Example (The CYK Parsing Algorithm)

- The grammar converted to CNF is

$$S \to SD \mid CE \mid SS \mid CB$$
$$A \to BS \mid SC \mid BC$$
$$B \to \mathbf{a}$$
$$C \to \mathbf{b}$$
$$D \to AS$$
$$E \to AB$$

The CYK Parsing Algorithm

Example (The CYK Parsing Algorithm)

- The CYK algorithm is much easier to perform if we use an $n \times n$ table.
- The entry in row i, column j is the set V_{ij}.
- We begin by initializing the diagonal elements (i, i) and then proceed inductively to the right.

Example

Example (The CYK Parsing Algorithm)

B					
	C				
		B			
			C		
				B	
					B

Example

Example (The CYK Parsing Algorithm)

B	A				
	C	S			
		B	A		
			C	S	
				B	$\varnothing$
					B

Example

Example (The CYK Parsing Algorithm)

B	A	A, E			
	C	S	A		
		B	A	A, E	
			C	S	$\varnothing$
				B	$\varnothing$
					B

Example

Example (The CYK Parsing Algorithm)

B	A	A, E	$\varnothing$		
	C	S	A	S, E	
		B	A	A, E	E
			C	S	$\varnothing$
				B	$\varnothing$
					B

Example

Example (The CYK Parsing Algorithm)

B	A	A, E	$\varnothing$	A, D	
	C	S	A	S, E	S
		B	A	A, E	E
			C	S	$\varnothing$
				B	$\varnothing$
					B

Example

Example (The CYK Parsing Algorithm)

B	A	A, E	$\varnothing$	A, D	A, E
	C	S	A	S, E	S
		B	A	A, E	E
			C	S	$\varnothing$
				B	$\varnothing$
					B

Example

Example (The CYK Parsing Algorithm)

- Because $S \notin V_{1,6}$, **ababaa** $\notin L(G)$

Example

Example (The CYK Parsing Algorithm)

- Because $S \notin V_{1,6}$, **ababaa** $\notin L(G)$
- Now perform the algorithm for $w =$ **babaaba**

Example

Example (The CYK Algorithm)

C						
	B					
		C				
			B			
				B		
					C	
						B

Example

Example (The CYK Algorithm)

C	S					
	B	A				
		C	S			
			B	$\varnothing$		
				B	A	
					C	S
						B

Example

Example (The CYK Algorithm)

C	S	A				
	B	A	A, E			
		C	S	$\varnothing$		
			B	$\varnothing$	$\varnothing$	
				B	A	A, E
					C	S
						B

Example

Example (The CYK Algorithm)

C	S	A	S, E			
	B	A	A, E	E		
		C	S	$\varnothing$	$\varnothing$	
			B	$\varnothing$	$\varnothing$	$\varnothing$
				B	A	A, E
					C	S
						B

Example

Example (The CYK Algorithm)

C	S	A	S, E	S		
	B	A	A, E	E	$\varnothing$	
		C	S	$\varnothing$	$\varnothing$	$\varnothing$
			B	$\varnothing$	$\varnothing$	$\varnothing$
				B	A	A, E
					C	S
						B

Example

Example (The CYK Algorithm)

C	S	A	S, E	S	A	
	B	A	A, E	E	$\varnothing$	$\varnothing$
		C	S	$\varnothing$	$\varnothing$	$\varnothing$
			B	$\varnothing$	$\varnothing$	$\varnothing$
				B	A	A, E
					C	S
						B

Example

Example (The CYK Algorithm)

C	S	A	S, E	S	A	S, E
	B	A	A, E	E	$\varnothing$	$\varnothing$
		C	S	$\varnothing$	$\varnothing$	$\varnothing$
			B	$\varnothing$	$\varnothing$	$\varnothing$
				B	A	A, E
					C	S
						B

Example

Example (The CYK Parsing Algorithm)

- Because $S \in V_{1,7}$, **babaaba** $\in L(G)$

Example

Example (The CYK Parsing Algorithm)

- Let the grammer G be

$$E \to E + E \mid E * E \mid (E) \mid \mathbf{a} \mid \mathbf{b}$$

- Use the CYK algorithm to show that $\mathbf{a}(+\mathbf{b} * \mathbf{c}) \notin L(G)$.

Outline

Collected

To be collected on Wed, Oct 12:

- Section 4.2 Exercises 14, 18.
- Section 4.3 Exercise 2, 5c.
- Section 5.1 Exercise 12b, 23.
- Section 6.1 Exercise 10.

Outline

Assignment

Assignment

- Section 6.3 Exercises 1, 4.