CS 536 Announcements for Tuesday, March 8, 2022

Last Time
- wrap up Java CUP
  - translating lists
  - handling unary minus
- intro to parsing
- review / midterm 1 questions

Today
- Bottom-up parsing
- CYK algorithm
- Chomsky normal form (CNF)

Next Time
- Wrap up CYK
- Top-down parsing

Parsing: two approaches

Top-down / "goal driven"
- start at start nonterminal
- grow parse tree downward until entire sequence is matched

Bottom-up / "data driven"
- start with terminals (sequence)
- generate ever larger subtrees until get to single tree whose root is the start nonterminal
Cocke – Younger – Kasami (CYK) algorithm

- Works bottom-up
- Time complexity: $O(n^3)$
- Requires grammar to be in Chomsky Normal Form

Chomsky Normal Form (CNF)

- All rules must be in one of two forms
  - $x \rightarrow T$
  - $x \rightarrow a \ b$
- Only rule allowed to derive epsilon is the start symbol $s$

Why CNF is helpful?

- Nonterminals in pairs
- Nonterminals (except start) can't derive epsilon

CYK: Dynamic Programming

$x \rightarrow T$

$x \rightarrow a \ b$
Running CYK

Track every viable subtree from leaf to root.

All subspans for a sequence (string) with 6 terminals
CYK Example

\[ f \rightarrow iw \]
\[ f \rightarrow iy \]
\[ w \rightarrow lx \]
\[ x \rightarrow nr \]
\[ y \rightarrow lr \]
\[ n \rightarrow ID \]
\[ n \rightarrow iz \]
\[ z \rightarrow cn \]
\[ i \rightarrow ID \]
\[ l \rightarrow ( \]
\[ r \rightarrow ) \]
\[ c \rightarrow , \]
Eliminating useless nonterminals

Avoid unnecessary work – remove useless rules

1. If a nonterminal cannot derive a sequence of terminal symbols, then it is useless
2. If a nonterminal cannot be derived from the start symbol, then it is useless

Nonterminals that cannot derive a sequence of terminal symbols

mark all terminal symbols
repeat
  if all symbols on the RHS of a production are marked
    mark the LHS nonterminal
until no more nonterminals can be marked

Example

s \rightarrow x \mid y
x \rightarrow ()
y \rightarrow (yy)

Nonterminals that cannot be derived from the start symbol

mark the start symbol
repeat
  if the LHS of a production is marked
    mark all RHS nonterminals
until no more nonterminals can be marked

Example

s \rightarrow a b
a \rightarrow + \mid - \mid \varepsilon
b \rightarrow \text{digit} \mid b \text{digit}
c \rightarrow . b
Chomsky Normal Form

Four steps

- eliminate epsilon productions
- eliminate unit productions
- fix productions with terminal on RHS
- fix productions with > 2 nonterminals on RHS

Eliminate (most) epsilon productions
If nonterminal A immediately derives epsilon

Example 1

\[
\begin{align*}
f & \rightarrow \text{ID ( } a \text{ )} \\
a & \rightarrow \epsilon \\
a & \rightarrow n \\
n & \rightarrow \text{ID} \\
n & \rightarrow \text{ID} , n
\end{align*}
\]

Example 2

\[
\begin{align*}
x & \rightarrow aXaYa \\
a & \rightarrow \epsilon \\
a & \rightarrow Z
\end{align*}
\]
Chomsky Normal Form (cont.)

Eliminate unit productions
Productions of the form $a \rightarrow b$ are called *unit productions*

Example

$\begin{align*}
  f & \rightarrow \text{ID}( a ) \\
  f & \rightarrow \text{ID}( ) \\
  a & \rightarrow n \\
  n & \rightarrow \text{ID} \\
  n & \rightarrow \text{ID}, n
\end{align*}$

Fix RHS terminals
For productions with terminals and something else on the RHS

Example

$\begin{align*}
  f & \rightarrow \text{ID}( n ) \\
  f & \rightarrow \text{ID}( ) \\
  n & \rightarrow \text{ID} \\
  n & \rightarrow \text{ID}, n
\end{align*}$
Chomsky Normal Form (cont.)

**Fix RHS nonterminals**
For productions with > 2 nonterminals on the RHS

Example