Compilerconstructie

najaar 2013

http://www.liacs.nl/home/rvvliet/coco/

Rudy van Vliet
kamer 124 Snellius, tel. 071-527 5777
rvvliet(at)liacs(dot)nl

college 1, dinsdag 3 september 2013

Overview
Why this course

It’s part of the general background of a software engineer

- How do compilers work?
- How do computers work?
- What machine code is generated for certain language constructs?
- Working on a non-trivial programming project

After the course

- Know how to build a compiler for a simplified program language
- Know how to use compiler construction tools, such as generators for scanners and parsers
- Be familiar with compiler analysis and optimization techniques
Course Outline

- In class, we discuss the theory using the ‘dragon book’ by Aho et al.
- The theory is applied in the practicum to build a compiler that converts Pascal code to MIPS instructions.

Course Outline

- In class, we discuss the theory using the ‘dragon book’ by Aho et al.
- The theory is applied in the practicum to build a compiler that converts Pascal code to MIPS instructions.

Earlier edition

- Dragon book has been revised in 2006
- In Second edition good improvements are made
  - Parallelism
    * ...  
    * Array data-dependence analysis
- First edition may also be used

A.V. Aho, R. Sethi, and J.D. Ullman,
Compilers: Principles, Techniques, and Tools,
Course Outline

• Contact
  – Room 124, tel. 071-5275777, rvvliet(at)liacs(dot)nl
  – Course website: http://www.liacs.nl/home/rvvliet/coco/
    Lecture slides, assignments, grades

• Practicum
  – 4 self-contained assignments
  – These assignments are done by groups of two persons
  – Assignments are submitted by e-mail
  – Assistants: Mathijs van de Nes (and Teddy Zhai)

• Written exam
  – 17 December 2013, 10:00–13:00
  – 11 March 2014, 14:00–17:00
Course Outline

- Grading:
  Average (50-50) of the grades from the written exam and the practicum

- You need to pass all 4 assignments to obtain a grade

- Final grade is only accepted if all grades are $\geq 5.5$

- Then, you obtain 6 EC

Studying only from the lecture slides may not be sufficient. Relevant book chapters will be given.
Course Outline

(tentative)
1. Overview
2. Lexical Analysis
3. Syntax Analysis Part 1
4. Syntax Analysis Part 2 (+ exercise class)
5. Assignment 1
6. Static Type Checking
7. Assignment 2
8. Intermediate Code Generation (+ exercise class)
9. Assignment 3
10. Code Generation
11. Code optimization (+ exercise class)
12. Assignment 4
13. spare date
14. Assignment 4 (extra session)
Practicum

• Assignment 1: Calculator

• Assignment 2: Parsing & Syntax tree

• Assignment 3: Intermediate code

• Assignment 4: Assembly generation

2 academic hours of Lab session + 3 weeks to complete (except assignment 1)
Short History of Compiler Construction

Formerly ‘a mystery’, today one of the best known areas of computing

1957 Fortran first compilers
   (arithmetic expressions, statements, procedures)

1960 Algol first formal language definition
   (grammars in Backus-Naur form, block structure, recursion, . . . )

1970 Pascal user-defined types, virtual machines (P-code)

1985 C++ object-orientation, exceptions, templates

1995 Java just-in-time compilation

We only consider imperative languages
Functional languages (e.g., Lisp) and logical languages (e.g., Prolog) require different techniques.
Compilers and Interpreters

- **Compilation:**
  Translation of a program written in a source language into a semantically equivalent program written in a target language.

- **Interpretation:**
  Performing the operations implied by the source program.
Compilers and Interpreters

- Compiler: Translates source code into machine code, with scanner, parser, . . ., code generator
- Interpreter: Executes source code ‘directly’, with scanner, parser
  Statements in, e.g., a loop are scanned and parsed again and again
Compilers and Interpreters

- Hybrid compiler (Java):
  - Translation of a program written in a source language into a semantically equivalent program written in an intermediate language (bytecode)
  - Interpretation of intermediate program by virtual machine, which simulates physical machine
Analysis-Synthesis Model of Compilation

There are two parts to compilation:

- Analysis
  - Determines the operations implied by the source program which are recorded in an intermediate representation, e.g., a tree structure

- Synthesis
  - Takes the intermediate representation and translates the operations therein into the target program
Other tools that use A-S Model

- Editors (syntax highlighting, text auto completion)

- Text formatters (\LaTeX{}, MS Word)
Compilation flow

- source program
  - Preprocessor
    - modified source program
      - Compiler
        - target assembly program
          - Assembler
            - relocatable machine code
              - Linker/Loader
                - target machine code
                  - library files
                    - relocatable object files
The Phases of a Compiler

source program / character stream

\[ \begin{align*}
\text{Lexical Analyser (scanner)} \\
\text{Syntax Analyser (parser)} \\
\text{Semantic Analyser} \\
\text{Intermediate Code Generator} \\
\text{Code optimizer} \\
\text{Code Generator} \\
\end{align*} \]

Symbol Table

target machine code
The Phases of a Compiler

Character stream:

position = initial + rate \times 60

Lexical Analyser (scanner)

Token stream:

⟨id, 1⟩ ⟨=⟩ ⟨id, 2⟩ ⟨+⟩ ⟨id, 3⟩ ⟨∗⟩ ⟨60⟩
The Phases of a Compiler

Token stream:

\[
\langle \text{id}, 1 \rangle \langle = \rangle \langle \text{id}, 2 \rangle \langle + \rangle \langle \text{id}, 3 \rangle \langle * \rangle \langle 60 \rangle
\]

Syntax Analyser (parser)

Parse tree / syntax tree:

```
stmt
  id = expr
    expr + term
      term term * factor
        factor factor number
          id id
```
The Phases of a Compiler

Syntax tree:

```
=  
|---+---
|   |---*
|   |   60
|   |   
|   |   |
|   | id,3
|   +---+
|       |---*
|       |   inttofloat
|       |   60
|       |
| id,2
|   +---+
|       |---+
|       |   id,1
```

Semantic Analyser

Syntax tree:

```
=  
|---+---
|   |---*
|   |   60
|   |   
|   |   |
|   | id,3
|   +---+
|       |---*
|       |   inttofloat
|       |   60
|       |
| id,2
|   +---+
|       |---+
|       |   id,1
```
The Phases of a Compiler

Syntax tree:

\[
\begin{aligned}
&\text{\(=\)} \\
&\quad \text{\(<\text{id}, 1>\)} \\
&\quad \quad \text{\(+\)} \\
&\quad \quad \quad \text{\(<\text{id}, 2>\)} \\
&\quad \quad \quad \quad \text{\(*\)} \\
&\quad \quad \quad \quad \quad \text{\(<\text{id}, 3>\)} \\
&\quad \quad \quad \quad \quad \quad \text{\(\text{inttofloat}\)} \\
&\quad \quad \quad \quad \quad \quad \quad \text{\(60\)}
\end{aligned}
\]

Intermediate Code Generator

Intermediate code (three-address code):

\[
\begin{align*}
\text{t1} &= \text{inttofloat}(60) \\
\text{t2} &= \text{id3} \times \text{t1} \\
\text{t3} &= \text{id2} + \text{t2} \\
\text{id1} &= \text{t3}
\end{align*}
\]
The Phases of a Compiler

Intermediate code (three-address code):

\[
\begin{align*}
t1 &= \text{inttofloat}(60) \\
t2 &= \text{id3} \times t1 \\
t3 &= \text{id2} + t2 \\
id1 &= t3
\end{align*}
\]

Code Optimizer

Intermediate code (three-address code):

\[
\begin{align*}
t1 &= \text{id3} \times 60.0 \\
id1 &= \text{id2} + t1
\end{align*}
\]
The Phases of a Compiler

Intermediate code (three-address code):

\[
\begin{align*}
t1 &= \text{id3} \times 60.0 \\
id1 &= \text{id2} + t1
\end{align*}
\]

Code Generator

Target code (assembly code):

\[
\begin{align*}
\text{LDF} & \quad \text{R2, id3} \\
\text{MULF} & \quad \text{R2, R2, #60.0} \\
\text{LDF} & \quad \text{R1, id2} \\
\text{ADDF} & \quad \text{R1, R1, R2} \\
\text{STF} & \quad \text{id1, R1}
\end{align*}
\]
The Grouping of Phases

- Front End:
  scanning, parsing, semantic analysis, intermediate code generation
  (source code → intermediate representation)

- Back End:
  code optimizing, code generation
  (intermediate representation → target machine code)

<table>
<thead>
<tr>
<th>language-dependent</th>
<th>machine-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Java</td>
<td>Pentium</td>
</tr>
<tr>
<td>C</td>
<td>PowerPC</td>
</tr>
<tr>
<td>Pascal</td>
<td>SPARC</td>
</tr>
</tbody>
</table>

24
Passes: Single-Pass Compilers

Phases work in an interleaved way

\[
\text{do} \\
\text{scan token} \\
\text{parse token} \\
\text{check token} \\
\text{generate code for token} \\
\text{while (not eof)}
\]

Portion of code is generated while reading portion of source program
Passes: Multi-Pass Compilers

Phases are separate ‘programs’, which run sequentially

characters $\rightarrow$ Scanner $\rightarrow$ tokens $\rightarrow$ Parser $\rightarrow$ tree $\rightarrow$ Semantic analyser $\rightarrow$ \ldots $\rightarrow$ code

Each phase reads from a file and writes to a new file.

Time vs memory

Why multi-pass?
  - If the language is complex
  - If portability is important

Today: often two-pass compiler
Compiler-Construction Tools

Software development tools are available to implement one or more compiler phases

- Scanner generators
- Parser generators
- Syntax-directed translation engines
- Automatic code generators
- Data-flow engines
The Structure of our compiler

Syntax directed translation:
The compiler uses the syntactic structure of the language to generate output
What is a grammar?

Context-free grammar is a 4-tuple with
- A set of nonterminals (syntactic variables)
- A set of tokens (terminal symbols)
- A designated start symbol (nonterminal)
- A set of productions: rules how to decompose nonterminals

Example: Context-free grammar for simple expressions:

\[ G = (\{list, digit\}, \{+,-,0,1,2,3,4,5,6,7,8,9\}, list, P) \]

with productions P:

\[
\begin{align*}
\text{list} & \rightarrow \text{list} + \text{digit} \\
\text{list} & \rightarrow \text{list} - \text{digit} \\
\text{list} & \rightarrow \text{digit} \\
\text{digit} & \rightarrow 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
\end{align*}
\]
Derivation

Given a context-free grammar, we can determine the set of all strings (sequences of tokens) generated by the grammar using derivations:

- We begin with the start symbol

- In each step, we replace one nonterminal in the current form with one of the right-hand sides of a production for that nonterminal
Derivation (Example)

\[
G = (\{list, digit\}, \{+, -, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}, list, \ P)
\]

list → list + digit | list − digit | digit

digit → 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

Example: 9-5+2

\[
\begin{align*}
list & \Rightarrow list + digit \\
& \Rightarrow list − digit + digit \\
& \Rightarrow digit − digit + digit \\
& \Rightarrow 9 − digit + digit \\
& \Rightarrow 9 − 5 + digit \\
& \Rightarrow 9 − 5 + 2
\end{align*}
\]

This is an example of leftmost derivation, because we replaced the leftmost nonterminal (underlined) in each step
Parse Tree
(derivation tree in FI2)

- The root of the tree is labelled by the start symbol

- Each leaf of the tree is labelled by a terminal (=token) or $\epsilon$ (=empty)

- Each interior node is labelled by a nonterminal

- If node $A$ has children $X_1, X_2, \ldots, X_n$, then there must be a production $A \rightarrow X_1X_2\ldots X_n$
Parse Tree (Example)

Parse tree of the string $9 - 5 + 2$ using grammar $G$

Yield of the parse tree: the sequence of leaves (left to right)

Parsing: the process of finding a parse tree for a given string

Language: the set of strings that can be generated by some parse tree
Ambiguity

Consider the following context-free grammar:

\[ G' = (\{string\}, \{+,-,0,1,2,3,4,5,6,7,8,9\}, string, P) \]

with productions \( P \)

\[
string \rightarrow string + string \mid string - string \mid 0 \mid 1 \mid \ldots \mid 9
\]

This grammar is ambiguous, because more than one parse tree generates the string \( 9 - 5 + 2 \)
Ambiguity (Example)

Parse trees of the string $9 - 5 + 2$ using grammar $G'$

$(9 - 5) + 2 = 6$

$9 - (5 + 2) = 2$
Associativity of Operators

By convention

\[
\begin{align*}
9 + 5 + 2 &= (9 + 5) + 2 \\
9 - 5 - 2 &= (9 - 5) - 2
\end{align*}
\]  
left associative

In most programming languages:

+ , − , * , / are left associative

** , = are right associative:

\[
\begin{align*}
a ** b ** c &= a ** (b ** c) \\
a = b = c &= a = (b = c)
\end{align*}
\]
Precedence of Operators

Consider: $9 + 5 \times 2$
Is this $(9 + 5) \times 2$ or $9 + (5 \times 2)$?

Associativity does not resolve this.

Precedence of operators: $\begin{array}{c}
+ \\
- \\
\times \\
/ \\
\end{array}$

increasing precedence

A grammar for arithmetic expressions: ...

Example:

$9 + 5 \times 2 \times 3 + 1 + 4 \times 7$
Precedence of Operators

Consider: $9 + 5 \times 2$
Is this $(9 + 5) \times 2$ or $9 + (5 \times 2)$?
Associativity does not resolve this.

Precedence of operators:

\[ + - \quad \downarrow \quad \text{increasing precedence} \]

\[ \times / \quad \downarrow \quad \text{precedence} \]

A grammar for arithmetic expressions:

\[
\begin{align*}
\text{expr} & \rightarrow \text{expr} + \text{term} \mid \text{expr} - \text{term} \mid \text{term} \\
\text{term} & \rightarrow \text{term} \times \text{factor} \mid \text{term}/\text{factor} \mid \text{factor} \\
\text{factor} & \rightarrow \text{digit} \mid (\text{expr}) \\
\text{digit} & \rightarrow 0 \mid 1 \mid \ldots \mid 9
\end{align*}
\]

Parse tree for $9 + 5 \times 2$ . . .
Syntax-Directed Translation

Using the syntactic structure of the language to generate output corresponding to some input

Two techniques:
- Syntax-directed definition
- Translation scheme

Example: translation of infix notation to postfix notation

<table>
<thead>
<tr>
<th>Infix</th>
<th>Postfix</th>
</tr>
</thead>
<tbody>
<tr>
<td>(9 - 5) + 2</td>
<td>95 - 2 +</td>
</tr>
<tr>
<td>9 - (5 + 2)</td>
<td>952 + -</td>
</tr>
</tbody>
</table>

What is $952 + -3^*$?
Syntax-Directed Translation

Using the syntactic structure of the language to generate output corresponding to some input

Two variants:
- Syntax-directed definition
- Translation scheme

Example: translation of infix notation to postfix notation

Simple infix expressions generated by

\[
\begin{align*}
expr & \rightarrow expr_1 + term \mid expr_1 - term \mid term \\
term & \rightarrow 0 \mid 1 \mid \ldots \mid 9
\end{align*}
\]
Syntax-Directed Definition

- Uses a context-free grammar to specify the syntactic structure of the language

- Associates a set of attributes with (non)terminals

- Associates with each production a set of semantic rules for computing values for the attributes

- The attributes contain the translated form of the input after the computations are completed (in example: postfix notation corresponding to subtree)
Syntax-Directed Definition (Example)

<table>
<thead>
<tr>
<th>Production</th>
<th>Semantic rule</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>expr → expr_1 + term</code></td>
<td>`expr.t = expr_1.t</td>
</tr>
<tr>
<td><code>expr → expr_1 − term</code></td>
<td>`expr.t = expr_1.t</td>
</tr>
<tr>
<td><code>expr → term</code></td>
<td><code>expr.t = term.t</code></td>
</tr>
<tr>
<td><code>term → 0</code></td>
<td><code>term.t = '0'</code></td>
</tr>
<tr>
<td><code>term → 1</code></td>
<td><code>term.t = '1'</code></td>
</tr>
<tr>
<td>[\ldots]</td>
<td>[\ldots]</td>
</tr>
<tr>
<td><code>term → 9</code></td>
<td><code>term.t = '9'</code></td>
</tr>
</tbody>
</table>

Result: annotated parse tree
Example: 9 − 5 + 2
Synthesized and Inherited Attributes

An attribute is said to be . . .

- **synthesized** if its value at a parse tree node $N$ is determined from attribute values at the children of $N$ (and at $N$ itself)

- **inherited** if its value at a parse tree node $N$ is determined from attribute values at the parent of $N$ (and at $N$ itself and its siblings)

We (mainly) consider synthesized attributes
Depth-First Traversal

• A syntax-directed definition does not impose an evaluation order of the attributes on a parse tree

• Different orders might be suitable

• *Tree traversal*: a specific order to visit the nodes of a tree (always starting from the root node)

• Depth-first traversal
  – Start from root
  – Recursively visit children *(in any order)*
  – Hence, visit nodes far away from the root as quickly as it can (DF)
A Possible DF Traversal

Postorder traversal

procedure visit (node N)
{
    for (each child C of N, from left to right)
    {
        visit (C);
    }
    evaluate semantic rules at node N;
}

Can be used to determine synthesized attributes / annotated parse tree
Translation Scheme

A translation scheme is a context-free grammar with semantic actions embedded in the bodies of the productions.

Example

\[
\text{rest} \rightarrow + \text{term} \ \text{rest}_1
\]

With semantic action:

\[
\text{rest} \rightarrow + \text{term} \ \{\text{print}(’+’)\} \ \text{rest}_1
\]
Translation Scheme

A translation scheme is a context-free grammar with semantic actions embedded in the bodies of the productions.

Example

\[ \text{rest} \rightarrow + \text{term} \{ \text{print}('+' ) \} \text{ rest}_1 \]

Corresponding effect on parse tree:
Translation Scheme (Example)

\[
\begin{align*}
expr & \rightarrow \ expr_1 + \ term \ \{\text{print('+'})\} \\
expr & \rightarrow \ expr_1 - \ term \ \{\text{print('-')}\} \\
expr & \rightarrow \ term \\
term & \rightarrow \ 0 \ \{\text{print('0')}\} \\
term & \rightarrow \ 1 \ \{\text{print('1')}\} \\
\ldots & \ldots \\
term & \rightarrow \ 9 \ \{\text{print('9')}\}
\end{align*}
\]

Example: parse tree for \(9 - 5 + 2\)

Implementation requires postorder traversal (LRW)
Parsing

• Process of determining if a string of tokens can be generated by a grammar

• For any context-free grammar, there is a parser that takes at most $O(n^3)$ time to parse a string of $n$ tokens

• Linear algorithms sufficient for parsing programming languages

• Two methods of parsing:
  – **Top-down** constructs parse tree from root to leaves
  – **Bottom-up** constructs parse tree from leaves to root

Cf. top-down PDA and bottom-up PDA in FI2
Parsing (Top-Down Example)

\[
\begin{align*}
stmt & \rightarrow expr ; \\
 & \mid if ( expr ) stmt \\
 & \mid for ( optexpr ; optexpr ; optexpr ) stmt \\
 & \mid other \\
optexpr & \rightarrow \epsilon \\
 & \mid expr
\end{align*}
\]

How to determine parse tree for

\[\text{for (; expr ; expr ) other}\]

Use lookahead: current terminal in input
Predictive Parsing

• Recursive-descent parsing is a top-down parsing method:
  – Executes a set of recursive procedures to process the input
  – Every nonterminal has one (recursive) procedure parsing the nonterminal’s syntactic category of input tokens

• Predictive parsing is a special form of recursive-descent parsing:
  – The lookahead symbol unambiguously determines the production for each nonterminal
Predictive Parsing (Example)

```c
void stmt()
{
    switch (lookahead)
    {
    case expr:
        match(expr); match(';'); break;
    case if:
        match(if); match('('); match(expr); match(')'); stmt(); break;
    case for:
        match(for); match('(');
        optexpr(); match('); optexpr(); match(';'); optexpr();
        match(')'); stmt(); break;
    case other:
        match(other); break;
    default:
        report("syntax error");
    }
}

void match(terminal t)
{
    if (lookahead==t) lookahead = nextTerminal;
    else report("syntax error");
}
```
Using **FIRST**

- Let $\alpha$ be string of grammar symbols
- $\text{FIRST}(\alpha)$ is the set of terminals that appear as first symbols of strings generated from $\alpha$

Simple example:

\[
\text{stmt} \rightarrow \text{expr} ; \\
\quad \mid \text{if (expr )stmt} \\
\quad \mid \text{for (optexpr ; optexpr ; optexpr )stmt} \\
\quad \mid \text{other}
\]

Right-hand side may start with nonterminal...
Using FIRST

- Let $\alpha$ be string of grammar symbols
- $\text{FIRST}(\alpha)$ is the set of terminals that appear as first symbols of strings generated from $\alpha$
- When a nonterminal has multiple productions, e.g.,

$$A \rightarrow \alpha \mid \beta$$

then $\text{FIRST}(\alpha)$ and $\text{FIRST}(\beta)$ must be disjoint in order for predictive parsing to work
Compilerconstructie

college 1
Overview

Chapters for reading: 1.1, 1.2, 2.1-2.5