Announcements

• Homework set 2 solutions will be posted by Thursday.

• The first programming project has been posted. Deadline: Thursday, October 25.

• Please check whether you have access to piazza on sakai. If not, let me know.
Review: Typing and Type Systems

Read Scott: Chapter 7.1 - 7.2; ALSU Chapter 6.5

type:

A set of values and meaningful operations on them.

Types provide semantic “sanity checks” (consistency checks). Types help identify:

• errors, if an operator is applied to an incompatible operand.
  ○ dereferencing of a non-pointer
  ○ adding a function name to something
  ○ incorrect number of parameters to a procedure
  ○ …

• which operation to use for overloaded names and operators, or what type coercion to use (e.g.: 3.0 + 1)

• identification of polymorphic functions. Such functions can be executed with arguments of several types (e.g.: list of objects of unknown type $\alpha$)
Type systems

Each language construct (operator, expression, statement, ...) has a type

**basic types:** integer, real, character, ...

**constructed types:** arrays, records, sets, pointers, functions

A **type system** is a collection of *rules* for assigning *type expressions* to operators, expressions, ... in the program. Type systems are language dependent.

A **type checker** implements the type system, i.e., deduces type expressions for program constructs based on the type inference rules of the type system. The type checker “computes” or “reconstructs” type expressions.
Example type rules

- If both operands of the arithmetic operators of addition, subtraction, and multiplication are of type integer, then the result is of type integer (Pascal definition).

Rule for + (analogue rules for − and ⋅):

\[
\begin{align*}
E \vdash e_1 : \text{integer} & \quad E \vdash e_2 : \text{integer} \\
\hline
E \vdash (e_1 + e_2) : \text{integer}
\end{align*}
\]

where \( E \) is a type environment that maps constants and variables to their types.

In combination with the following axiom template in the type system \( \{c : \alpha\} \vdash c : \alpha \)

we can now infer that \((2 + 3)\) is of type integer:

\[
\begin{align*}
E \vdash 2 : \text{integer} & \quad E \vdash 3 : \text{integer} \\
\hline
E \vdash (2 + 3) : \text{integer}
\end{align*}
\]

where \( E = \{2 : \text{integer}, 3 : \text{integer}\} \).

In general, type deduction proofs work bottom up.
Example type rules

- The result of the unary & operator is a pointer to the object referred to by the operand. If the operand is of type “foo”, then the type of the result is a “pointer to foo”. (C and C++ definition)

\[
E \vdash e : \alpha \\
\hline
E \vdash \& e : \text{pointer}(\alpha)
\]

- Two expressions can only be compared if they have the same types. The result is of type boolean.

\[
E \vdash e_1 : \alpha \quad E \vdash e_2 : \alpha \\
\hline
E \vdash (e_1 = e_2) : \text{boolean}
\]

In the examples, integer and $\alpha$ are type expressions.
Type expressions  

inductive definition

1. A basic type is a type expression. A special basic type, `typeError` will signal an error. A basic type `void` denotes an untyped statement.

2. Since type expressions may be named, a type name is a type expression.

3. Type expressions may contain variables whose values are type expressions (e.g.: useful for languages without type declarations, or polymorphism).

4. A `type constructor` applied to type expressions is a type expression. Examples:

   (a) arrays
   (b) cartesian products
   (c) records
   (d) pointers
   (e) functions
Type checking

The purpose of type checking is to prevent type errors. Type errors can occur during the execution of a program, for instance, when a function is applied with an argument of the wrong type or if a variable with a non-function type is called.

“How much” type checking is done is part of the definition of the programming language.

Type checking can be performed

- at compile-time (only)
- at compile-time and program execution time, or
- at program execution time (only).
Type checking

- A *strongly typed* language guarantees that the compiler will accept only programs that execute without type errors, i.e., the compiler ensures that no type error will *remain undetected*. Strongly typed languages can use static and dynamic type checking.

  Standard example: array bounds checking.
  
  
  \[
  \begin{align*}
  \text{table: array[0..255] of char;} \\
  \text{i: integer} \\
  \text{table[i] cannot be guaranteed at compile time to fall in the range of 0 to 255}
  \end{align*}
  \]

- In principle, type checking can be done entirely dynamically. However, this requires *type tags* for each value and each function application has to check for matching tags at run-time.
**Type checking — discussion**

**compile–time:**

+ points out time errors early  
+ no overhead at program execution time (run–time)  
– cannot always be done  
? parts of the program that may never be executed are still checked (too restrictive)

**program execution time:**

+ allows more flexibility (e.g.: type may change depending on the use of a variable)  
– overhead in terms of space and time  
? part of program that are not executed are not checked  
+ rapid prototyping  
– harder to debug

Goal: Strongly typed languages with as much type checking as possible at compile–time
Representation of type expressions

A convenient way to represent a type expression is to use a graph.

Example:

```
      X
     / \  
char  char
```

\((\text{char} \times \text{char}) \rightarrow \text{int}\)

This makes comparisons of type expressions easier, or helps implementing operations on type expressions such as “unification”. Please see problem 9 in homework set 2.
Type names

Type expressions can be named and such names can occur in type expressions (e.g.: typedef in C).

When are two type expressions with type names the same?

*structural equivalence:*

- either same basic types (e.g., int is equivalent to int), or
- formed by application of the same type constructor to structurally equivalent types, where
  - type names are abbreviations for type expressions

In some sense, structural equivalence deletes names in composite constructs and replaces them by the type structures that they represent ("unroll" the tree or DAG).

*name equivalence:*

- Two type expressions are equivalent iff they are identical
- each name is assumed to represent a different type.
type link1 = \uparrow A \neq \text{type link2} = \uparrow A
Structural type equivalence algorithm

function sequiv (s,t): boolean;
begin
  if s and t are the same basic type then
    return true
  else if s = array (s1, s2) and t = array (t1, t2) then
    return sequiv (s1, t1) and sequiv (s2, t2)
  else if s = s1 × s2 and t = t1 × t2 then
    return sequiv (s1, t1) and sequiv (s2, t2)
  else if s = pointer (s1) and t = pointer (t1) then
    return sequiv (s1, t1)
  else if s = s1 → s2 and t = t1 → t2 then
    return sequiv (s1, t1) and sequiv (s2, t2)
  else
    return false
end

Problem: Recursive data types
Type variables

Type expressions may contain variables (type variables) whose values are type expressions.

Type variables are used for implicitly typed languages or languages with polymorphic types.

Programming languages can be

- explicitly typed — every object is declared with its type (type checking)
- implicitly typed — type of object is derived from its use (type reconstruction)
- monomorphic — every object has a unique, single type
- polymorphic — allows objects to have more than one type (e.g., nil, and & in C)
Type variables — examples

Recall:

\[
E \vdash e_1 : \text{integer} \quad E \vdash e_2 : \text{integer} \\
E \vdash (e_1 + e_2) : \text{integer}
\]

where \(E\) is a type environment. In other words, “+” has the type expression \((\text{integer} \times \text{integer}) \rightarrow \text{integer}\).

What are the types of the variables \(a\) and \(b\) in the following program:

\[
\text{read}(a); \\
\text{read}(b); \\
... \quad a + b \quad ...;
\]

Here is an idea: Guess the types of variables you don’t know about and use unification to match guesses with rules.

\[
\begin{align*}
\text{read}(a) & \quad \{a: \alpha\} \\
\text{read}(b) & \quad \{a: \alpha, \ b: \beta\} \\
a + b & \quad \text{unify(} \alpha, \text{integer) } \\
& \quad \text{unify(} \beta, \text{integer) } \\
& \quad \text{apply type rule; result integer}
\end{align*}
\]
Type variables — examples

• Polymorphic **cons**:

\[
\begin{align*}
E \vdash e_1 : \alpha & \quad E \vdash e_2 : list(\alpha) \\
\hline
E \vdash (\text{cons } e_1 e_2) : list(\alpha)
\end{align*}
\]

**cons** has the type expression
\[
\forall \alpha. (\alpha \times list(\alpha)) \rightarrow list(\alpha)
\]

• Polymorphic **nil**:

\[
E \vdash \text{nil} : list(\alpha)
\]

**nil** has the type expression \( \forall \alpha. list(\alpha) \)

**Questions:**

- What is the type of \((\text{cons } a \text{ nil})\)?
- What is the type of \((\text{cons } 1 \text{ nil})\)?
- What is the type of \((\text{cons } a \text{ 1})\)?
Unification

A unifier is a substitution of variables by expressions such that, when applied to two expressions $expr_1$ and $expr_2$, both expressions become syntactically identical. The most general unifier ($mgu$) makes as few and as general bindings as possible.

Example:

$\text{times}(Z, \text{times}(Y, 7)) = \text{times}(4, W)$

$\Theta_1 = \{Z \rightarrow 4, Y \rightarrow \text{plus}(3, 5), W \rightarrow \text{times}(\text{plus}(3, 5), 7)\}$

$\Theta_2 = \{Z \rightarrow 4, W \rightarrow \text{times}(Y, 7)\}$

Question: Which of the two substitutions is more general?

Note: mgu $\theta$ is unique modulo renaming of variables
Type systems with polymorphism

How can we infer the type of a polymorphic function such as `length`?

```haskell
fun length (l) =
    if null(l) then 0
    else length(tl(l)) + 1;
```

- Need type variables to represent unknown types during type reconstruction (type inference).
- Need type environment \( E \) that
  - Keeps track of type expressions for program variables and type variables
  - Is initialized to predefined types of constants and functions
Algorithm for typing functions

High–level view of algorithm

1. Add fresh type variables to E for function name, arguments, and result. Add resulting type expression for function to E.

2. Analyze the program (parse tree) bottom up:

   - If constant, look–up type in E.
   - If variable, look–up type in E. If no entry in E, generate entry with fresh type variable.
   - If function application, look–up function in environment. ”Apply” application inference rule by unification of type expressions of formal and actual parameters and result.

\[
E \vdash e_1 : \alpha \rightarrow \beta \quad E \vdash e_2 : \alpha \\
\hline
E \vdash e_1(e_2) : \beta
\]

Note: If function is polymorphic (\(\forall \alpha. \ldots\)) generate fresh type variable for bound variable and remove \(\forall\) in resulting type expression.

If at any point unification fails, algorithm returns type error.
Algorithm for typing functions — example

Type inference for function \texttt{length}

Initial type environment:

\[ E = \{ \begin{array}{l}
0: \text{integer}, \\
1: \text{integer}, \\
\text{if}: \forall \alpha. \text{boolean} \times \alpha \times \alpha \to \alpha \\
\text{null}: \forall \alpha. \text{list} \langle \alpha \rangle \to \text{boolean} \\
\text{tl}: \forall \alpha. \text{list} \langle \alpha \rangle \to \text{list} \langle \alpha \rangle \\
+ : \text{integer} \times \text{integer} \to \text{integer} 
\end{array} \} \]

\begin{verbatim}
fun length (l) = 
  if null(l) then 0 
  else length(tl(l)) + 1;
\end{verbatim}

See ALSU 6.5.4, page 391 — 395
### Algorithm for typing functions — example

<table>
<thead>
<tr>
<th>Type expressions</th>
<th>Type environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l: \gamma$</td>
<td>$l: \gamma$</td>
</tr>
<tr>
<td>$\text{null}: \text{list}(\alpha_1) \rightarrow \text{boolean}$</td>
<td>$\text{null}(l): \text{boolean}$</td>
</tr>
<tr>
<td>$\text{null}(l): \text{boolean}$</td>
<td>$\gamma = \text{list}(\alpha_1)$</td>
</tr>
<tr>
<td>$0: \text{integer}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$1: \text{list}(\alpha_1)$</td>
<td>$\alpha_1 = \alpha_2$</td>
</tr>
<tr>
<td>$\text{tl}: \text{list}(\alpha_2) \rightarrow \text{list}(\alpha_2)$</td>
<td>$\text{tl}(l): \text{list}(\alpha_1)$</td>
</tr>
<tr>
<td>$\text{length}: \text{list}(\alpha_1) \rightarrow \delta$</td>
<td>$\text{length}(\text{tl}(l)): \delta$</td>
</tr>
<tr>
<td>$\text{length} + 1: \text{integer}$</td>
<td>$\delta = \text{integer}$</td>
</tr>
<tr>
<td>$\text{if}: \text{boolean} \times \alpha_3 \times \alpha_3 \rightarrow \alpha_3$</td>
<td>$\alpha_3 = \text{integer}$</td>
</tr>
<tr>
<td>$\text{if}(...): \text{integer}$</td>
<td></td>
</tr>
<tr>
<td>$\text{length}: \text{list}(\alpha_1) \rightarrow \text{integer}$</td>
<td></td>
</tr>
</tbody>
</table>

Since we don’t make any assumptions about $\alpha_1$, we get the polymorphic type for length: $\forall \alpha. \text{list}(\alpha) \rightarrow \text{integer}$. 
Project 1 – Polymorphic Type Reconstruction

The really-TINY language is defined as follows

\[ x \in \text{variables} \]
\[ n \in \text{integers} \]
\[ c ::= n \mid \#t \mid \#f \mid \text{add1} \mid \text{sub1} \mid \text{zero}? \mid \text{and} \mid \text{or} \mid \text{not} \]
\[ e ::= c \mid x \mid (\text{lambda} (x) e) \mid (e e) \]

\( \text{add1} \) adds the value 1 to its argument. \( \text{sub1} \) subtracts the value 1 from its argument.

Type rules:

\[ E \vdash n : \text{int} \quad E \vdash \#t : \text{bool} \quad E \vdash \#f : \text{bool} \]
\[ E \vdash \text{add1} : (\text{int} \rightarrow \text{int}) \quad E \vdash \text{sub1} : (\text{int} \rightarrow \text{int}) \quad E \vdash \text{zero}? : (\text{int} \rightarrow \text{bool}) \]
\[ E \vdash \text{and} : (\text{bool} \rightarrow (\text{bool} \rightarrow \text{bool})) \quad E \vdash \text{or} : (\text{bool} \rightarrow (\text{bool} \rightarrow \text{bool})) \]
\[ E \vdash \text{not} : (\text{bool} \rightarrow \text{bool}) \]
\[ E \vdash x : E(x) \quad \text{with } E(x) \in \text{type\_expressions} \]
\[ E \vdash x : \alpha \quad E[x \mapsto \alpha] \vdash e : \beta \]
\[ E \vdash (\text{lambda} (x) e) : \alpha \rightarrow \beta \]
\[ E \vdash e_1 : (\alpha \rightarrow \beta) \quad E \vdash e_2 : \alpha \]
\[ E \vdash (e_1 \ e_2) : \beta \]
Project 1 – Polymorphic Type Reconstruction

The provided Scheme function \textit{parse} maps a really-TINY program into an AST representation. Your type reconstructor should take as input an AST, a type environment, and a set of constraints:

\begin{verbatim}
(define TR
  (lambda (ast E C)
    ...
  )
)

(define TRec
  (lambda (m)
    ...
    ;; extract type expression from TR call
    (TR (parse m) init_E init_C) ...
  ))
\end{verbatim}

Good luck, and start early!
Review – Formal Languages

Alphabet: finite set of symbols – \( \{a, b, c\} \in \Sigma \)

String (or word): finite sequence of symbols from an alphabet – \( \{w, x, y\} \in \Sigma^* \);
\( \epsilon \in \Sigma^* \) denotes the empty word

Language: set of strings over an alphabet – \( L \subseteq \Sigma^* \)

Two problems:

- How to **specify** a formal language \( L \)?
  Specifications have to be finite.

- How to **recognize** that string \( w \in L \)?
  Procedure should be efficient (terminate?).

<table>
<thead>
<tr>
<th>specification</th>
<th>recognizing automaton</th>
<th>complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular expressions</td>
<td>DFA</td>
<td>( O(n) )</td>
</tr>
<tr>
<td>CFG</td>
<td>PDA</td>
<td>( O(n) / O(n^3) )</td>
</tr>
<tr>
<td>CSG</td>
<td>LBA</td>
<td>P-SPACE COMPLETE</td>
</tr>
<tr>
<td>Arbitrary rewrite systems</td>
<td>TM</td>
<td>r.e.</td>
</tr>
</tbody>
</table>
Compilers

source code → front end → IR → opt. → IR → back end → target code

**front end** produce an intermediate representation (**IR**) for the program.

**optimizer** transforms the code in IR form into an equivalent program that may run more efficiently.

**back end** transforms the code in IR form into native code for the target machine.

The IR encodes knowledge that the compiler has derived about the source program.
The role of the parser

Parser

- call scanner to get new token
- perform context-free syntax analysis
- guide the context-sensitive analysis
  \(\rightarrow\) SDT / attribute grammars
- construct an intermediate representation
  \(\rightarrow\) SDT / attribute grammars
- produce meaningful error messages
- attempt error correction
Syntax analysis

Context-free syntax is specified with a grammar.

Formally, a context-free grammar $G$ is a four-tuple $(T, NT, S, P)$

$T$ is the set of terminal symbols in the grammar. For our purposes, the set of terminals is equivalent to the set of tokens returned by the lexical analyzer.

$NT$ is a set of syntactic variables that denote sets of (sub)strings occurring in the language. These are used to impose a structure on the grammar.

$S$ is a distinguished nonterminal ($S \in NT$) that denotes the entire set of strings in $L(G)$. This is sometimes called a goal symbol or start symbol. In an ECFG (extended/augmented CFG), $S$ cannot appear on the right hand side of some production.

$P$ is a set of productions that specify the way that terminals and non-terminals can be combined to form strings in the language. Each production must have a single non-terminal on its left hand side.
Next class: Top-down and Bottom-up Parsing

For next time, please read:
ASU Ch 4.1 – 4.9