Assignment 3: Semantics & Intermediate Code

CS 464/5607: Compiler Design

Spring 2026

1 Introduction

In this assignment, you will implement two critical phases of the compiler pipeline:

1. **Phase 3: Semantic Analysis**: You will traverse the Abstract Syntax Tree (AST) to verify context-sensitive properties. This includes ensuring variables are declared before use, checking type compatibility in assignments, and verifying function arguments.
2. **Phase 4: Intermediate Representation (IR) Generation**: Once the program is validated, you will translate the AST into a linear, machine-independent Intermediate Representation (Three-Address Code) that serves as the bridge to final code generation.

2 Setup and Files

The complete implementation of the lexer and the parser is provided for this assignment

- `lib/`: Contains the **provided reference implementation** of Phase 1 and 2.
  - `lexer.l`: The reference Lexer.
  - `parser.y`: The reference Bison parser.
  - `ast.cpp / ast.h`: The reference AST implementation.

Part 1: Semantic Analysis

3 Error Reporting

To facilitate automated testing, semantic errors must be reported in a strictly standardized format.

3.1 The `errors.h` Interface

We have provided `errors.h` and `errors.c`, which define the `ErrorType` enum. **Note:** The specific error codes (e.g., `ERR_UNDECLARED_VAR`, `ERR_TYPE_MISMATCH_OP`) are documented in the comments within `errors.h`. Please review that file to understand which code corresponds to which semantic violation.

3.2 Using `report_error`

You must strictly use the provided helper function for all error output:

```
void report_error(ErrorType type, int lineno);
```

Crucial: Do not use `printf` manually for errors. The test suite relies on the specific string format generated by this function. Passing the correct `ErrorType` ensures your compiler matches the "Golden Output" required for grading.

4 The Symbol Table

The Symbol Table is the central data structure for Semantic Analysis. It tracks identifiers (variables, functions, arrays) and their associated attributes (types, scope, parameters) throughout the program.

4.1 Files

- `symbol_table.h`: Defines the API and structures.
- `symbol_table.cpp`: You must implement the logic here.

4.2 API Overview

You are required to implement the following functions to manage scope and symbols:

- `init_symbol_table()`: Initializes the global scope and data structures.
- `push_scope()`: Enters a new scope (e.g., entering a function body or `if` block).
- `pop_scope()`: Exits the current scope and removes local symbols.
- `insert_symbol(...)`: Adds a new symbol to the current scope. You must handle checks for redeclaration here (e.g., declaring `int x` twice in the same block).
- `lookup_symbol(...)`: Searches for a symbol starting from the current scope and moving outward to global scope.
- `lookup_local_symbol(...)`: Searches for a symbol *only* in the current scope (useful for detecting redeclarations).

5 Semantic Analysis

5.1 Entry Point

```
int semantic_analysis(ASTNode* root);
```

This function is called by the driver after parsing is complete. It should traverse the AST, perform all necessary checks, and return the **total number of semantic errors** found.

5.2 Implementation Hints

You will need to write a traversal function that visits every node in the AST.

- **Scope Management:** Ensure you call `push_scope()` and `pop_scope()` when entering/leaving blocks (functions, loops, conditionals).
- **Semantic Errors**
  You must implement checks for the following specific error conditions, as defined in `errors.h`:
  - **Undeclared Variable:** Attempting to use a variable that has not been declared in the current or global scope.
  - **Undeclared Function:** Attempting to call a function that has not been defined.

- **Variable Redeclaration:** Declaring a variable with the same name more than once in the same scope.
- **Function Redeclaration (No Overloading):** Defining a function with a name that is already used, regardless of parameter count or return type. Function overloading is **not** permitted.
- **Assignment Type Mismatch:** Assigning a value to a variable of an incompatible type. Implicit casting is disallowed. You cannot assign a `float` to an `int` or an `int` to a `float`.
- **Operation Type Mismatch:** Using incompatible operands in a binary operation. Operations between mixed types (e.g., `int + float`) are not allowed. Both operands must match exactly.
- **Return Type Mismatch:** Returning a value that does not match the function's declared return type.
- **Argument Count Mismatch:** Calling a function with fewer or more arguments than defined.
- **Argument Type Mismatch:** Passing an argument that does not match the expected type for that parameter position.
- **Not a Function:** Attempting to call an identifier that is not a function.
- **Not an Array:** Attempting to index an identifier that is not declared as an array.
- **Invalid Array Index:** Using a non-integer expression as an array index.

Suggested Helpers: It is highly recommended to create helper functions to keep your code clean. Consider helpers for:

- Determining the result type of an expression node.
- Checking compatibility between two types.
- Validating function arguments against formal parameters.

Part 2: Intermediate Representation (IR) Generation

Once the AST has been validated and confirmed to be free of semantic errors, the compiler proceeds to the final phase of this assignment: generating the **Intermediate Representation (IR)**.

6 Overview of IR

An Intermediate Representation is a machine-independent version of the code that sits between the high-level source and the low-level machine code. IR decomposes complex expressions into a sequence of simple instructions. This linear structure makes it much easier to perform optimization and final code generation.

7 The IR Structure (`ir.h`)

The interface for the IR is defined in `ir.h`. You will need to inspect this file closely to understand the available operations.

7.1 Operations and Structure

The `IROp` enum defines the instruction set. It includes arithmetic (`IR_ADD`, `IR_SUB`), control flow (`IR_IFZ`, `IR_GOTO`), and function management (`IR_CALL`, `IR_RET`).
Note: Please refer to `ir.h` for the comprehensive list of supported operations.
The instructions are stored as a linked list of `IRInst` structures:

```
typedef struct IRInst {
    IROp op;              // The operation (e.g., IR_ADD)
    char *arg1;           // First operand (e.g., "x")
    char *arg2;           // Second operand (e.g., "5")
    char *result;         // Result destination (e.g., "t0")
    struct IRInst *next;
} IRInst;
```

7.1.1 Example Translation

A high-level statement like `x = a + 5;` might be represented internally as:

```
// Represents: t0 = a + 5 (ADD a 5 t0)
create_instruction(IR_ADD, "a", "5", "t0");

// Represents: x = t0 (ASSIGN t0 x)
create_instruction(IR_ASSIGN, "t0", NULL, "x");
```

7.2 API Implementation (`ir.cpp`)

You must implement the functions declared in `ir.h` inside the `ir.cpp` file. These include:

- `create_instruction`: Allocates and initializes a new `IRInst`.
- `append_instruction`: Adds an instruction to the end of the linked list.
- `new_temp()`: Generates a unique temporary variable name (e.g., `t0`, `t1`).
- `new_label()`: Generates a unique label name (e.g., `L0`, `L1`).

8 IR Translation Guide

This section details the standard patterns you must use to translate source code into Intermediate Representation (IR).

8.1 1. Arithmetic and Assignment

Source Code:

```
int x = a + b * 5;
```

Generated IR:

```
// Perform multiplication (higher precedence)
MUL b 5 t0      // t0 = b * 5

// Perform addition
ADD a t0 t1       // t1 = a + t0

// Assign result to variable
ASSIGN t1 x
```

8.2 2. Control Flow (If-Else)

Source Code:

```
if (x > 10) {
    y = 1;
} else {
    y = 2;
}
```

Generated IR Pattern: Generated IR Pattern:

```
// Evaluate Condition
GT x 10 t0              // t0 = (x > 10)

// Branching
IFZ t0 L1               // If false, jump to L1 (Else)

// "Then" Block
ASSIGN 1 y
GOTO L0                 // Skip the Else block

// "Else" Block
LABEL L1
ASSIGN 2 y

// Convergence Point
LABEL L0
```

8.3 3. Loops

Loops require two labels: one to restart the iteration and one to exit the loop.

Source Code:

```
while (x < 5) {
    x = x + 1;
}
```

Generated IR Pattern:

```
LABEL L0                // Mark start of loop

// Evaluate Condition
LT x 5 t0               // t0 = (x < 5)

// Conditional Exit
IFZ t0 L1               // If condition false, jump out

// Loop Body
ADD x 1 t1
ASSIGN t1 x

// Jump back to start
GOTO L0

LABEL L1                // Exit label
```

8.4 4. Function Calls

Function arguments must be pushed using `PARAM` instructions *before* the `CALL` instruction.

Source Code:

```
int add(int a, int b) {
    return a+b;
}

int main() {
    int result = add(10, 20);
}
```

Generated IR:

```
FUNCTION add: // function definition
POP_PARAM b // pop the parameters off the stack in reverse order
POP_PARAM a
ADD a b t0
RETURN t0 // return instruction

FUNCTION main:
PARAM 10 // push first arg
PARAM 20 // push second arg
CALL add 2 t1  // call "add" function with 2 parameters and save the
    result in t1
ASSIGN t1 result
```

9 Generating IR (`ir_generator.cpp`)

The core logic for this phase resides in `ir_generator.cpp`. You must implement the function:

```
IRInst* generate_ir_from_ast(ASTNode* root);
```

9.1 Implementation Hints

- **Traversal:** Similar to semantic analysis, you will traverse the AST recursively. However, instead of checking for errors, you will be building a list of instructions.
- **Temporaries:** For binary operations (like `3 * x`), you cannot store the result directly in the final variable. You must generate a temporary variable using `new_temp()` to hold the intermediate result.
- **Control Flow:** For `if` and `while` nodes, you will need to generate labels using `new_label()` and insert `IR_IFZ` (Branch if Zero) and `IR_GOTO` instructions to manage the execution flow.

Note: For Part B (IR Generation), you may assume that all variable names in the source program are globally unique. You do not need to handle complex name mangling for scoping; simply using the variable name provided in the AST node (e.g., `node->data`) is sufficient.

10 Building and Testing

10.1 Running Tests

We have provided a Python automation script. To build your project and run the tests, simply execute:

```
python run_tests.py
```

This script will automatically run the `makefile` to compile your code and then execute the binary against the test suite.

10.2 Testing with Custom Inputs

You are encouraged to create your own test cases to debug specific issues or test edge cases.

1. Create a text file (e.g., `my_test.txt`) with your source code.
2. Open your terminal and run `make` to ensure your compiler is built.
3. Run the compiler manually by passing the file path as an argument:

```
./build/compiler my_test.txt
```

This will print the output of all phases (Syntax, Semantics, IR) directly to your terminal, allowing you to debug specific errors.

10.3 Platform Specifics (Windows)

If you are developing natively on Windows (using MinGW/Git Bash) rather than WSL or Linux, you must make two small modifications to handle file extensions correctly.

1. Modify `Makefile`

Windows executables require the `.exe` extension. Update the `TARGET` variable:

```
# Change:
TARGET = $(BUILD_DIR)/compiler

# To:
TARGET = $(BUILD_DIR)/compiler.exe
```

2. Modify `run_tests.py`

The Python automation script needs to know the exact name of the executable to launch it.

```
# Change:
EXECUTABLE = os.path.join("build", "compiler")

# To:
EXECUTABLE = os.path.join("build", "compiler.exe")
```

10.4 Grading Distribution

- **70% Visible Test Cases:** These tests are provided in the `tests/` folder. Passing them ensures your compiler handles standard semantic rules correctly.
- **30% Hidden Test Cases:** These test edge cases, robust error recovery, and complex scoping scenarios.

11 Submission Instructions

...