Translating programming languages for intermediate codes

Ioan Daniel HUNYADI, Mircea MUŞAN
Department of Informatics, University "Lucian Blaga" of Sibiu
daniel.hunyadi@ulbsibiu.ro, mircea.musan@ulbsibiu.ro

Many of the important data structures used in a compiler are intermediate representations of the program being compiled. Often these representations take the form of trees, with several node types, each of which has different attributes. Tree representations can be described with grammars, just like programming languages. For each grammar rule, there is one constructor that belongs to the class for its left-hand-side symbol. I simply extend the abstract class with a "concrete" class for each grammar rule. Each grammar rule has right-hand-side components that must be represented in the data structures.
Keywords*: compiler, lexical analysis, abstract syntax, intermediate representation, abstract machine language*

1 Introduction

The semantic analyses phase of a compiler must translate abstract syntax into abstract machine code. It can do this after type-checking, or at the same time.
An intermediate representation (IR) is a kind of abstract machine language that can express the target-machine operations without committing to too much machine-specific details. But it is also independent of the details of the source language. The front-end of the compiler does lexical analysis, parsing, semantic analyses, and translation to intermediate representation. The back-end does optimization of the intermediate representation and translation to machine language.
A portable compiler translates the source language into IR ant then translates the IR into machine language, as illustrated in Fig.1. Now only N front ends and M back ends are required. Such an implementation task is more reasonable.

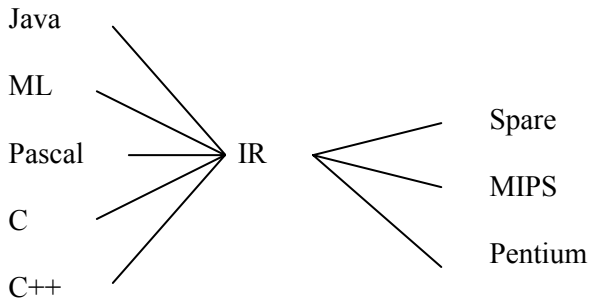


Fig.1. Compilers for four languages and three target machines with an IR.

Even when only one front-end and one back-end are being built, a good IR cam modularize the task, so that the front end is not complicated with machine-specific details, and the back-end is not bothered with information specific to one source language. Many different kinds of IR are used in compilers. For this compiler I have chosen simple expression syntax.

2. Problem Formulation

The intermediate representation tree language is defined by the package `Tree`, containing abstract classes `Stm` and `Exp` and their subclasses.
A good intermediate representation has several qualities:

- It must be convenient for the semantic analyses phase to produce.
- It must be convenient to translate into real machine language, for all the desired target machines.
- Each construct must have a clear and simple meaning, so that optimizing transforma-

tions that rewrite the intermediate representation can easily be specified and implemented. Individual pieces of abstract syntax can be complicated things, such as array subscripts, procedure calls, and so on. And individual "real machine" instructions can also have a complicated effect. Unfortunately, it is not always the case that complex pieces of the abstract syntax correspond exactly to the complex instructions that a machine can execute.

3. Problem Solution

Therefore, the intermediate representation should have individual components that describe only extremely simple things: a single fetch, store, add, move, or jump. Then any "chunky" piece of abstract syntax can be translated into just the right set of abstract machine instructions.

```
package Tree;
abstract class Exp
CONST(int value)
NAME(Label label)
TEMP(Temp.Temp temp)
BINOP(int binop, Exp left,Exp right)
MEM(Exp exp)
CALL(Exp func, ExpList args)
ESEQ(Stm stm, Exp exp)

abstract class Stm
MOVE(Exp dst, Exp src)
EXT(Exp exp)
JUMP(Exp exp,Temp.LabelList targets)
CJUMP(int rel,Exp left,Exp right, Label
iftrue,Label iffalse)
SEQ(Stm left, Stm right)
LABEL(Label label)
```

Here is a description of the meaning of each tree operator. First, the expression (Exp), which stand for the computation of some value (possibly with side effects):

`CONST(i)` – The integer constant *i*.

`NAME(n)` – the symbolic constant *n* (corresponding to an assembly language label)

`TEMP(t)` – Temporary *t*. A temporary in the abstract machine is similar to a register in a real machine. However, the abstract machine has an infinite number of temporaries.

`BINOP(o,e1,e2)` – The application of binary operator *o* to operands *e1*, *e2*. Subexpression *e1* is evaluated before *e2*. The integer arithmetic operator are PLUS , MINUS, MUL, DIV; the integer bitwise logical operators are AND, OR, XOR; the integer logical shift operators are LSHIFT, RSHIFT; the integer arithmetic right-shift is ARSHIFT.

`MEM(e)` – The content of *wordSize* bytes of memory starting at address *e* (where *wordSize* is defined in the Frame module). Note that when MEM is used as the left child of a MOVE, it means "store", but anywhere else it means "fetch".

`CALL(f,l)` – A procedure call: the application of function *f* to argument list *l*. The subexpression *f* is evaluated before the arguments which are evaluated left to right.

`ESEQ(s,e)` – The statement *s* is evaluated for side effects, then *e* is evaluated for a result.

The statements (`stm`) of the tree language perform side effects and control flow:

MOVE(TEMP t,e) – Evaluate *e* and move it into temporary *t*.

MOVE(MEM(e1),e2) – Evaluate *e1*, yielding address *a*. The evaluate *e2*, and store the result into *wordSize* bytes of memory starting at *a*.

EXP(e) – Evaluate *e* and discard the results.

JUMP(e,labs) – Transfer control (jump) to address *e*. The destination *e* may be a literal label, as in NAME (*lab*), or it may be an address calculated by any other kind of expression. For example, a C-language `switch(i)` statement may be implemented by doing arithmetic on *i*. The list of labels *labs* specifies all the possible location that the expression *e* can evaluate to; this is necessary for dataflow analysis later.

CJUMP(o,e1,e2,t,f) – Evaluate *e1, e2* in that order, yielding values *a, b*. Then compare *a, b* using the relational operator *o*. If the result is `true`, jump to *t*; otherwise jump to *f*.

SEQ(s1,s2) – The statement *s1* followed by *s2*.

LABEL(n) – Define the constant value of name *n* to be the current machine code address. This is like a label definitions in assembly language. The value NAME(n) may be the target of jumps, calls, etc.

It is almost possible to give a formal semantic to the `Tree` language. However, there is no provision in this language for procedure

and function definitions – we can specify only the body of each function. The procedure entry and exit sequences will be added later as special "glue" that is different for each target machine.

Translation of abstract syntax expressions into intermediate tree is reasonably straightforward; but there are many cases to handle. I will cover the translation of various language construct, including many from MiniJava.

The MiniJava grammar has clearly distinguished statements and expression. In languages such as C, the distinction is blurred. For example, an assignment in C can be used as an expression. When translating such languages, we will have to ask the following question. What should the representation of an abstract syntax expression be in `Tree` language? At first it seems obvious that it should be `Tree.Exp`. This is true only for certain kind of expressions, the ones that compute a value. Expressions that return no value are more naturally represented by `Tree.Stm`. And expressions with boolean values, such as `a>b`, might best be represented as an conditional jump – a combination of `Tree.Stm` and a pair of destinations represented by `Temp.Labels`.

It is better instead to ask, "how might the expression be used?" Then I can make the right kind of methods for an object-oriented interface to expressions. Both for MiniJava and other languages, I end up with `Translate.Exp`, not the same class as `Tree.Exp`, having three methods:

```
package Translate;
public abstract class Exp{
             abstract Tree.Exp unEx();
abstract Tree.Stm unNx();
     abstract Tree.Stm unCx
      (Temp.Label t, Temp.Label f);
  }
```

`Ex` – stands for an "expression", represented as a `Tree.Exp`.

`Nx` – stands for "no result", represented as a `Tree` statement.

`Cx` – stands for "conditional", represented as a function from label-pair to statement.

For example, the MiniJava statement

```
if (a<b && c<d)
{ //true block }
else
{ //false block }
```

might translate to a Translate.Exp whose unCx method is roughly like

```
Tree.Stm unCx(Label t, Label f)
{ Label z=new Label();
  return new SEQ(new
    CJUMP(CJUMP.LT,a,b,z,f), new
    SEQ(new LABEL(z),new
    CJUMP(CJUMP.LT,c,d,t,f)));
}
```

The abstract class `Translate.Exp` can be instantiated by several subclasses: `Ex` for an ordinary expression that yields a single value, `Nx` for an expression that yields no value, and `Cx` for a "conditional" expression that jumps to either *t* or *f*:

```
class Ex extends Exp{
  Tree.Exp exp ;
  Ex(Tree.Exp e) {exp=e;}
  Tree.Exp unEx() {return exp;}
  Tree.Stm unNx() {...?...}
  Tree.Stm unCx(
        Label t, Label f) {...?...}
}
class Nx extends Exp {
  Tree.Stm stm;
  Nx(Tree.Stm s) {stm=s;}
  Tree.Exp unEx() {...?...}
  Tree.Stm unNx() {return stm;}
  Tree.Stm unCx(
        Label t, Label f) {...?...}
}
abstract class Cx extends Exp{
Tree.Exp unEx(){
   Temp r=new Temp();
   Label t=new Label();
   Label f=new Label();
}
abstract Tree.Stm unCx(Label t,
   Label f);
Tree.Stm unNx(){...}
}
```

Each kind of `Translate.Exp` class must have similar conversion methods. The `unCx` method is still abstract. But the `unEx` and `unNx` methods can still be implemented in terms of the `unCx` method.

3.1. Structured *l*-values

An *l*-value is the result of an expression that can occur on the *left* of an assignment statement. An *r*-value is one that can only appear on the right of an assignment. That is, an *l*-value denotes a *location* that can be assigned to, and an *r*-value does not. Of course, an *l*-value can occur on the right of an assignment statement. In this case the contents of the location are implicitly taken.

All the variables and *l*-values in MiniJava are scalar, since it has only one component. Even a MiniJava array or object variable is really a pointer.

In C or Pascal there are structured *l*-value – structs in C, records in Pascal - that are not scalar. In a C compiler, the `access` type would require information about the size of the variables. Then, the `MEM` operator of the `TREE` intermediate language would need to be extends with a notation of size:

```
package Tree;
abstract class Exp
 MEM(Exp exp, int size)
```

The translation of a local variable into an IR tree would look like

```
MEM(+(TEMP fp, CONST kn),S)
```

where the *S* indicates the size of the object to be fetched or stored (depending on whether this tree appears on the left or right of a MOVE).

Leaving out the size on MEM nodes makes the MiniJava compiler easier to implement, but limits the generality of its intermediate representation.

3.2. Subscripting and field selection

To select field *f* of a record *l*-value *a* (to calculate *a.f*), simply add the constant field offset of *f* to the address *a*.

An array variable *a* is an *l*-value; so is an array subscript expression *a[i]*, even if *i* is not an *l*-value. To calculate the *l*-value a[i] from *a*, we do arithmetic on the address of *a*. Thus, in a Pascal compiler, the translation of an *l*-value (particularly a structured *l*-value) should not be something like in Fig. 2, but should instead be the `Tree` expression representing the base address of the array (Fig. 3).

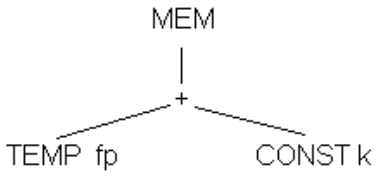


Fig.2.

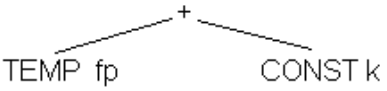
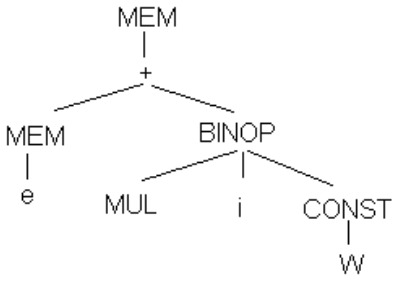


Fig.3.

In the MiniJava Language, there are no structured, or "large" *l*-value. This is because all object and array values are really pointers to object and array structures. The "base address" of the array is really the contents of a pointer variable, so `MEM` is required to fetch this base address.

Thus, if *a* is a memory-resident array variable represented as `MEM(e)`, then the contents of address *e* will be a one-word pointer value *p*. The contents of addresses *p, p+W, p+2W,...,* will be the elements of the array. Thus, *a[i]* is just *l*-values and MEM nodes, like in Fig. 4.




```
MEM{+(MEM(e),
BINOP(MUL,i,CONST(W)))
```

Fig.4.

Technically, an *l*-value should be represented as an address (without the top MEM node in the diagram above). Converting an *l*-value to an *r*-value (when it is used an expression) means *fetching* from that address. Assigning to an *l*-value means *storing* to that address. We are attaching the MEM node to the *l*-value before knowing whether it is to be fetched or stored. This works only because in the `Tree` intermediate representation, MEM

means both store (when used as the left child of a MOVE) and fetch (when used elsewhere).

3.3 Conditionals

The result of a comparison operator will be a `Cx` expression: a statement *s* that will jump to any true-destination and false-destination you specify.

Making "simple" `Cx` expression from comparison operators is easy with the `CJUMP` operator. However, the whole point of the `Cx` representation is that conditional expressions can be combined easily with the MiniJava operator &&. Therefore, an expression such as x<5 will be translated as `Cx(`s_1`)`, where

$$s_1(t, f) = CJUMP(LT, x, CONST(5), t, f)$$

for any labels *t* and *f*.

To do this, I extend the `Cx` class to make a subclass `RelCx` that has private fields to hold the left and right expressions (in this case *x* and 5) and the comparison operator (in this case `Tree.CJUMP.LT`). Then we override the `unCx` method to generate the `CJUMP` from these data. It is not necessary to make `unCx` and `unNx` methods, since these will be inherited from the parent `Cx` class.

The most straightforward thing to do with an *if* expression

if *e1* **then** *e2* **else** *e3*

is to treat *e1* as a `Cx` expression, and *e2* and *e3* as `Ex` expression. That is, use the `unCx` methos of *e1* and the `unEx` of *e2* and *e3*. Make two labels *t* and *f* to which the conditional will branch. Allocate a temporary *r*, and after label *t*, move *e2* to *r*. After label *f*, move *e3* to *r*. Both branches should finish by jumping to a newly created "join" label.

This will produce perfectly correct result. However, the translated code may not be very efficient at all. If *e2* and *e3* are both "statements" (expressions that return no value), then their representation is likely to be `Nx`, not `Ex`. Applying `unEx` to them will work – a coercion will automatically be applied – but it might be better to recognize this case specially.

The translation of an **if** requires a new subclass of `Exp`:

```
class IfThenElseExp extends Exp{
  Exp cond,a,b;
  Label t=new Label();
  Label f=new Label();
  Label join=new Label();
  IfThenElseExp(Exp cc, Exp aa,
           Exp bb)
  {
     cond=cc; a=aa;  b=bb;
  }
  Tree.Stm unCx(Label tt, Label ff)
  {…}
  Tree.Exp unEx(){…}
  Tree.Stm unNx(){…}
}
```

The labels *t* and *f* indicate the beginning of the then-clause and else-clause, respectively. The labels *tt* and *ff* are quite different: these are the places to which conditions inside then-clause (or else-clause) must jump, depending on the truth of those subexpressions.

4. Conclusion

To simplify the implementation of the translator, we may do without the `Ex, Nx, Cx` constructors. The entire translation can be done with ordinary value expression. This means that there is only one `Exp` class. This class contains one field of type `Tree.Exp` and only an `unEx()` method. Instead of `Nx(s)`, use `Ex(EXEQ(s.CONST 0))`. For conditionals, instead of a Cx, use an expression that just evaluates to 1 or 0.

The intermediate representation trees produced from this kind of naïve translation will be bulkier and slower than a "fancy" translation. But they *will* work correctly, and in principle a fancy back-end optimizer might be able to clean up the clumsiness. In any case, a clumsy but correct translator is better than a fancy one that doesn't work.

References:

[CHAL95] Chambers C., Leavens G.T., Typechecking and modules for multimethods, *ACM Trans. on Programming Languages and Systems 17(6), 1995, pp.805-843*

[CHET94] Chen W., Turau B., Efficient dynamic look-up strategy for multi-methods, *European Conference an Object Oriented Programming (ECOOP '94),* 1994

[LIMA97] Lipton R.J., Martino P.J., On the complexity of a set-union problem, *Proc. 38th Annual Symposium on Foundations of Computer Science*, IEEE Computer Society Press 110-115, 1997

[STRO97] Stroustrup P.B., Programming Language, *Third ed. Addison-Wesley* , Reading, MA, 1997