Attribute Grammars and their Applications

Krishnaprasad Thirunarayan
Wright State University, USA

INTRODUCTION

Attribute grammars are a framework for defining semantics of programming languages in a syntax-directed fashion. In this chapter, we define attribute grammars, and then illustrate their use for language definition, compiler generation, definite clause grammars, design and specification of algorithms, and so forth. Our goal is to emphasize its role as a tool for design, formal specification and implementation of practical systems, so our presentation is example rich.

BACKGROUND

The lexical structure and syntax of a language is normally defined using regular expressions and context-free grammars respectively (Aho, Lam, Sethi & Ullman et al., 2007,). Knuth (1968) introduced attribute grammars to specify static and dynamic semantics of a programming language in a syntax-directed way.

Let $G = (N, T, P, S)$ be a context-free grammar for a language L (Aho et al., 2007). N is the set of non-terminals. T is the set of terminals. P is the set of productions. Each production is of the form A::= α, where A $\in$ N and $\alpha \in$ (N U T)*. S $\in$ N is the start symbol. An *attribute grammar* AG is a triple (G, A, AR), where G is a context-free grammar for the language, A associates each grammar symbol X $\in$ N U T with a set of attributes, and AR associates each production R $\in$ P with a set of attribute computation rules (Paakki, 1995). A(X), where X $\in$ (N U T), can be further

partitioned into two sets: synthesized attributes S(X) and inherited attributes I(X). AR(R), where R $\in$ P, contains rules for computing inherited and synthesized attributes associated with the symbols in the production R.

Consider the following attribute grammar that maps bit strings to numbers. CFG = ({N},{0,1},P,N), where P is the left column of productions shown below. The number semantics is formalized by associating a synthesized attribute *val* with N, and providing rules for computing the value of the attribute *val* associated with the left-hand side N (denoted N_l) in terms of the value of the attribute *val* associated with the right-hand side N (denoted N_r), and the terminal.

N ::= 0	N.*val* = 0
N ::= 1	N.*val* = 1
N ::= N0	N_l.*val* = 2 * N_r.*val*
N ::= N1	N_l.*val* = 2 * N_r.*val* + 1

An attribute grammar involving only synthesized attributes is called an *S-attributed grammar*. It is straightforward to parse a binary string using this grammar and then compute the value of the string using a simple top-down left-to-right traversal of the abstract syntax tree.

The above attribute grammar is not unique. One can construct a different S-attributed grammar for the same language and the same semantics as follows.

N ::= 0	N.*val* = 0, N.*len* = 1
N ::= 1	N.*val* = 1, N.*len* = 1
N ::= 0N	N_l.*val* = N_r.*val*;
	N_l.*len* = N_r.len +1
N ::= 1N	N_l.*val* = $2^{N_r.len}$ + N_r.*val*;
	N_l.*len* = N_r.len +1

Attribute grammars can be devised to specify different semantics associated with the same language. For instance,

the bit string can be interpreted as a fraction by associating an inherited attribute *pow* to capture the left context—the length of the bit string between left of a non-terminal and the binary point, to determine the local value or the weight of the bit.

```
F ::= . N     F.val = N.val, N.pow  = 1
N ::= 0       N.val = 0
N ::= 1       N.val = (1 / 2^ N.pow)
N ::= 0N      Nₗ.val = Nᵣ.val;
              Nᵣ.pow = Nₗ.pow +1
N ::= 1N      Nₗ.val = (1 / 2^ N.pow) + Nᵣ.val;
              Nᵣ.pow = Nₗ.pow +1
```

Each production is associated with attribute computation rules to compute the synthesized attribute *val* of the left-hand side non-terminal and the inherited attribute *pow* of the right-hand side non-terminal (we leave it as an exercise for the interested reader to devise an S-attributed grammar to capture this semantics).

APPLICATIONS OF ATTRIBUTE GRAMMARS

Attribute grammars provide a *modular framework* for formally specifying the semantics of a language based on its context-free grammar (or in practice, for conciseness, on its Extended Backus Naur formalism representation (Louden, 2003)). By *modular*, we emphasize its role in structuring specification that is incremental with respect to the productions. That is, it is possible to develop and understand the attribute computation rules one production at a time. By *framework*, we emphasize its role in structuring a specification, rather than conceptualizing the meaning. For instance, denotational semantics, axiomatic semantics, and operational semantics of a language can all be specified using the attribute grammar formalism by appropriately choosing the attributes (Louden, 2003). In practice, different programming languages interpret the same syntax/construct differently, and attribute grammars provide a framework for defining and analyzing subtle semantic differences.

In this section, we illustrate the uses and the subtleties associated with attribute grammars using examples of contemporary interest. Traditionally, attribute grammars have been used to specify various compiler activities formally. We show examples involving (i) type checking/inference (static semantics), (ii) code generation, and (iii) collecting distinct variables in a straight-line program. Attribute grammars can also be used to specify compiler generator activities. We show parser generator examples specifying the computation of (i) nullable non-terminals, (ii) first sets, and (iii) follow sets, in that sequence (Aho et al., 2007). Definite clause grammars enable attribute grammars satisfying certain restrictions to be viewed as executable specifications (Bratko, 2001). We exemplify this in SWI-Prolog. The essence of attribute grammars can be seen to underlie several database algorithms. We substantiate this by discussing the magic sets for optimizing bottom-up query processing engine. We also discuss how attribute grammars can be used for developing and specifying algorithms for information management (Thirunarayan, Berkovich, & Sokol, 2005).

Type Checking, Type Inference, and Code Generation

Consider a simple prefix expression language containing terminals {n, x, +}. The type of variable n is int, and the type of variable x is double. The binary arithmetic operation + returns an int result if both the operands are int, otherwise it returns a double. The type of a prefix expression can be specified as follows.

```
E ::= n       E.typ = int
E ::= x       E.typ = double
E ::= + E E            E.typ = if Eᵣ₁.typ = Eᵣ₂.typ
                       then Eᵣ₁.typ else double
```

The corresponding executable specification in Prolog can be obtained by defining a binary relation 'typ' between prefix expression terms and their types as follows. Observe that each line of specification that ends in a "." is an axiom (first two are Prolog facts, while last two are Prolog rules), E, F, T, T1, and T2 are universally quantified Prolog variables, ":-" stands for *logical if*, and "," stands for *logical and*.

```
typ(i,int).
typ(x,double).
typ(+(E,F),T) :- typ(E,T), typ(F,T).
typ(+(E,F),real) :- typ(E,T1), typ(F,T2),  T1 \= T2.
```

A type checking query '?- typ(+(n,x),int).' verifies if the expression '+(n,x)' is of type int, while a type inference query '?- typ(+(n,x),T).' determines the type of the expression '+(n,x)'.

Attribute grammar specifying the translation of an equivalent expression language containing infix + into Java bytecode in the context of the instance method definition: 'class { double f(int n, double i) { return E;}}' is as follows:

```
E ::= n            E.code = [iload_1]
E ::= x            E.code = [dload_2]
E ::= E + E        E.code = if Eᵣ₁.typ = int
                   then if Eᵣ₂.typ = int
                       then Eᵣ₁.code@Eᵣ₂.code@[iadd]
                       else Eᵣ₁.code@[i2d]@Eᵣ₂.code@[dadd]
                   else if Eᵣ₂.typ = int
                       then Eᵣ₁.code@Eᵣ₂.code@[i2d,dadd]
                       else Eᵣ₁.code@Eᵣ₂.code@[dadd]
```

The attribute *typ* has been specified earlier. The attribute *code* is bound to a list of Java bytecode. '@' refers to list append operation. Java compiler maps the formal parameters n and x to register 1, and register pair 2 and 3, respectively.

(The double value requires two registers.) iload_1 (dload_2) stands for pushing the value of the int n (double x) on top of the stack; dadd (iadd) stands for popping the top two double (int) values from the stack, adding them, and pushing the result on top of the stack; and i2d stands for coercing an int value into a double value (Lindholm & Yellin, 1999). (Note that + is left associative in Java.) In practice, the code generator has to cater to variations on whether the method is static or instance, whether the formal arguments require 4 or more registers, whether the arguments are of primitive types or reference types, etc, and all this can be made explicit via attribute grammars.

Collecting Distinct Identifiers

We use the example of collecting distinct identifiers in an expression to illustrate the influence of primitive data types available for specifying the semantics on the ease of writing specifications, and the rules of thumb to be used to enable sound and complete attribute computation in one-pass using top-down left-to-right traversal of the abstract syntax tree.

<exp> ::= <var> | <exp> + <exp>

Suppose we have ADT SET available to us as a primitive. We associate synthesized attributes *id* and *ids* with <var> and <exp> respectively, to obtain the following attribute grammar. ('∪' refers to set-union.)

<exp> ::= <var>   <exp>.*ids* = { <var>.*id* }
<exp> ::= <exp> + <exp>   <exp>.*ids* = <exp>.*ids* ∪ <exp>.*ids*

Instead, if we have only ADT LIST available to us, we associate synthesized attribute *envo* and inherited attribute *envi* of type list of symbols with <exp>, to obtain the following attribute grammar.

<exp> ::= <var>   <exp>.*envo* = if <var>.id ∈ <exp>.*envi*
                           then <env>.*envi*
                           else cons(<var>.*id*, <env>.*envi*)
<exp> ::= <exp> + <exp>   <exp$_{r1}$>.*envi* = <exp$_l$>.*envi*
                           <exp$_{r2}$>.*envi* = <exp$_{r1}$>.*envo*
                           <exp$_l$>.*envo* = <exp$_{r2}$>.*envo*

Observe that, given the definition of the attributes and the production rule, one can *automatically* determine the left-hand sides of the attribute computation rules required. *There is one rule for each synthesized attribute of the left-hand side non-terminal, and one rule for each inherited attribute of the right-hand side symbol (non-terminal).*

<exp>   ::= <exp>   +   <exp>
↓ envi   ↓ *envi*   ↓ *envi*
↑ *envo*   ↑ envo   ↑ envo

For the above production, the attributes shown in italics need to be determined using the attributes given in italics.

To enable one-pass top-down left-to-right computation of the attributes, each inherited attribute of the right-hand side symbol can depend on all the attributes associated with preceding right-hand side symbols and the inherited attribute of the left-hand side non-terminal. Similarly, the synthesized attribute of the left-hand side non-terminal can depend on all the attributes associated with all the right-hand side symbols and the inherited attribute of the left-hand side non-terminal. Effectively, the inherited attribute associated with the left-hand side non-terminal provides context information from the parent and the left siblings, while the synthesized attributes of the right-hand side non-terminals provide information from their descendants.

From modularity perspective, if the attribute computation rules associated with each production satisfies the above constraints, it can be argued using induction principle that all the attributes associated with each node in the abstract syntax tree will be well-defined after one-pass of computation.

In practical compiler construction, a programming language is specified to a parser generator as an S-attributed grammar that transforms a program into its abstract syntax tree. This abstract syntax tree is traversed multiple times for semantic analysis and code generation. The APIs of the meta-language provide primitive functions and data types for defining and implementing semantic actions (attribute computations). For example, bottom-up parser generator Bison is based on C/C++ libraries, top-down parser generator ANTRL is based on Java APIs, and so forth.

Definite Clause Grammars

Prolog's definite clause grammars can be used to execute the set-based and the list-based attribute grammars discussed earlier. Each DCG rule resembles a production with the predicate name corresponding to the non-terminal and the formal arguments corresponding to the attributes. The input string is encoded as a list of symbols (tokens). Upon loading, the DCG rules are automatically translated into ordinary Prolog rules using difference-list implementation (Bratko, 2001). As a result, the predicate exp in the query has two additional arguments than are found in the DCG specification. The semantic action code inside curly braces incorporates calls to SWI-Prolog library functions.

The DCG corresponding to the set-based attribute grammar is as follows.

```
exp(Ids)   →   aexp(Ids).
exp(Ids)   →   aexp(Ids1), ['+'] , exp(Ids2), {union(Ids1,Ids2,Ids)}.
aexp([Id]) →   [Id], {atom(Id)}.

/* ?- exp(Vs, [x, '+', y, '+', z, +, y],[]). */
/* Vs = [x, z, y] ; */
```

The DCG corresponding to the list-based attribute grammar is as follows.

```
exp(Envi,Envo)   →  aexp(Envi,Envo).
```

```
exp(Envi,Envo)        →      aexp(Envi,EnvT), ['+'] , exp(EnvT,Envo).
aexp(Envi,Envo)       →      [Id], {atom(Id)},
                             {member(Id,Envi) -> Envo = Envi;
                                    Envo = [Id | Envi]}.
```

/* ?- exp([],Vs, [x, '+', y, '+', z, +, y],[]). */

/* Vs = [z, y, x] ; */

The sample queries and the results have been commented out. Refer to (Bratko, 2001) for DCG details.

Specifying Compiler Generator Operations

Attribute grammars can be used to formalize and implement parser generators. For instance, consider the computation of nullable non-terminals, first-sets associated with non-terminals and follow-sets associated with non-terminals. A non-terminal is *nullable* if it can derive a null string. The *first-set* associated with a non-terminal is the set of tokens that can begin a string derivable from the non-terminal. The *follow-set* associated with a non-terminal is the set of tokens that can come after the non-terminal in a sentential form. These can be specified for an expression grammar along the lines indicated below. (Only partial specification has been given.)

```
S ::= T E
E ::= ε | + S
T ::= x | y
```

S.*nullable* = T.*nullable* and E.*nullable*
E.*nullable* = true
T.*nullable* = false

S.*first-set* = if T.*nullable* then T.*first-set* $\cup$ E.*first-set* else T.*first-set*
E.*first-set* = {ε , +}
T.*first-set* = {x , y}

T.*follow-set* = if E.*nullable* then S.*follow-set* $\cup$ E.*first-set* else E.*first-set*
S.*follow-set* = E.*follow-set*
E.*follow-set* = S.*follow-set*

The dependencies among the various attributes can be exploited to sequence their computation in three phases: compute *nullable* first, followed by *first-sets*, followed by *follow-sets* (Bochmann, 1976). Each phase can potentially require multiple iterations to converge to a fixed-point. The *nullable* is a synthesized attribute and requires multiple bottom-up pass. The *first-set* is also a synthesized attribute computed using multiple bottom-up left-to-right pass. The *follow-set* is an inherited attribute computed using multiple top-down right-to-left pass.

Optimizing Bottom-up Database Query Evaluation

Top-down query evaluation and bottom-up query evaluation are two well-known deductive database query implementation strategies (Ramakrishnan & Sudarshan, 1991). In top-down approach, the query solution tree is grown from the root (goal) to the leaves (data), applying the datalog (function-free Horn-logic or Prolog) rules from left to right. It is potentially incomplete in the presence of left-recursive rules, but it is efficient because the bindings in the queries are propagated to the base relations (data). In bottom-up approach, the query solution tree is grown from the leaves (data) to the root (goal), applying the datalog rules from right to left. It is complete, but can be inefficient because the search is not goal directed. Memoing techniques can be used to improve top-down strategy by making its search complete, while magic predicates can be employed to improve bottom-up strategy by making it more focused and efficient (Beeri & Ramakrishnan, 1987). The ideas underlying the definition of magic predicates and sideways information processing are reminiscent of attribute computation rules involving inherited and synthesized attributes as explained and illustrated in the following.

Consider the definition of the ancestor relation based on the parent relation, and the ancestor query with the first argument bound.

```
ancestor(X,Y) :- parent(X,Y).
ancestor(X,Y) :- parent(X,Z), ancestor(Z,Y).
?- ancestor(john, N).
```

In order to explore only a small fragment of the necessary parent facts to compute the query answers using naïve bottom-up approach, it is important to restrict the "firing" of the second rule. This can be done by transforming the second rule using *magic* predicates so that it is satisfied only for the ancestors of john as follows.

```
magic(john).
magic(Z) :- magic(X), parent(X,Z).
ancestor(X,Y) :- magic(X), parent(X,Y).
ancestor(X,Y) :- magic(X), parent(X,Z), ancestor(Z,Y).
```

The rules for *magic* predicate are analogous to the computation of attributes. The first fact corresponds to the inherited attribute value corresponding to the first argument of ancestor initialized by the top-level query. The second rule corresponds to the computation of additional inherited attribute values corresponding to the first argument of ancestor obtained through sideways information passing of the synthesized attribute from the parent relation. Beeri and Ramakrishnan (1987) provide several additional examples embodying attribute grammars ideas for query optimization.

Specifying Algorithms for Customized Information Extraction

In this section, we discuss an information extraction problem of industrial significance that has benefited from attribute grammar based algorithm specification. We skip the detailed specification *per se* due to space constraints.

A typical materials and process spec (from authoring organizations such as GE, Pratt and Whitney, ASTM, AMS, and so forth) contains requirements for making and testing a variety of alloys. A typical customer order specifies the desired material in terms of the specs it must conform to, and a collection of domain-specific product parameters such as product type, spec class, product dimension and cross-section, etc. A fundamental operation of interest on a spec is the determination of applicable fragments of the spec for a customer order. The goal of coarse-grain extraction is to convert a spec into a form that can be evaluated against the order parameters to determine applicable fragments.

To balance the commercial viability of the extraction task and its tractability, a spec is transformed into a possibly, nested sequence of conditioned notes of the form "If CONDITIONS Then [Note = " ... "]", where the note contains contiguous block of spec text. These extractions are cheap to produce because the detailed requirements are still in text.

In order to formalize conditioned notes, we need to propose a structure for the conditional expression and the note it qualifies. The conditional expression can be formed using characteristic names (e.g., spec class, alloy, product type, etc.), constants, relational symbols (e.g., `=`, `$>$`, etc.), and boolean connectives (e.g., `and`, `or`, etc) in the standard way. The note can be defined in quanta of (sub)-sections and paragraphs. Thus, there are two important technical problems to be solved for carrying out extraction: (1) identification of the values of a characteristic that can appear as a condition, and (2) transformation of the relevant spec text into a sequence of conditioned notes.

The conditioned notes can be specified in terms of the scope rules of applicability of characteristic-value pairs to the spec text fragments. For instance, to associate a spec class with a section or a paragraph, we use the following heuristic: Every (sub-)section is conditioned on all spec classes named in section 'scope.' Explicit spec class references in a paragraph override the default condition. Otherwise, a paragraph inherits the condition from its left sibling (earlier paragraph), or transitively from its parent (enclosing (sub-)section). The rationale behind the heuristic is that, when the conditionals in an extraction are evaluated against the given condition values, it should generate all applicable fragments of the spec.

To abstract the algorithmic details of the extraction, the structure of a spec can be captured using EBNF as:

```
<document> ::= <document-header> <section>+
<section> ::= <sectionNumber> <sectionHeading> <paragraph>+ <section>*
```

A

and the computation of the conditioned notes involving spec classes can be given using attribute grammars. There are many other qualifiers of interest besides spec classes such as products, product types, alloys, and so forth. (See Thirunarayan et al., 2005 for a detailed attribute grammar specification.)

FUTURE TRENDS

Information flow ideas underlying attribute grammars can provide a general framework for designing and specifying algorithms. For example, Neven (2005), and Neven and Den Bussche (2002) demonstrate the influence of attribute grammars on query languages.

Web technologies such as XML/XSLT that are based on adorned context-free grammars can benefit from techniques developed for attribute grammars (Harold, 2004). Conceptually, an XML document consists of annotations, where each annotation consists of an associated XML-element that reflects the semantic category to which the corresponding text fragment belongs, and the associated XML-attributes that are bound to relevant semantic values. Overlaying domain-specific XML tags on a text document enables abstraction, formalization, and in-place embedding of machine-processable semantics. In the future, we can expect the annotated data to be interpreted by viewing it as a function/procedure call, and defining the XML-element as a function/procedure in a language such as XSLT or Water for associating different collections of behaviors with XML-elements (Thirunarayan, 2005).

CONCLUSION

Historically, attribute grammars were developed in the context of compiler construction. We provided several examples illustrating the application of attribute grammars for static analysis of programs, for program translation, and for specifying certain phases of a parser generator. We also provided general principles for developing attribute grammar specifications for efficient one-pass computation of attributes. We introduced Prolog's definite clause grammars to enable attribute grammars to be viewed as executable specifications. We showed how the information flow ideas implicit in the attribute computation rules can be exploited to optimize bottom-up evaluation of datalog programs. Finally, we discussed the application of attribute grammars for specifying information extraction algorithms, and its future role in XML technologies.

REFERENCES

Aho, A. V., Lam, M., Sethi, R., & Ullman, J. D. (2007). *Compilers: principles, techniques, and tools*. Addison Wesley.

Beeri, C., & Ramakrishnan, R. (1987). On the Power of Magic, *Proceedings of the 6th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, (pp. 269-283).

Bratko, I. (2001). *Prolog Programming for Artificial Intelligence*, (3rd ed.). Addison Wesley.

Bochmann, G. V. (1976). Semantic evaluation from left to right. *Communications of the ACM, 19*(2), 55-62.

Harold, E. R. (2004). *XML in a nutshell,* (3rd ed.). O'Reilly.

Knuth, D. E. (1968). Semantics of context-free languages. *Mathematical Systems Theory, 2*(2), 127-145. (Corrigenda: *Mathematical Sysems Theory 5*(1),1971, 95-96.)

Lindholm, T., & Yellin, F. (1999). *Programming Java virtual machine specification,* (2nd ed.). Addison-Wesley.

Louden, K. C. (2003). *Programming languages: Principles and practice*, (2nd ed.). Thomson – Course Technology.

Neven, F. (2005). Attribute grammars for unranked trees as a query language for structured documents. *Journal of Computer and System Sciences, 70,* 221-257.

Neven, F., & Den Bussche, J. V. (2002). Expressiveness of structured document query languages based on attribute grammars. *Journal of the ACM*, *49*(1), 56-100.

Paakki, J. (1995). Attribute grammar paradigms—a high-level methodology in language implementation, *ACM Computing Surveys, 27*(2)*,* 196-255.

Ramakrishnan, R., & Sudarshan, S. (1991). Top-Down versus Bottom-Up Revisited. *In Proceedings of the 1991 International Symposium on Logic Programming*, (pp. 321-336).

Thirunarayan, K. (1984). *A Compiler-Generator Based on Attributed Translation Grammars*, M. E. Thesis, Indian Institute of Science, Bangalore. (Advisors: Priti Shankar and Y. N. Srikant)

Thirunarayan, K. (2005). On embedding machine-processable semantics into documents. *IEEE Transactions on Knowledge and Data Engineering*, *17*(7), 1014-1018.

Thirunarayan, K., Berkovich, A., & Sokol, D. (2005). An information extraction approach to reorganizing and summarizing specifications. *Information and Software Technology Journal*, *47*(4), 215-232, 2005.

KEY TERMS

Attribute computation rules: Rules used to compute attributes, usually defined in terms of other attributes and standard primitives.

Attribute grammar: An attribute grammar is an extension of context-free grammar that enables definition of context-sensitive aspects of a language and its translation.

DCG: A definite clause grammar is a Prolog built-in mechanism for implementing attribute grammars efficiently using difference lists.

EBNF: Extended Backus Naur formalism is an extension of context-free grammar with regular expression operations for defining context-free languages. It provides a more concise syntax specification.

Inherited attributes: These attributes pass information from root to the leaves of a parse tree, or sideways among siblings.

Machine-processable semantics: Metadata added to the documents to enable machines to understand and reason with text or multi-media content.

Synthesized attributes: These attributes pass information from leaves to the root of a parse tree.

XML/XSLT: Extensible markup l,anguage is a meta-language for creating markup languages. XML is a subset of SGML. XHTML is an instance of XML. Extensible stylesheet language transformations is a language for manipulating XML documents.