Introduction to Functional Programming

John Harrison

University of Cambridge

Lecture 10

ML examples II: Recursive Descent Parsing

Topics covered:

- The parsing problem
- Recursive descent
- Parsers in ML
- Higher order parser combinators
- Efficiency and limitations.

Grammar for terms

We would like to have a parser for our terms, so that we don't have to write them in terms of type constructors.

$$
\begin{array}{rcl}
\mathit{term} & \longrightarrow & \mathit{name}\texttt{(}\mathit{termlist}\texttt{)} \\
 & | & \mathit{name} \\
 & | & \texttt{(}\mathit{term}\texttt{)} \\
 & | & \mathit{numeral} \\
 & | & \texttt{-}\mathit{term} \\
 & | & \mathit{term}\ \texttt{+}\ \mathit{term} \\
 & | & \mathit{term}\ \texttt{*}\ \mathit{term} \\
\mathit{termlist} & \longrightarrow & \mathit{term}\texttt{,}\mathit{termlist} \\
 & | & \mathit{term}
\end{array}
$$

Here we have a grammar for terms, defined by a set of production rules.

Ambiguity

The task of *parsing*, in general, is to reverse this, i.e. find a sequence of productions that could generate a given string.

Unfortunately the above grammar is *ambiguous*, since certain strings can be produced in several ways, e.g.

$$\begin{array}{rcl} term & \longrightarrow & term \; \texttt{+} \; term \\ & \longrightarrow & term \; \texttt{+} \; term \; \texttt{*} \; term \end{array}$$

and

$$\begin{array}{rcl} term & \longrightarrow & term \; \texttt{*} \; term \\ & \longrightarrow & term \; \texttt{+} \; term \; \texttt{*} \; term \end{array}$$

These correspond to different 'parse trees'. Effectively, we are free to interpret $x \; \texttt{+} \; y \; \texttt{*} \; z$ either as $x \; \texttt{+} \; (y \; \texttt{*} \; z)$ or $(x \; \texttt{+} \; y) \; \texttt{*} \; z$.

Encoding precedences

We can encode operator precedences by introducing extra categories, e.g.

$$
\begin{array}{rcl}
atom & \longrightarrow & name\texttt{(}termlist\texttt{)} \\
 & | & name \\
 & | & numeral \\
 & | & \texttt{(}term\texttt{)} \\
 & | & \texttt{-}atom \\
mulexp & \longrightarrow & atom\ \texttt{*}\ mulexp \\
 & | & atom \\
term & \longrightarrow & mulexp\ \texttt{+}\ term \\
 & | & mulexp \\
termlist & \longrightarrow & term\texttt{,}termlist \\
 & | & term
\end{array}
$$

Now it's unambiguous. Multiplication has higher precedence and both infixes associate to the right.

Recursive descent

A *recursive descent* parser is a series of mutually recursive functions, one for each syntactic category (*term*, *mulexp* etc.).

The mutually recursive structure mirrors that in the grammar.

This makes them quite easy and natural to write — especially in ML, where recursion is the principal control mechanism.

For example, the procedure for parsing terms, say `term` will, on encountering a `-` symbol, make a recursive call to itself to parse the subterm, and on encountering a name followed by an opening parenthesis, will make a recursive call to `termlist`. This in itself will make at least one recursive call to `term`, and so on.

Parsers in ML

We assume that a parser accepts a list of input characters or tokens of arbitrary type.

It returns the result of parsing, which has some other arbitrary type, and also the list of input objects not yet processed. Therefore the type of a parser is:

$$(\alpha)list \rightarrow \beta \times (\alpha)list$$

For example, when given the input characters `(x + y) * z` the function `atom` will process the characters `(x + y)` and leave the remaining characters `* z`. It might return a parse tree for the processed expression using our earlier recursive type, and hence we would have:

```
atom "(x + y) * z" =
  Fn("+",[Var "x"; Var "y"]),"* z"
```

Parser combinators

In ML, we can define a series of *combinators* for plugging parsers together and creating new parsers from existing ones.

By giving some of them infix status, we can make the ML parser program look quite similar in structure to the original grammar.

First we declare an exception to be used where parsing fails:

```
exception Noparse;;
```

`p1 ++ p2` applies `p1` first and then applies `p2` to the remaining tokens; `many` keeps applying the same parser as long as possible.

`p >> f` works like `p` but then applies `f` to the result of the parse.

`p1 || p2` tries `p1` first, and if that fails, tries `p2`. These are automatically infix, in decreasing order of precedence.

Definitions of the combinators

```
let prefix ++ parser1 parser2 input =
  let result1,rest1 = parser1 input in
  let result2,rest2 = parser2 rest1 in
  (result1,result2),rest2;;

let rec many parser input =
  try let result,next = parser input in
      let results,rest = many parser next in
      (result::results),rest
  with Noparse -> [],input;;

let prefix >> parser treatment input =
  let result,rest = parser input in
  treatment(result),rest;;

let prefix || parser1 parser2 input =
  try parser1 input
  with Noparse -> parser2 input;;
```

Auxiliary functions

These are the general functions we will use:

```
let rec itlist f =
     fun [] b -> b
       | (h::t) b -> f h (itlist f t b);;

let uncurry f(x,y) = f x y;;
let K x y = x;;
let C f x y = f y x;;

let o f g x = f(g x);;
#infix "o";;

let explode s =
  let rec exap n l =
    if n < 0 then l else
    exap (n - 1) ((sub_string s n 1)::l) in
  exap (string_length s - 1) [];;
```

Atomic parsers

We need a few primitive parsers to get us started.

```
let some p =
  fun [] -> raise Noparse
    | (h::t) -> if p h then (h,t)
                else raise Noparse;;

let a tok =
  some (fun item -> item = tok);;

let finished input =
  if input = [] then 0,input
  else raise Noparse;;
```

The first two accept something satisfying `p`, and something equal to `tok`, respectively. The last one makes sure there is no unprocessed input.

Lexical analysis

First we want to do lexical analysis, i.e. split the input characters into tokens. This can also be done using our combinators, together with a few character discrimination functions. First we declare the type of tokens:

```
type token = Name of string
           | Num of string
           | Other of string;;
```

We want the lexer to accept a string and produce a list of tokens, ignoring spaces, e.g.

```
#lex "sin(x + y) * cos(2 * x + y)";;
 - : token list =
 [Name "sin"; Other "("; Name "x"; Other "+";
  Name "y"; Other ")"; Other "*"; Name "cos";
  Other "("; Num "2"; Other "*"; Name "x";
  Other "+"; Name "y"; Other ")"]
```

Definition of the lexer

```
let lex =
  let several p = many (some p) in
  let lowercase_letter s = "a" <= s & s <= "z" in
  let uppercase_letter s = "A" <= s & s <= "Z" in
  let letter s =
    lowercase_letter s or uppercase_letter s in
  let alpha s = letter s or s = "_" or s = "'" in
  let digit s = "0" <= s & s <= "9" in
  let alphanum s = alpha s or digit s in
  let space s = s = " " or s = "\n" or s = "\t" in
  let collect(h,t) = h^(itlist (prefix ^) t "") in
  let rawname =
     some alpha ++ several alphanum
     >> (Name o collect) in
  let rawnumeral =
     some digit ++ several digit
     >> (Num o collect) in
  let rawother = some (K true) >> Other in
  let token =
    (rawname || rawnumeral || rawother) ++
    several space >> fst in
  let tokens = (several space ++ many token) >> snd in
  let alltokens = (tokens ++ finished) >> fst in
  fst o alltokens o explode;;
```

Parsing terms

In order to parse terms, we start with some basic parsers for single tokens of a particular kind:

```
let name =
  fun (Name s::rest) -> s,rest
    | _ -> raise Noparse;;

let numeral =
  fun (Num s::rest) -> s,rest
    | _ -> raise Noparse;;

let other =
  fun (Other s::rest) -> s,rest
    | _ -> raise Noparse;;
```

Now we can define a parser for terms, in a form very similar to the original grammar. The main difference is that each production rule has associated with it some sort of special action to take as a result of parsing.

The term parser (take 1)

```
let rec atom input
      = (name ++
         a (Other "(") ++ termlist ++ a (Other ")")
             >> (fun (((f,_),a),_) -> Fn(f,a))
      || name
             >> (fun s -> Var s)
      || numeral
             >> (fun s -> Const s)
      || a (Other "(") ++ term ++ a (Other ")")
             >> (snd o fst)
      || a (Other "-") ++ atom
             >> snd) input
and mulexp input
      = (atom ++ a(Other "*") ++ mulexp
             >> (fun ((a,_),m) -> Fn("*",[a;m]))
      || atom) input
and term input
      = (mulexp ++ a(Other "+") ++ term
             >> (fun ((a,_),m) -> Fn("+",[a;m]))
      || mulexp) input
and termlist input
      = (term ++ a (Other ",") ++ termlist
             >> (fun ((h,_),t) -> h::t)
      || term
             >> (fun h -> [h])) input;;
```

Examples

Let us package everything up as a single parsing function:

```
let parser =
  fst o (term ++ finished >> fst) o lex;;
```

To see it in action, we try with and without the printer (see above) installed:

```
#parser "sin(x + y) * cos(2 * x + y)";;
- : term =
 Fn
  ("*",
   [Fn ("sin", [Fn ("+", [Var "x"; Var "y"])]);
    Fn ("cos", [Fn ("+", [Fn ("*", [Const "2";
                                    Var "x"]);
                          Var "y"])])])
#install_printer "print_term";;
- : unit = ()
#parser "sin(x + y) * cos(2 * x + y)";;
- : term = `sin(x + y) * cos(2 * x + y)`
```

Automating precedence parsing

We can easily let ML construct the ‘fixed-up’ grammar from our dynamic list of infixes:

```
let rec binop op parser input =
  let atom1,rest1 as result = parser input in
  if not rest1 = [] & hd rest1 = Other op then
    let atom2,rest2 = binop op parser (tl rest1) in
    Fn(op,[atom1; atom2]),rest2
  else result;;

let findmin l = itlist
  (fun (_,pr1 as p1) (_,pr2 as p2) ->
     if pr1 <= pr2 then p1 else p2) (tl l) (hd l);;

let rec delete x (h::t) =
  if h = x then t else h::(delete x t);;

let rec precedence ilist parser input =
  if ilist = [] then parser input else
  let opp = findmin ilist in
  let ilist' = delete opp ilist in
  binop (fst opp) (precedence ilist' parser) input;;
```

By using this function, we can make the main parser simpler and more general.

The term parser (take 2)

```
let rec atom input
     = (name ++
        a (Other "(") ++ termlist ++ a (Other ")")
            >> (fun (((f,_),a),_) -> Fn(f,a))
     || name
            >> (fun s -> Var s)
     || numeral
            >> (fun s -> Const s)
     || a (Other "(") ++ term ++ a (Other ")")
            >> (snd o fst)
     || a (Other "-") ++ atom
            >> snd) input
and term input = precedence (!infixes) atom input
and termlist input
     = (term ++ a (Other ",") ++ termlist
            >> (fun ((h,_),t) -> h::t)
     || term
            >> (fun h -> [h])) input;;
```

This will dynamically construct the precedence parser using the list of infixes active when it is actually used. Now the basic grammar is simpler.

Backtracking and reprocessing

Some productions for the same syntactic category have a common prefix. Note that our production rules for *term* have this property:

$$\begin{array}{rcl} term & \longrightarrow & name(termlist) \\ & | & name \\ & | & \cdots \end{array}$$

We carefully put the longer production first in our actual implementation, otherwise success in reading a name would cause the abandonment of attempts to read a parenthesized list of arguments.

However, this backtracking can lead to our processing the initial name twice.

This is not very serious here, but it could be in `termlist`.

An improved treatment

We can easily replace:

```
let ...
and termlist input
     = (term ++ a (Other ",") ++ termlist
              >> (fun ((h,_),t) -> h::t)
     || term
              >> (fun h -> [h])) input;;
```

with

```
let ...
and termlist input
    = term ++
      many (a (Other ",") ++ term >> snd)
              >> (fun (h,t) -> h::t) input;;
```

This gives another improvement to the parser, which is now more efficient and slightly simpler. The final version is:

The term parser (take 3)

```
let rec atom input
      = (name ++
          a (Other "(") ++ termlist ++ a (Other ")")
               >> (fun (((f,_),a),_) -> Fn(f,a))
      || name
               >> (fun s -> Var s)
      || numeral
               >> (fun s -> Const s)
      || a (Other "(") ++ term ++ a (Other ")")
               >> (snd o fst)
      || a (Other "-") ++ atom
               >> snd) input
and term input = precedence (!infixes) atom input
and termlist input
       = (term ++ many (a (Other ",") ++ term >> snd)
               >> (fun (h,t) -> h::t)) input;;
```

General remarks

With care, this parsing method can be used effectively. It is a good illustration of the power of higher order functions.

The code of such a parser is highly structured and similar to the grammar, therefore easy to modify.

However it is not as efficient as LR parsers; CAML-Yacc is capable of generating good LR parsers automatically.

Recursive descent also has trouble with *left recursion*. For example, if we had wanted to make the addition operator left-associative in our earlier grammar, we could have used:

$$\begin{array}{rcl} term & \longrightarrow & term \; \texttt{+} \; mulexp \\ & | & mulexp \end{array}$$

The naive transcription into ML would loop indefinitely. However we can often replace such constructs with explicit repetitions.