Handout 4 (Sulzmann & Lu Algorithm)

So far our algorithm based on derivatives was only able to say yes or no depending on whether a string was matched by regular expression or not. Often a more interesting question is to find out *how* a regular expression matched a string? Answering this question will also help us with the problem we are after, namely tokenising an input string. The algorithm we will be looking at for this was designed by Sulzmann & Lu in a rather recent paper. A link to it is provided on KEATS, in case you are interested.[1]

In order to give an answer for how a regular expression matched a string, Sulzmann and Lu introduce *values*. A value will be the output of the algorithm whenever the regular expression matches the string. If not, an error will be raised. Since the first phase of the algorithm by Sulzmann & Lu is identical to the derivative based matcher from the first coursework, the function *nullable* will be used to decide whether as string is matched by a regular expression. If *nullable* says yes, then values are constructed that reflect how the regular expression matched the string. The definitions for regular expressions r and values v is shown next to each other below:

regular expressions			values		
r	$::=$	$\varnothing$	v	$::=$	
	$\mid$	ϵ			$Empty$
	$\mid$	c		$\mid$	$Char(c)$
	$\mid$	$r_1 \cdot r_2$		$\mid$	$Seq(v_1, v_2)$
	$\mid$	$r_1 + r_2$		$\mid$	$Left(v)$
				$\mid$	$Right(v)$
	$\mid$	r^*		$\mid$	$[v_1, \ldots v_n]$

The point is that there is a very strong correspondence between them. There is no value for the $\varnothing$ regular expression, since it does not match any string. Otherwise there is exactly one value corresponding to each regular expression with the exception of $r_1 + r_2$ where there are two values, namely $Left(v)$ and $Right(v)$ corresponding to the two alternatives. Note that r^* is associated with a list of values, one for each copy of r that was needed to match the string. This means we might also return the empty list $[]$, if no copy was needed.

To emphasise the connection between regular expressions and values, I have in my implementation the convention that regular expressions are written entirely with upper-case letters, while values just start with a single upper-case character. My definition of values in Scala is below. I use this in the REPL of Scala; when I use the Scala compiler I need to rename some constructors, because Scala on Macs does not like classes that are called `EMPTY` and `Empty`.

```
abstract class Val
case object Empty extends Val
```

[1] In my humble opinion this is an interesting instance of the research literature: it contains a very neat idea, but its presentation is rather sloppy. In earlier versions of their paper, students and I found several rather annoying typos in their examples and definitions.

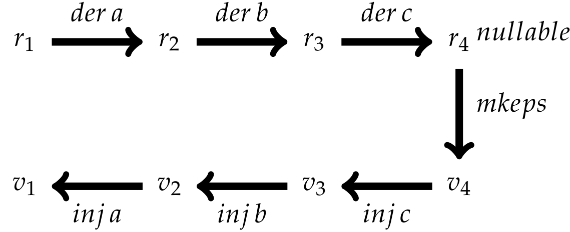

Figure 1: The two phases of the algorithm by Sulzmann & Lu.

```
case class Char(c: Char) extends Val
case class Seq(v1: Val, v2: Val) extends Val
case class Left(v: Val) extends Val
case class Right(v: Val) extends Val
case class Stars(vs: List[Val]) extends Val
```

Graphically the algorithm by Sulzmann & Lu can be illustrated by the picture in Figure 1 where the path from the left to the right involving $der/nullable$ is the first phase of the algorithm and $mkeps/inj$, the path from right to left, the second phase. This picture shows the steps required when a regular expression, say r_1, matches the string abc. We first build the three derivatives (according to a, b and c). We then use $nullable$ to find out whether the resulting regular expression can match the empty string. If yes we call the function $mkeps$.

The $mkeps$ function calculates a value for how a regular expression has matched the empty string. Its definition is as follows:

$$
\begin{array}{lcl}
mkeps(\epsilon) & \stackrel{\text{def}}{=} & Empty \\
mkeps(r_1 + r_2) & \stackrel{\text{def}}{=} & \text{if } nullable(r_1) \\
 & & \text{then } Left(mkeps(r_1)) \\
 & & \text{else } Right(mkeps(r_2)) \\
mkeps(r_1 \cdot r_2) & \stackrel{\text{def}}{=} & Seq(mkeps(r_1), mkeps(r_2)) \\
mkeps(r^*) & \stackrel{\text{def}}{=} & []
\end{array}
$$

There are no cases for ϵ and c, since these regular expression cannot match the empty string. Note also that in case of alternatives we give preference to the regular expression on the left-hand side. This will become important later on.

The second phase of the algorithm is organised so that it will calculate a value for how the derivative regular expression has matched a string whose first character has been chopped off. Now we need a function that reverses this “chopping off” for values. The corresponding function is called inj for injection. This function takes three arguments: the first one is a regular expression for which we want to calculate the value, the second is the character we want to inject and the third argument is the value where we will inject the character. The result of this function is a new value. The definition of inj is as follows:

$$
\begin{array}{ll}
inj\,(c)\,c\,Empty & \stackrel{\text{def}}{=} Char\,c\\
inj\,(r_1 + r_2)\,c\,Left(v) & \stackrel{\text{def}}{=} Left(inj\,r_1\,c\,v)\\
inj\,(r_1 + r_2)\,c\,Right(v) & \stackrel{\text{def}}{=} Right(inj\,r_2\,c\,v)\\
inj\,(r_1 \cdot r_2)\,c\,Seq(v_1, v_2) & \stackrel{\text{def}}{=} Seq(inj\,r_1\,c\,v_1, v_2)\\
inj\,(r_1 \cdot r_2)\,c\,Left(Seq(v_1, v_2)) & \stackrel{\text{def}}{=} Seq(inj\,r_1\,c\,v_1, v_2)\\
inj\,(r_1 \cdot r_2)\,c\,Right(v) & \stackrel{\text{def}}{=} Seq(mkeps(r_1), inj\,r_2\,c\,v)\\
inj\,(r^*)\,c\,Seq(v, vs) & \stackrel{\text{def}}{=} inj\,r\,c\,v\;::\;vs
\end{array}
$$

This definition is by recursion on the regular expression and by analysing the shape of the values. Therefore there are, for example, three cases for sequnece regular expressions. The last clause for the star regular expression returns a list where the first element is $inj\,r\,c\,v$ and the other elements are vs. That means $_ :: _$ should be read as list cons.

To understand what is going on, it might be best to do some example calculations and compare them with Figure 1. For this note that we have not yet dealt with the need of simplifying regular expressions (this will be a topic on its own later). Suppose the regular expression is $a \cdot (b \cdot c)$ and the input string is abc. The derivatives from the first phase are as follows:

$$
\begin{array}{ll}
r_1: & a \cdot (b \cdot c)\\
r_2: & \epsilon \cdot (b \cdot c)\\
r_3: & (\varnothing \cdot (b \cdot c)) + (\epsilon \cdot c)\\
r_4: & (\varnothing \cdot (b \cdot c)) + ((\varnothing \cdot c) + \epsilon)
\end{array}
$$

According to the simple algorithm, we would test whether r_4 is nullable, which in this case it is. This means we can use the function $mkeps$ to calculate a value for how r_4 was able to match the empty string. Remember that this function gives preference for alternatives on the left-hand side. However there is only ϵ on the very right-hand side of r_4 that matches the empty string. Therefore $mkeps$ returns the value

$$v_4 : \; Right(Right(Empty))$$

The point is that from this value we can directly read off which part of r_4 matched the empty string. Next we have to "inject" the last character, that is c in the running example, into this value v_4 in order to calculate how r_3 could have matched the string c. According to the definition of inj we obtain

$$v_3 : \; Right(Seq(Empty, Char(c)))$$

This is the correct result, because r_3 needs to use the right-hand alternative, and then ϵ needs to match the empty string and c needs to match c. Next we need to inject back the letter b into v_3. This gives

$$v_2 : \; Seq(Empty, Seq(Char(b), Char(c)))$$

which is again the correct result for how r_2 matched the string bc. Finally we need to inject back the letter a into v_2 giving the final result

$$v_1 : \; Seq(Char(a), Seq(Char(b), Char(c)))$$

This now corresponds to how the regular expression $a \cdot (b \cdot c)$ matched the string abc.

There are a few auxiliary functions that are of interest when analysing this algorithm. One is called *flatten*, written $|_|$, which extracts the string "underlying" a value. It is defined recursively as

$$\begin{array}{lcl}
|Empty| & \stackrel{\text{def}}{=} & [] \\
|Char(c)| & \stackrel{\text{def}}{=} & [c] \\
|Left(v)| & \stackrel{\text{def}}{=} & |v| \\
|Right(v)| & \stackrel{\text{def}}{=} & |v| \\
|Seq(v_1, v_2)| & \stackrel{\text{def}}{=} & |v_1| \,@\, |v_2| \\
|[v_1, \ldots, v_n]| & \stackrel{\text{def}}{=} & |v_1| \,@ \ldots @\, |v_n|
\end{array}$$

Using flatten we can see what is the string behind the values calculated by *mkeps* and *inj* in our running example are:

$$\begin{array}{ll}
|v_4|: & [] \\
|v_3|: & c \\
|v_2|: & bc \\
|v_1|: & abc
\end{array}$$

This indicates that *inj* indeed is injecting, or adding, back a character into the value. Next we look at how simplification can be included into this algorithm.

Simplification

Generally the matching algorithms based on derivatives do poorly unless the regular expressions are simplified after each derivative step. But this is a bit more involved in the algorithm of Sulzmann & Lu. So what follows might require you to read several times before it makes sense and also might require that you do some example calculations. As a first example consider the last derivation step in our earlier example:

$$r_4 = der\, c\, r_3 = (\varnothing \cdot (b \cdot c)) + ((\varnothing \cdot c) + \epsilon)$$

Simplifying this regular expression would just give us ϵ. Running *mkeps* with this regular expression as input, however, would then provide us with *Empty* instead of $Right(Right(Empty))$ that was obtained without the simplification. The problem is we need to recreate this more complicated value, rather than just *Empty*.

This will require what I call *rectification functions*. They need to be calculated whenever a regular expression gets simplified. Rectification functions take a value as argument and return a (rectified) value. Let us first take a look again at our simplification rules:

$$\begin{array}{lcl}
r \cdot \varnothing & \mapsto & \varnothing \\
\varnothing \cdot r & \mapsto & \varnothing \\
r \cdot \epsilon & \mapsto & r \\
\epsilon \cdot r & \mapsto & r \\
r + \varnothing & \mapsto & r \\
\varnothing + r & \mapsto & r \\
r + r & \mapsto & r
\end{array}$$

Applying them to r_4 will require several nested simplifications in order end up with just ϵ. However, it is possible to apply them in a depth-first, or inside-out, manner in order to calculate this simplified regular expression.

The rectification we can implement this by letting simp return not just a (simplified) regular expression, but also a rectification function. Let us consider the alternative case, $r_1 + r_2$, first. By going depth-first, we first simplify the component regular expressions r_1 and r_2. This will return simplified versions (if they can be simplified), say r_{1s} and r_{2s}, but also two rectification functions f_{1s} and f_{2s}. We need to assemble them in order to obtain a rectified value for $r_1 + r_2$. In case r_{1s} simplified to $\varnothing$, we continue the derivative calculation with r_{2s}. The Sulzmann & Lu algorithm would return a corresponding value, say v_{2s}. But now this value needs to be "rectified" to the value

$$Right(v_{2s})$$

The reason is that we look for the value that tells us how $r_1 + r_2$ could have matched the string, not just r_2 or r_{2s}. Unfortunately, this is still not the right value in general because there might be some simplifications that happened inside r_2 and for which the simplification function retuned also a rectification function f_{2s}. So in fact we need to apply this one too which gives

$$Right(f_{2s}(v_{2s}))$$

This is now the correct, or rectified, value. Since the simplification will be done in the first phase of the algorithm, but the rectification needs to be done to the values in the second phase, it is advantageous to calculate the rectification as a function, remember this function and then apply the value to this function during the second phase. So if we want to implement the rectification as function, we would need to return

$$\lambda v.\, Right(f_{2s}(v))$$

which is the lambda-calculus notation for a function that expects a value v and returns everything after the dot where v is replaced by whatever value is given.

Let us package this idea with rectification functions into a single function (still only considering the alternative case):

$simp(r)$:
  case $r = r_1 + r_2$
    let $(r_{1s}, f_{1s}) = simp(r_1)$
      $(r_{2s}, f_{2s}) = simp(r_2)$
    case $r_{1s} = \varnothing$: return $(r_{2s}, \lambda v.\, Right(f_{2s}(v)))$
    case $r_{2s} = \varnothing$: return $(r_{1s}, \lambda v.\, Left(f_{1s}(v)))$
    case $r_{1s} = r_{2s}$: return $(r_{1s}, \lambda v.\, Left(f_{1s}(v)))$
    otherwise: return $(r_{1s} + r_{2s}, f_{alt}(f_{1s}, f_{2s}))$

We first recursively call the simplification with r_1 and r_2. This gives simplified regular expressions, r_{1s} and r_{2s}, as well as two rectification functions f_{1s} and f_{2s}. We next need to test whether the simplified regular expressions are $\varnothing$ so as to make further simplifications. In case r_{1s} is $\varnothing$, then we can return r_{2s} (the other alternative). However for this we need to build a corresponding rectification function, which as said above is

$$\lambda v.\, Right(f_{2s}(v))$$

The case where $r_{2s} = \varnothing$ is similar: We return r_{1s} and rectify with $Left(_)$ and the other calculated rectification function f_{1s}. This gives

$$\lambda v.\, Left(f_{1s}(v))$$

The next case where $r_{1s} = r_{2s}$ can be treated like the one where $r_{2s} = \varnothing$. We return r_{1s} and rectify with $Left(_)$ and so on.

The otherwise-case is slightly more complicated. In this case neither r_{1s} nor r_{2s} are $\varnothing$ and also $r_{1s} \neq r_{2s}$, which means no further simplification can be applied. Accordingly, we return $r_{1s} + r_{2s}$ as the simplified regular expression. In principle we also do not have to do any rectification, because no simplification was done in this case. But this is actually not true: There might have been simplifications inside r_{1s} and r_{2s}. We therefore need to take into account the calculated rectification functions f_{1s} and f_{2s}. We can do this by defining a rectification function f_{alt} which takes two rectification functions as arguments and applies them according to whether the value is of the form $Left(_)$ or $Right(_)$:

$f_{alt}(f_1, f_2) \stackrel{\text{def}}{=}$
  $\lambda v.$ case $v = Left(v')$: return $Left(f_1(v'))$
    case $v = Right(v')$: return $Right(f_2(v'))$

The other interesting case with simplification is the sequence case. In this case the main simplification function is as follows

$$
\begin{array}{l}
simp(r): \qquad \text{(continued)}\\
\quad \text{case } r = r_1 \cdot r_2\\
\qquad \text{let } (r_{1s}, f_{1s}) = simp(r_1)\\
\qquad\quad (r_{2s}, f_{2s}) = simp(r_2)\\
\qquad \text{case } r_{1s} = \varnothing\text{: return } (\varnothing, f_{error})\\
\qquad \text{case } r_{2s} = \varnothing\text{: return } (\varnothing, f_{error})\\
\qquad \text{case } r_{1s} = \epsilon\text{: return } (r_{2s}, \lambda v.\, Seq(f_{1s}(Empty), f_{2s}(v)))\\
\qquad \text{case } r_{2s} = \epsilon\text{: return } (r_{1s}, \lambda v.\, Seq(f_{1s}(v), f_{2s}(Empty)))\\
\qquad \text{otherwise: return } (r_{1s} \cdot r_{2s}, f_{seq}(f_{1s}, f_{2s}))
\end{array}
$$

whereby in the last line f_{seq} is again pushing the two rectification functions into the two components of the Seq-value:

$$
\begin{array}{l}
f_{seq}(f_1, f_2) \stackrel{\text{def}}{=}\\
\qquad \lambda v.\ \text{case } v = Seq(v_1, v_2)\text{: return } Seq(f_1(v_1), f_2(v_2))
\end{array}
$$

Note that in the case of $r_{1s} = \varnothing$ (similarly r_{2s}) we use the function f_{error} for rectification. If you think carefully, then you will realise that this function will actually never been called. This is because a sequence with $\varnothing$ will never recognise any string and therefore the second phase of the algorithm would never been called. The simplification function still expects us to give a function. So in my own implementation I just returned a function which raises an error. In the case where $r_{1s} = \epsilon$ (similarly r_{2s}) we have to create a sequence where the first component is a rectified version of $Empty$. Therefore we call f_{1s} with $Empty$.

Since we only simplify regular expressions of the form $r_1 + r_2$ and $r_1 \cdot r_2$ we do not have to do anything else in the remaining cases. The rectification function will be just the identity, which in lambda-calculus terms is just

$$\lambda v.\, v$$

This completes the high-level version of the simplification function, which is also shown again in Figure 2. This can now be used in a *lexing function* as follows:

$$
\begin{array}{lcl}
lex\, r\, [] & \stackrel{\text{def}}{=} & \text{if } nullable(r) \text{ then } mkeps(r)\\
 & & \text{else } error\\
lex\, r\, c{::}s & \stackrel{\text{def}}{=} & \text{let } (r_{simp}, f_{rect}) = simp(der(c, r))\\
 & & inj\, r\, c\, f_{rect}(lex\, r_{simp}\, s)
\end{array}
$$

This corresponds to the *matches* function we have seen in earlier lectures. In the first clause we are given an empty string, [], and need to test wether the regular expression is *nullable*. If yes we can proceed normally and just return the value calculated by *mkeps*. The second clause is for strings where the first character is c and the rest of the string is s. We first build the derivative of r with respect to c; simplify the resulting regulare expression. We continue lexing with the simplified regular expression and the string s. Whatever will be returned as value, we sill rectify using the f_{rect} from the simplification and finally inject c back into the (rectified) value.

$simp(r)$:
  case $r = r_1 + r_2$
    let $(r_{1s}, f_{1s}) = simp(r_1)$
      $(r_{2s}, f_{2s}) = simp(r_2)$
    case $r_{1s} = \varnothing$: return $(r_{2s}, \lambda v. Right(f_{2s}(v)))$
    case $r_{2s} = \varnothing$: return $(r_{1s}, \lambda v. Left(f_{1s}(v)))$
    case $r_{1s} = r_{2s}$: return $(r_{1s}, \lambda v. Left(f_{1s}(v)))$
    otherwise: return $(r_{1s} + r_{2s}, f_{alt}(f_{1s}, f_{2s}))$

  case $r = r_1 \cdot r_2$
    let $(r_{1s}, f_{1s}) = simp(r_1)$
      $(r_{2s}, f_{2s}) = simp(r_2)$
    case $r_{1s} = \varnothing$: return $(\varnothing, f_{error})$
    case $r_{2s} = \varnothing$: return $(\varnothing, f_{error})$
    case $r_{1s} = \epsilon$: return $(r_{2s}, \lambda v. Seq(f_{1s}(Empty), f_{2s}(v)))$
    case $r_{2s} = \epsilon$: return $(r_{1s}, \lambda v. Seq(f_{1s}(v), f_{2s}(Empty)))$
    otherwise: return $(r_{1s} \cdot r_{2s}, f_{seq}(f_{1s}, f_{2s}))$

  otherwise:
    return $(r, \lambda v. v)$

Figure 2: The simplification function that returns a simplified regular expression and a rectification function.

Records and Tokenisation

Algorithm by Sulzmann, Lexing