15-399 Supplementary Notes: Regular Expression Matching as Deduction

Robert Harper

April 14, 2004

In class we associated with each regular expression $\mathbf{r}$ an unrestricted context $\Gamma_{\mathbf{r}}(s, f)$ with a designated "start" and "end" predicate symbol unique to that context. The context is chosen so that the following *adequacy theorem* holds:[1]

Theorem 0.1 *Let $\Gamma_{\mathbf{r}}(s, f)$ be the context, start, and end predicates associate with regular expression $\mathbf{r}$. Then $x \in \mathcal{L}(\mathbf{r})$ iff $\Gamma_{\mathbf{r}}(s, f); \bullet \Vdash \forall y.s(x \cdot y) \multimap f(y)$.*

The proof of adequacy proceeds in the forward direction by induction on the structure of $\mathbf{r}$, in each case exhibiting the required derivation. In the backward direction we relied on a normalization theorem for DILL that allows us to proceed by analyzing the structure of a normal proof of the quantified formula. The sketch of the proof given in class was unnecessarily turgid; the purpose of this note is to give a clearer proof.

First, let us review the definition of $\Gamma_{\mathbf{r}}(s, f)$ given in Figure 1. The only significant difference to what we did in class is in the treatment of the regular expression $\mathbf{0}$, which matches nothing. Here we have two axioms, rather than one. This is done to ensure the following invariants for each $\Gamma_{\mathbf{r}}(s, f)$:

1. The start, s, and end, f, symbol of $\Gamma_{\mathbf{r}}(s, f)$ is unique to that context.
2. There is precisely one assumption governing the start symbol, s, and it has the form $\forall \ldots .(s(\ldots) \multimap \ldots)$.
3. There is precisely one assumption governing the end symbol, f, and it has the form $\forall \ldots .(\ldots \multimap f(\ldots))$.

Second, the backward direction of adequacy follows from the following lemma:

Lemma 0.1 *If $\Gamma_{\mathbf{r}}(s, f), y; s(z){\downarrow} \Vdash f(y){\uparrow}$, then $z = x \cdot y$ for some $x \in \mathcal{L}(\mathbf{r})$.*

Proof: We proceed by induction on the structure of $\mathbf{r}$, analyzing the form of normal proofs of the antecedent in each case.

- Suppose that $\mathbf{r} = \mathbf{r}_1\,\mathbf{r}_2$. We have by induction

[1] Throughout variables range over the type of strings of letters of a fixed alphabet.

$$
\begin{array}{rcl}
\Gamma_{\mathbf{1}}(s,f) & = & \forall y.s(y) \multimap f(y) \\
\Gamma_{\mathbf{a}}(s,f) & = & \forall y.s(a \cdot y) \multimap f(y) \\
\Gamma_{\mathbf{0}}(s,f) & = & \forall y.s(y) \multimap \top \\
 & & \forall y.\mathbf{0} \multimap f(y) \\
\Gamma_{\mathbf{r}_1\,\mathbf{r}_2}(s,f) & = & \forall y.s(y) \multimap s_1(y) \\
 & & \Gamma_{\mathbf{r}_1}(s_1,f_1) \\
 & & \forall y.f_1(y) \multimap s_2(y) \\
 & & \Gamma_{\mathbf{r}_2}(s_2,f_2) \\
 & & \forall y.f_2(y) \multimap f(y) \\
\Gamma_{\mathbf{r}_1+\mathbf{r}_2}(s,f) & = & \forall y.s(y) \multimap s_1(y) \& s_2(y) \\
 & & \Gamma_{\mathbf{r}_1}(s_1,f_1) \\
 & & \Gamma_{\mathbf{r}_2}(s_2,f_2) \\
 & & \forall y.f_1(y) \oplus f_2(y) \multimap f(y) \\
\Gamma_{\mathbf{r}_1\cap\mathbf{r}_2}(s,f) & = & \forall y.s(y) \multimap s_1(y) \otimes s_2(y) \\
 & & \Gamma_{\mathbf{r}_1}(s_1,f_1) \\
 & & \Gamma_{\mathbf{r}_2}(s_2,f_2) \\
 & & \forall y.f_1(y) \otimes f_2(y) \multimap f(y) \\
\Gamma_{\mathbf{r}^*}(s,f) & = & \forall y.s(y) \multimap f(y) \& s_1(y) \\
 & & \Gamma_{\mathbf{r}_1}(s_1,f_1) \\
 & & \forall y.f_1(y) \multimap s(y) \\
\Gamma_{\top}(s,f) & = & \forall x.\forall y.s(x \cdot y) \multimap f(y)
\end{array}
$$

Figure 1: Translation of Regular Expressions to Contexts

1. If $\Gamma_{\mathbf{r}_1}(s_1, f_1), y_1; s_1(z_1){\downarrow} \Vdash f_1(y_1){\uparrow}$, then $z_1 = x_1 \cdot y_1$ for some $x_1 \in \mathcal{L}(\mathbf{r}_1)$.
2. If $\Gamma_{\mathbf{r}_2}(s_2, f_2), y_2; s_2(z_2){\downarrow} \Vdash f_2(y_2){\uparrow}$, then $z_2 = x_2 \cdot y_2$ for some $x_2 \in \mathcal{L}(\mathbf{r}_2)$.

Assume $\Gamma_{\mathbf{r}}(s, f), y; s(z) \Vdash f(y){\uparrow}$; we are to show that $z = x \cdot y$ with $x \in \mathcal{L}(\mathbf{r})$. Consulting the definition of $\Gamma_{\mathbf{r}}$, the derivation must start with

$$\Gamma_{\mathbf{r}}(s, f), y; s(z){\downarrow} \Vdash s_1(z){\downarrow}$$

and end with

$$\Gamma_{\mathbf{r}}(s, f), y; f_2(y){\downarrow} \Vdash f(y){\downarrow}.$$

In between we must have

$$\Gamma_{\mathbf{r}}(s, f), y; f_1(w){\downarrow} \Vdash s_2(w){\downarrow},$$

for some w, since that is the only assumption linking $\mathbf{r}_1$ to $\mathbf{r}_2$. It follows that we must have

$$\Gamma_{\mathbf{r}}(s, f), y; s_2(w){\downarrow} \Vdash f_2(y){\downarrow},$$

from which we obtain by induction that $w = x_2 \cdot y$ with $x_2 \in \mathcal{L}(\mathbf{r}_2)$. We must also have

$$\Gamma_{\mathbf{r}}(s, f), y; s_1(z){\downarrow} \Vdash f_1(w){\downarrow},$$

from which it follows by induction that $z = x_1 \cdot w$ for some $x_1 \mathcal{L}(\mathbf{r}_1)$. This means that $z = x_1 \cdot x_2 \cdot y = x \cdot y$, where $x = x_1 \cdot x_2 \in \mathcal{L}(\mathbf{r})$, as desired.

- Suppose that $\mathbf{r} = \mathbf{r}_1 + \mathbf{r}_2$. We have by induction

1. If $\Gamma_{\mathbf{r}_1}(s_1, f_1), y_1; s_1(z_1){\downarrow} \Vdash f_1(y_1){\uparrow}$, then $z_1 = x_1 \cdot y_1$ for some $x_1 \in \mathcal{L}(\mathbf{r}_1)$.
2. If $\Gamma_{\mathbf{r}_2}(s_2, f_2), y_2; s_2(z_2){\downarrow} \Vdash f_2(y_2){\uparrow}$, then $z_2 = x_2 \cdot y_2$ for some $x_2 \in \mathcal{L}(\mathbf{r}_2)$.

Assume $\Gamma_{\mathbf{r}}(s, f), y; s(z) \Vdash f(y){\uparrow}$; we are to show that $z = x \cdot y$ where either $x \in \mathcal{L}(\mathbf{r}_1)$ or $x \in \mathcal{L}(\mathbf{r}_2)$. Consulting the definition of $\Gamma_{\mathbf{r}}$, the derivation must start with

$$\Gamma_{\mathbf{r}}(s, f), y; s(z){\downarrow} \Vdash s_1(z) \& s_2(z){\downarrow}$$

and end with

$$\Gamma_{\mathbf{r}}(s, f), y; f_1(y) \oplus f_2(y){\downarrow} \Vdash f(y){\downarrow}.$$

The latter implies that we have

$$\Gamma_{\mathbf{r}}(s, f), y; f_1(y){\downarrow} \Vdash f(y){\downarrow},$$

and

$$\Gamma_{\mathbf{r}}(s, f), y; f_2(y){\downarrow} \Vdash f(y){\downarrow}.$$

To fill the gap we either must have

$$\Gamma_{\mathbf{r}}(s,f), y; s_1(z) \& s_2(z){\downarrow} \Vdash s_1(z){\downarrow}$$

and

$$\Gamma_{\mathbf{r}}(s,f), y; s_1(z){\downarrow} \Vdash f_1(y),$$

or we must have

$$\Gamma_{\mathbf{r}}(s,f), y; s_1(z) \& s_2(z){\downarrow} \Vdash s_2(z){\downarrow}$$

and

$$\Gamma_{\mathbf{r}}(s,f), y; s_2(z){\downarrow} \Vdash f_2(y).$$

In the former case we have by induction that $z = x \cdot y$ for some $x \in \mathcal{L}(\mathbf{r}_1)$, and in the latter we have $z = x\dot{y}$ for some $x \in \mathcal{L}(\mathbf{r}_2)$, as desired.

The other cases are handled similarly. □