LECTURES 4: FORMAL LANGUAGES AND COMPLEXITY
October 8, 1998
Bud Mishra

1 Pumping Lemma; Non-regular Languages

$$\mathcal{L} = \{0^n 1^n | n \geq 0\}$$

Not possible to construct a DFA, M such that $\mathcal{L}(M) = \mathcal{L}$. Any such DFA must remember how many 0's have been seen so far in order to recognize valid strings. But since the machine has only $|Q| =$ finitely many states if we give it the following sets of strings:

$$\epsilon, 0, 00, \ldots, 0^a, \ldots, 0^b, \ldots . 0^{|Q|},$$

there must be a pair of distinct strings 0^a and 0^b $(a < b)$ such that they both bring the machine to the same state q':

$$\delta(q_0, 0^a) = \delta(q_0, 0^b) = q'$$

Now there are two possibilities, either $\delta(q', 1^a) \in F$ or $\delta(q', 1^a) \notin F$. If the first possibility holds then $0^b 1^a \in \mathcal{L}(M)$ and in the second case $0^a 1^a \notin \mathcal{L}(M)$. Thus, no matter what, $\mathcal{L}(M) \neq \mathcal{L}$.

Theorem. 1.1 *If $\mathcal{L}$ is a regular language, then there is a number p (the pumping length that only depends on $\mathcal{L}$) where*
$\forall \sigma \in \mathcal{L}$ if $|\sigma| \geq p$ then σ can be divided into substrings $\sigma = uvw$ such that

1. $|v| > 0$.

2. $|uv| \leq p$.

3. $\forall i \geq 0$ $uv^i w \in \mathcal{L}$.

PROOF.
See Sipser.

Suppose you (Jane) now wish to show (to Dick) that the following language

$$\mathcal{L} = \{ww | w \in \{0,1\}^*\} \neq \text{Reg}$$

We can imagine a conversation between Dick and Jane that proceeds as below:

> *Dick*: I claim that $\mathcal{L}$ is regular.
>
> *Jane*: If $\mathcal{L}$ is regular then what is its pumping length?
>
> *Dick*: Of course it is p. (Perhaps some specific number)
>
> *Jane*: Since $0^p10^p1 \in \mathcal{L}$, how would you break it up into $\sigma = 0^p10^p1 = uvw$, with $|uv| \leq p$, $|v| > 0$ so that it can be pumped?
>
> *Dick*: I have no choice but to take $u = 0^\alpha$ and $v = 0^\beta$ with $\beta > 0$ and $\alpha + \beta \leq p$. Then
>
> $$w = 0^{p-\alpha-\beta}10^p1.$$
>
> *Jane*: By pumping lemma, you also accept uw (uv^uw with $i = 0$); but
>
> $$uw = 0^{p-\beta}10^p1 \notin \mathcal{L}.$$
>
> Thus I prove that your claim that there is a DFA for M (with p or fewer states) is false. Also notice that, I have not made any specific use of the numerical value of p and my argument does not depend on what value of p you could have given. In general, there is no DFA to accept $\mathcal{L}$.

2 Myhill-Nerode Theorem

Let $\mathcal{L}$ = an arbitrary language. $\mathcal{L} \subseteq \Sigma^*$.

Define a relation $R_L \subseteq \Sigma^* \times \Sigma^*$ as follows:

$$xR_Ly \quad \text{iff} \quad \forall z\ xz \in \mathcal{L} \Leftrightarrow yz \in \mathcal{L}.$$

Note that R_L is reflexive, symmetric and transitive. (e.g., $xz \in \mathcal{L} \Leftrightarrow yz \in \mathcal{L} \Leftrightarrow wz \in \mathcal{L}$ showing that xR_Ly and yR_Lw implies that xR_Lw.)

Thus R_L is an equivalence relation, and partitions Σ^* into a set of equivalence classes. The *index* of R_L is the cardinality of the set of equivalence classes. We are most interested in equivalence relations with *finite index*.

Now consider a DFA

$$M = (Q, \Sigma, \delta, q_0, F).$$

Define a relation $R_M \subseteq \Sigma^* \times \Sigma^*$ as follows:

$$x R_M y \quad \text{iff} \quad \delta(q_0, x) = \delta(q_0, y).$$

Note that R_M is reflexive, symmetric and transitive, and thus, R_M is an equivalence relation, partitioning Σ^* into a *finite* set of equivalence classes. Clearly R_M has finite index.

Definition. 2.1 A *right invariant relation* (with respect to concatenation) is an equivalence relation $R \subseteq \Sigma^* \times \Sigma^*$ such that

$$xRy \Rightarrow xzRyz.$$

Claim: R_M is a right invariant equivalence relation of finite index.

$$\begin{aligned} xR_My &\Rightarrow \delta(q_0, x) = \delta(q_0, y) \\ &\Rightarrow \delta(\delta(q_0, x), z) = \delta(\delta(q_0, y), z) \\ &\Rightarrow \delta(q_0, xz) = \delta(q_0, yz) \\ &\Rightarrow xzR_Myz. \end{aligned}$$

Theorem. 2.2 *The following three statements are equivalent:*

1. *The set $\mathcal{L} \subseteq \Sigma^*$ is regular.*

2. *$\mathcal{L}$ is the union of some of the equivalence classes of a right invariant equivalence relation of finite index.*

3. *Let equivalence relation R_L be defined by:*

   $$xR_Ly \quad \textit{iff} \quad \forall z \in \Sigma^* \; xz \in \mathcal{L} \Leftrightarrow yz \in \mathcal{L}.$$

   Then R_L is of finite index.

PROOF.

(1) $\Rightarrow$ (2)

Let $\mathcal{L} \subseteq \Sigma^*$ be regular, accepted by a DFA M.

$$M(\mathcal{L}) = (Q, \Sigma, \delta, q_0, F).$$

Let $R_M \subseteq \Sigma^* \times \Sigma^*$, defined as

$$x R_M y \quad \text{iff} \quad \delta(q_0, x) = \delta(q_0, y).$$

We know that R_M is a right invariant equivalence relation of finite index.

Let

$$[q]_{R_M} = \{\sigma | \delta(q_0, \sigma) = q\},$$

an equivalence class contacting the strings that take M to a state q from the start state q_0.

Clearly,

$$\mathcal{L} = \bigcup_{q \in F} [q]_{R_M},$$

finite union of some of the equivalence classes.

(2) $\Rightarrow$ (3)

R_2 be any equivalence relation satisfying (2). We show that R_2 is a *refinement* of R_L.

$$[x]_{R_2} \subseteq [x]_{R_L};$$

Every equivalence class of R_2 is entirely contained in an equivalence class of R_L. Hence

$$\text{Index of } R_L \leq \text{Index of } R_2 = \text{Finite}.$$

This is true as

$$\begin{aligned} xR_2y &\Rightarrow \forall z \in \Sigma^* \; xzR_2yz \quad \text{(By right invariance)} \\ &\Rightarrow \forall z \in \Sigma^* \; xz \in \mathcal{L} \Leftrightarrow yz \in \mathcal{L} \quad (\mathcal{L} = \text{union of equivalence classes}) \\ &\Rightarrow xR_Ly. \end{aligned}$$

Hence

$$[x]_{R_2} = \{y | xR_2y\} \subseteq \{y | xR_Ly\} = [x]_{R_L}.$$

(3) $\Rightarrow$ (1) First note that R_L is right invariant:

$$
\begin{aligned}
xR_Ly &\Rightarrow \forall v = zw \; xv \in \mathcal{L} \Leftrightarrow yv \in \mathcal{L} \\
&\Rightarrow \forall w \in \Sigma^* \; xzw \in \mathcal{L} \Leftrightarrow yzw \in \mathcal{L} \\
&\Rightarrow \forall z \in \Sigma^* \; xzR_Lyz
\end{aligned}
$$

Thus R_L is a right invariant equivalence relation of finite index. The finite set of equivalence classes of R_L can be enumerated as

$$Q = \{[x_1], [x_2], \ldots, [x_n]\}, \quad [x_i] = \{y | x_i R_L y\}.$$

We construct a DFA M as follows:

$$M = (Q, \Sigma, \delta, q_0, F),$$

where

$$
\begin{aligned}
q_0 &= [\epsilon] \\
F &= \{[x_i] | x_i \in \mathcal{L}\} \\
\delta &: Q \times \Sigma \to Q \\
&: ([x_i], a) \mapsto [x_i a]
\end{aligned}
$$

Note that if $\sigma = a_1 a_2 \cdots a_n \in \mathcal{L}$ then

$$
\begin{aligned}
\delta([\epsilon], a_1) &= [a_1] \\
\delta([a_1], a_2) &= [a_1 a_2] \\
\delta([a_1 a_2], a_3) &= [a_1 a_2 a_3] \\
\delta([a_1 \cdots a_{n-1}], a_n) &= [a_1 a_2 \cdots a_n] \in F.
\end{aligned}
$$

3 Regular Expressions

See Sipser.