Summary on Lecture 1, March 28th, 2016

Languages and Finite State Machines

Warm-up: Languages. Let $\Sigma = \{a_1, \ldots, a_k\}$ be an alphabet. We denote by Σ^n the set of words (strings) over Σ of length n. There is a special *empty string* which is denoted by λ. We accept the convention: $\Sigma^0 = \{\lambda\}$. We will use the notations:

$$\Sigma^+ = \bigcup_{n>0} \Sigma^n, \quad \Sigma^* = \bigcup_{n\geq 0} \Sigma^n.$$

We say that a subset $A \subset \Sigma^*$ is a *language*. Two words (strings) $w = x_1 \ldots x_n$, $w' = x'_1 \ldots x'_{n'}$ in the laguage A are equal iff $n = n'$ and $x_i = x'_i$ for each $i = 1, \ldots, n$.

If $w = x_1 \ldots x_n$ is a word, then the length $\|w\| = n$. We also define $\|\lambda\| = 0$. There is an obvious concatenation of words: if $w = x_1 \ldots x_n$, $w' = x'_1 \ldots x'_{n'}$ then

$$ww' = x_1 \ldots x_n x'_1 \ldots x'_{n'}, \quad \|ww'\| = \|w\| + \|w'\|.$$

We accept the convention: $w\lambda = \lambda w = w$. In particular, $\lambda\lambda = \lambda$. For each word $w \in A$, we have that its power w^ℓ is well-defined.

Let $v = ww'$ be a concatenation of two words. Then we say that w is *proper prefix of* v if $w \neq \lambda$, and w' is *proper suffix of* v if $w' \neq \lambda$. In the case if the words w or w' could be empty, we call them *prefix of* v or *suffix of* v respectively.

We have seen many examples of alphabets and languages. Here some of them:

- Binary alphabet $\Sigma = \{0, 1\}$, $A^* = \Sigma^*$.
- English Language $\Sigma = \{a, b, \ldots, z, A, B, \ldots, Z\}$, $A = \Sigma^*$.
- It is known that DNA is constructed from four main types of molecules: **adenine** (A), **cytosine** (C), **guanine** (G), and **thymine** (T). Sequences of these molecules, and so strings over the alphabet $\Sigma = \{A, C, G, T\}$ form the basis of genes.

Let $A, B \subset \Sigma^*$ be two languages. We can form new language AB, the conteniation of A and B, as follows:

$$AB = \{ \ ab \ | a \in A, \quad b \in B \ \}.$$

Example. Let $\Sigma = \{x, y, z\}$, and $A = \{x, xy, z\}$, $B = \{\lambda, y\}$. Then

$$\begin{aligned} AB &= \{x, xy, z, xyy, zy\}, \\ BA &= \{x, xy, z, yx, yxy, yz\}. \end{aligned}$$

We have that $|AB| \neq |BA|$. In general, one can show that $|AB| \leq |A| \cdot |B|$.

Theorem 1. *Let $A, B, C \subset \Sigma^*$ be languages. Then*

(a) $A\{\lambda\} = \{\lambda\}A = A$ (b) $(AB)C = A(BC)$ (c) $A(B \cup C) = AB \cup AC$

(d) $(B \cup C)A = BA \cup CA$ (e) $A(B \cap C) = AB \cap AC$ (f) $(B \cap C)A = BA \cap CA$

Exercise. Prove Theorem 1.

For a language $A \subset \Sigma^*$, we also define its powers A^ℓ:

$$A^0 = \{\lambda\}, \quad A^1 = A, \quad A^{\ell+1} = \{ \ ab \mid a \in A, \ b \in A^\ell \ \}.$$

We also define its closures A^+ and A^* as follows:

$$A^+ = \bigcup_{\ell>0} A^\ell, \quad A^* = \bigcup_{\ell\geq 0} A^\ell.$$

The languages A^+ and A^* are called *positive closure* and *Kleene closure* of A respectively.

Examples. We consider two languages over $\Sigma = \{0, 1\}$:

(1) The language $\{1\}\{0, 1\}^*$ represents binary natural numbers.

(2) Binary strings containing the substring 1011 can be represented by the language

$$\{0, 1\}^* 1011 \{0, 1\}^*.$$

More examples. Let $\Sigma = \{x, y\}$.

(1) Let $A = \{xx, xy, yx, yy\}$. Then A^* is the language over Σ, in which all words have even length.

(2) Let $A = \{xx, xy, yx, yy\}$ be as above, and $B = \{x, y\}$. Then BA^* is the language Σ, in which all words have odd length.

(3) The language $\{x\}\{x, y\}^*$ over Σ contains all words from Σ^* for which x is a prefix, and the language $\{x\}\{x, y\}^+$ over Σ contains all words from Σ^* for which x is a proper prefix,

(4) The language $\{x, y\}^*\{yy\}$ over Σ contains all words from Σ^* for which yy is a suffix, and the language $\{x, y\}^+\{yy\}$ over Σ contains all words from Σ^* for which yy is a proper suffix.

(5) The language $\{x, y\}^*\{xxyy\}\{x, y\}^*$ over Σ consists of all words from Σ^* which contain a substring $xxyy$.

(6) The language $\{x\}^*\{y\}^*$ over Σ consists of all words from Σ^* which have some number (possibly zero) of x following by some number (possibly zero) of y. Notice that $\{x\}^*\{y\}^* \subset \{x, y\}^*$, but $\{x\}^*\{y\}^* \neq \{x, y\}^*$. Indeed, $w = xyx$ is in $\{x, y\}^*$, but not in $\{x\}^*\{y\}^*$.

Lemma 1. *Let $A, B \subset \Sigma^*$ be two languages. If $A \subset B$, then $A^\ell \subset B^\ell$ for each $\ell \geq 0$.*

Exercise. Prove Lemma 1.

Theorem 2. *Let $A, B \subset \Sigma^*$ be languages. Then*

(a) $A \subset AB^*$

(b) $A \subset B^*A$

(c) $A \subset B \Rightarrow A^+ \subset B^+$

(d) $A \subset B \Rightarrow A^* \subset B^*$

(e) $AA^* = A^*A = A^+$

(f) $A^*A^* = A^* = (A^*)^* = (A^*)^+ = (A^+)^*$

(g) $(A \cup B)^* = (A^* \cup B^*)^* = (A^*B^*)^*$.

Exercise. Prove Theorem 2.