Formal Languages and Parsing

CS 462

Jeffrey Shallit

LaTeXed by *Sibelius Peng*

Preface

Disclaimer Much of the information on this set of notes is transcribed directly/indirectly from the lectures of CS 462 during Winter 2022 as well as other related resources. I do not make any warranties about the completeness, reliability and accuracy of this set of notes. Use at your own risk.

For any questions, send me an email via `https://notes.sibeliusp.com/contact`.

You can find my notes for other courses on `https://notes.sibeliusp.com/`.

Sibelius Peng

Contents

1

CS 462 notation

- Natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$ and we use letters $i, j, k, \ell, m, n \in \mathbb{N}$.
- Finite string/word: a map from $[0, n-1]$ (an interval) to Σ (a finite alphabet of symbols)

  $w[i]$ is ith symbol of w
- infinite strings/words: a map from $\mathbb{N}$ to Σ. We denote infinite strings by bold-face:

$$\mathbf{w} = \mathbf{w}[0]\mathbf{w}[1]\mathbf{w}[2] \cdots$$

- Σ^* is the set of all finite words over Σ.
- Σ^ω is the set of all infinite words over Σ. Also written $\Sigma^{\mathbb{N}}$.
- $\Sigma^\infty = \Sigma^* \cup \Sigma^{\mathbb{N}}$.

Finite words typically denote by s, t, u, v, w, x, y, z

1.1 Some refreshers from CS 360/365

- x is a **prefix** of z if there exists y such that $z = xy$
- x is a **suffix** of z if there exists y such that $z = yx$
- x is a **subword** (factor) of z if there exists w, y such that $z = wxy$.
- x is a **subsequence** of z if x can be obtained from z by striking out zero or more symbols.

> Remark:
> Does substring mean contiguous (like subword)? or noncontiguous (like subsequence)? This definition depends the author of the book.
>
> Empty string ϵ is a first-class string like any other string and is not ruled out unless done so explicitly.

Then we have "proper" prefix, suffix, etc. If $z = xy$ and $x \neq z$, then x is a **proper prefix** of z.

1.2 Some notations

A shorthand for subword:

$$w[a..b] = w[a]w[a+1]\cdots w[b]$$

Concatenation of strings:

which is not commutative in general. Because we write concatenation in a multiplicative way, we can raise strings to powers: $x^n = \underbrace{xx\cdots x}_{n \text{ times}}$, or formally

$$\begin{aligned} x^0 &= \epsilon \\ x^n &= x \cdot x^{n-1} \qquad n \geq 1 \\ x^{m+n} &= x^m x^n \end{aligned}$$

A word is not of the form z^n, $n \geq 2$, $z \neq \epsilon$ is called **primitive**. The set of binary primitive words are denoted

$$P_2 = \{0, 1, 01, 10, 001, 010, 011, \ldots\}$$

One open question: is P_2 context-free? Probably not! But no one knows a proof.

1.3 Other operations on words

We define perfect shuffle on x and y, for $|x| = |y| = n$ as

$$x \text{ Ш } y = x[1]y[1]x[2]y[2]\cdots x[n]y[n]$$

where Ш is the Russian "sha". For example,

$$\text{term Ш hoes} = \text{theorems}$$

Single symbols are denoted by $a, b, c \in \Sigma$.

The reversal: x^R, symbols of x in reverse order. If you feel stressed, we can reverse it and get

$$(\text{stressed})^R = (\text{desserts})$$

Palindromes: $x = x^R$.

Ordering

lexicographic order We define it for the words of same length, $|x| = |y|$. Then $x < y$ means[1] there exists i such that $1 \leq i \leq n = |x| = |y|$, and $x[j] = y[j]$ for $j < i$ and $x[i] < y[i]$. $x \leq y$ means $x = y$ or $x < y$.

radio order $x < y$ in radix order, if $|x| < |y|$ or $|x| = |y|$ and $x < y$ in lexicographic order. For example,

$$\{0, 1, 2\}^* = \{\epsilon, 0, 1, 2, 00, 01, 02, 10, \ldots\}$$

cyclic shift of a string One example is `eat, ate, tea`

If x, y are cyclic shifts of each other, we say they are conjugates. Formally, x, y are conjugates if there exists u, v such that $x = uv$ and $y = vu$.

[1] need underlying order on Σ. For example, $a < b < c < \cdots$, $0 < 1 < 2 < \cdots$

Borders A word w is **bordered** if it has a proper nonempty prefix that is also a suffix. Otherwise, it's **unbordered**. One example is `entanglement`, whose border is `ent`. Also, we can have overlapping border: `alfalfa`.

1.4 Properties of infinite words

periodicity of infinite words Let $x \in \Sigma^+$, finite nonempty words over Σ. Then we can define

$$x^\omega = xxx\cdots$$

If $z = x^\omega$ for some x, we say z is **purely periodic**. If $z = yx^\omega$ for some finite y, then z is **ultimately periodic**.

2

Combinatorics on words

2.1 The theorems of Lyndon-Schützenberger

Suppose we have an equation from number theory,

$$x^2 + xy = y^2 - 1$$

and let's find solution in natural numbers:

$$\begin{aligned} x &= 0 \quad y = 1 \\ x &= 1 \quad y = 2 \\ x &= 3 \quad y = 5 \end{aligned}$$

Then we can guess the solutions are $x = F_{2n}, y = F_{2n+1}$ for $n \geq 0$.

Now we can consider equations in words: $x, y, z \in \Sigma^+$ (nonempty)

1. $xy = yx$ characterizes commuting words
2. $xy = yz$ characterizes bordered words

For the second equation, one solution would be $x = \texttt{alf}, y = \texttt{alfa}, z = \texttt{lfa}$.

Theorem 2.1

Suppose $x, y, z \in \Sigma^+$, $xy = yz$ if and only if $\exists u \in \Sigma^+$m $v \in \Sigma^*$, $e \geq 0$ such that

$$\begin{aligned} x &= uv \\ z &= vu \\ y &= (uv)^e u = u(vu)^e \end{aligned}$$

This theorem gives complete characterization to the equation.

Proof:
$\Leftarrow$ is easy to see:

$$xy = uv(uv)^e v = (uv)^e uvu = yz$$

For $\Rightarrow$, we prove by induction on $|y|$.

Base case $|y| = 1$. Let $y = a$, a single symbol. Then we have

$$xa = az$$

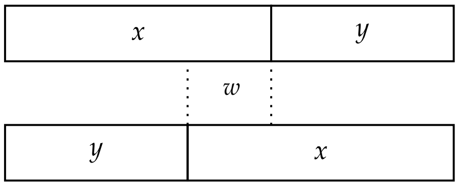
and then we find that $\exists x', z'$ such that $x = ax'$ and $z = z'a$. Then

$$ax'a = az'a$$

So $x' = z'$. Then we can take $u = a, v = x' = z', e = 0$. Then we are done with the base case.

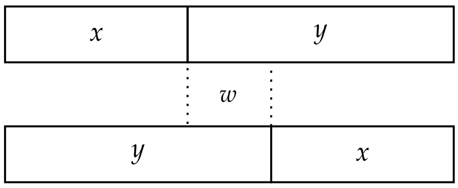
Now induction step. We discuss by cases (imposing length conditions) to break the symmetry.

Case I $|x| \geq |y|$.



We define w (could be empty) as in the picture. Then let $u = y, v = w, e = 0$.

Case II $|x| < |y|$.



We define w as in the picture. We observe that $w \neq \epsilon$, otherwise $|x| = |y|$. Also $x \neq \epsilon, z \neq \epsilon$. Then we observe that

$$y = wz = xw$$

which is our original equation with w playing the role of y. In order to apply induction, we need $|w| < |y|$, which is achieved by $x \neq \epsilon$. So induction says $\exists u, v, e,\ x = uv, z = vu, w = (uv)^e u$. Sub it back in, we get

$$wz = y = (uv)^e uvu = (uv)^{e+1} u$$

□

Consider the equation $x^2 = y^3$ in $\mathbb{N}$. We can parametrize the solution by $x = z^2, y = z^2$. This suggests the equation $x^2 = y^3$ over Σ^* only has the solution

$$x = z^3 = zzz$$
$$y = z^2 = zz$$

When does $xy = yx$? In other words, when does a word commute? Recall a classic theorem in linear algebra: two diagonalizable matrices commute if and only if they are simultaneously diagonalizable.

Theorem 2.2

Let $x, y \in \Sigma^+$. (Nonempty) Then the following 8 conditions are equivalent.

(1) There exist $z \in \Sigma^+$, and integers $k, \ell > 0$ such that $x = z^k$, $y = z^\ell$.

(2) $x^\omega = y^\omega$.

(3) There exist integers $i, j > 0$ such that $x^i = y^j$.

(4) $xy = yx$.

(5) There exist integers $r, s > 0$ such that $x^r y^s = y^s x^r$.

(6) Define the morphism $h : \{a, b\}^* \to \Sigma^*$: $h(a) = x, h(b) = y$. Then there exist two distinct words, $u, v \in \{a, b\}^*$ such that $h(u) = h(v)$.

(7) $x\{x, y\}^* \cap y\{x, y\}^* \neq \varnothing$.

(8) $x\{x, y\}^\omega \cap y\{x, y\}^\omega \neq \varnothing$.

Proof:

(1) ⇒ (2) $x^\omega = (z^k)^\omega = z^\omega = (z^\ell)^\omega = y^\omega$.

(2) ⇒ (3) Let $i = |y|, j = |x|$. Then consider the prefix of length ij of x^ω and y^ω. They have to be the same, and this implies $x^i = y^j$.

(3) ⇒ (4) WLOG, assume $|y| \leq |x|$. Then there exists w such that $x = yw$. Then note that

$$y^j = x^i = (yw)^i = y^j = y(wy)^{i-1}w$$

Take off y at the front:

$$y^{j-1} = (wy)^{i-1}w$$

Add y at the back:

$$y^j = (wy)^{i-1}wy = (wy)^i$$

Observe that

$$(yw)^i = (wy)^i$$

Look at the first $|y| + |w|$ symbols gives us $yw = wy$. Then sub $x = yw$

$$x = yw = wy$$

Then append y at the back

$$xy = (yw)y = y(wy) = yx$$

(4) ⇒ (5)

(5) ⇒ (6)

(6) ⇒ (7)

(7) ⇒ (8)

□