Words and Languages (part I)

Prof. Dan A. Simovici

UMB

1. Words

2. Prefixes, Suffixes, and Infixes

3. Numbering Words

An alphabet is a finite, nonempty set. We refer to the elements of an alphabet as the *symbols* of the alphabet.

Example

The sets $B = \{0, 1\}$, $\{a, b, c, \ldots, z\}$ are alphabets.
The set $\mathbb{N} = \{0, 1, 2, \cdots\}$ is not an alphabet.

Definition

A word of length n on an alphabet A is a sequence of length n of symbols of this alphabet (i.e., an element of $\mathbf{Seq}_n(A)$).
The length of a word w is denoted by $|w|$.

We write $w(0) \cdots w(n-1)$ for the word $(w(0), \ldots, w(n-1))$, where $w(0), \ldots, w(n-1)$ are symbols of an alphabet A.

Example

Let $A = \{a, b, c\}$ be an alphabet. The word (a, a, b, a, a, c) is written as $aabaac$.

Any C program is a word over the basic alphabet of this language that includes small and capital letters, as well as special symbols, such as parentheses, brackets, braces, spaces, new line characters, quotation marks, etc. Not all these characters are visible; in other words, some characters (such as spaces) appear on paper only as white spaces. For example, the famous C program

```
#include <stdio.h> main() { printf("hello, world\n"); }
```

can be looked at as a word.

Under the notation previously introduced, a word w of length 1 is denoted by $w(0)$. Hence, we use the same notation for a symbol from the alphabet and the word of length one whose entry is the symbol.

Counting Words

If A is an alphabet with $|A|$ symbols, the number of words of length n is $|A|^n$.

In particular, there exists exactly 1 word of length 0, namely, the null word that contains no symbols. This word is denoted by λ.

The length of a word w is denoted by $|w|$.

The reversal w^R of a word w is

$$w^R = \begin{cases} \lambda & \text{if } w = \lambda \\ a_{n-1} \cdots a_0 & \text{if } w = a_0 \cdots a_{n-1} \end{cases}$$

A palindrome is a word w such that $w^R = w$.

Example

The word $w = abba$ over the alphabet $A = \{a, b\}$ is a palindrome.

If A is an alphabet, then we write A^* for the set of all words over A, and we write A^+ for $\textbf{Seq}^+(A)$, the set of all non-null words over A. Note that $A^* = A^+ \cup \{\lambda\}$.
Also, we have

$$\begin{aligned} A^* &= \bigcup_{n=0}^{\infty} A^n, \\ A^+ &= \bigcup_{n=1}^{\infty} A^n, \end{aligned}$$

Word Concatenation

Definition

Let $u = u_0 u_1 \cdots u_{m-1}$ and $v = v_0 v_1 \cdots v_{n-1}$ be two words, where $u_0, \ldots, u_{m-1}, v_0, \ldots, v_{n-1}$ are symbols over the alphabet A. The concatenation or the product of u and v is the word uv defined as

$$uv = u_0 u_1 \cdots u_{m-1} v_0 v_1 \cdots v_{n-1}.$$

Note that:

- $vu = v_0 v_1 \cdots v_{n-1} u_0 \cdots u_{m-1}$, so $|uv| = |vu| = |u| + |v|$.
- $uv \neq vu$ in general; indeed, if $u = abb$ and $v = ba$ observe that $uv = abbba$ and $vu = baabb$.

The word concatenation is associative, that is,

$$u(vw) = (uv)w$$

for every $u, v, w \in A^*$.
Indeed suppose that

$$\begin{aligned} u &= u_0 u_1 \cdots u_{p-1}, \\ v &= v_0 u_1 \cdots v_{q-1}, \\ w &= w_0 w_1 \cdots w_{r-1}. \end{aligned}$$

We have

$$\begin{aligned} u(vw) &= u v_0 v_1 \cdots v_{q-1} w_0 w_1 \cdots w_{r-1} \\ &= u_0 u_1 \cdots u_{p-1} v_0 v_1 \cdots v_{q-1} w_0 w_1 \cdots w_{r-1}, \\ (uv)w &= u_0 u_1 \cdots u_{p-1} v_0 u_1 \cdots v_{q-1} w \\ &= u_0 u_1 \cdots u_{p-1} v_0 v_1 \cdots v_{q-1} w_0 w_1 \cdots w_{r-1}, \end{aligned}$$

Powers of Words

Let A be an alphabet and let $x \in A^*$. The powers of x are defined as

$$\begin{aligned} x^0 &= \lambda, \\ x^{n+1} &= x^n x. \end{aligned}$$

Note that for every $x \in A^*$ and $n \in \mathbb{N}$ we have $|x^n| = n\,|x|$. Thus, $|\lambda^n| = 0$.

Occurrences of Symbols in Words

Let $u = u_0 u_1 \cdots u_{n-1} \in A^*$ and let $a \in A$ be a symbol. The number of occurrences of a in A is

$$n_a(u) = |\{j \mid u_j = a\}|.$$

Example

Let $A = \{a, b, c\}$ be an alphabet and let $u = aababbb \in A^*$. We have

$$n_a(u) = 3, n_b(u) = 4, \text{ and } n_c(u) = 0.$$

If $u, v \in A^*$ and $a \in A$, we have

$$n_a(uv) = n_a(u) + n_a(v).$$

Definition

Let t be a word, $t \in A^*$. A *prefix* of t is a word u such that $t = uw$ for some $w \in A^*$. The prefix u is *proper* if $u \neq \lambda$ and $u \neq t$.
A *suffix* of t is a word y such that $t = xy$ for some $x \in A^*$. The suffix y is *proper* if $y \neq \lambda$ and $y \neq t$.
An *infix* of t is a word w such that $t = vwx$ for some $v, x \in A^*$. The infix w is *proper* if $w \neq \lambda$ and $w \neq t$.

Notations

The sets of prefixes, infixes and suffixes of a word t are denoted by $\mathrm{PREF}(t)$, $\mathrm{INFIX}(t), \mathrm{SUFF}(t)$ the set of infixes, the set of infixes, and the set of suffixes of t, respectively.
Also, the sets of proper prefixes, proper infixes and proper suffixes of a word t are denoted by $\mathrm{PREFpr}(t)$, $\mathrm{INFIXpr}(t)$, $\mathrm{SUFFpr}(t)$, respectively.

Example

Let $A = \{a, b, c\}$, and consider the word $t = accabac$. The word acc is a prefix of t, $abac$ is a suffix of t, and $ccab$ is an infix of t.

Since $\lambda x = \lambda x \lambda = x \lambda = x$, every word x is a prefix, a suffix, and an infix of itself. Similarly, the null word is a prefix, an infix, and a suffix of every word.

We denote by $x_{i,j}$ the infix $y = a_i \cdots a_{j-1}$ of $x = a_0 \cdots a_{n-1}$ for $0 \leq i < j \leq n$.

In other words, $x_{i,j}$ is the infix that begins with a_i and ends with a_{j-1}.

We extend this notation by defining $x_{i,j} = \lambda$ when $j \leq i$. Thus, $x_{i,i} = \lambda$ for every i, $0 \leq i \leq n-1$; $x_{0,j} \in \text{PREF}(x)$ for $0 \leq j \leq n$; and $x_{i,n} \in \text{SUFF}(x)$ for $0 \leq i \leq n$. If $i \leq j$, we have $|x_{i,j}| = j - i$.

Definition

Let $x_{i,j}, x_{p,q}$ be two infixes of a word x. Then, $x_{i,j}, x_{p,q}$ are *disjoint* if $j \leq p$.

Example

Let $x = abacbabbbabaca$ be a word over the alphabet $\{a, b, c\}$. The infixes $x_{2,7} = acbab$ and $x_{7,10} = bab$ are disjoint.

Let $A = \{a_0, \ldots, a_{n-1}\}$ be an alphabet containing n symbols. Words over A can be encoded as natural numbers; in other words, we can define a bijection $\phi_A : A^* \longrightarrow \mathbb{N}$ by

$$\phi_A(x) = \begin{cases} 0 & \text{if } x = \lambda \\ n\phi_A(y) + i + 1 & \text{if } x = ya_i \end{cases}$$

for every $x \in A^*$. It is easy to verify that

$$\begin{aligned} \phi_A(a_{i_0} \cdots a_{i_{k-1}}) &= n^{k-1}(i_0 + 1) + n^{k-2}(i_1 + 1) + \cdots \\ &\quad + n(i_{k-2} + 1) + i_{k-1} + 1. \end{aligned}$$

Example

If $A = \{a_0, a_1, a_2\}$, $x = a_0 a_1 a_0 a_2$, and $y = a_2 a_2 a_2$, then

$$\begin{aligned} \phi_A(x) &= 3^3 \cdot 1 + 3^2 \cdot 2 + 3^1 \cdot 1 + 3 = 51 \\ \phi_A(y) &= 3^2 \cdot 3 + 3^1 \cdot 3 + 3 = 39. \end{aligned}$$

- Note that $\phi_A(x)$ is not the representation of the number $i_0 \cdots i_{k-1}$ in base n, since we are using the "digits" $1, 2, \ldots, n$ rather than $0, 1, \ldots, n-1$.

Let A^k be the set of words of length k over the alphabet $A = \{a_0, \dots, a_{n-1}\}$.
The least value of $\phi_A(x)$ for $x \in A^k$ is:

$$n^{k-1} + n^{k-2} + \cdots + n + 1 = \frac{n^k - 1}{n - 1}$$

The largest value of $\phi_A(x)$ for $x \in A^k$ is:

$$n^k + n^{k-1} + \cdots + n = \frac{n^{k+1} - n}{n - 1}$$

Thus, there are

$$\frac{n^{k+1} - n}{n - 1} - \frac{n^k - 1}{n - 1} + 1 = n^k$$

consecutive natural numbers in the interval $\left[\frac{n^k-1}{n-1}, \frac{n^{k+1}-n}{n-1}\right]$ codes of words in A^k, that is, a number equal to $|A^k|$.

Let I_k be the set of natural numbers in the interval $\left[\frac{n^k-1}{n-1}, \frac{n^{k+1}-n}{n-1}\right]$. For the consecutive sets I_k and I_{k+1} we have

$$\begin{aligned} I_k &= \left\{ n \middle| n \in \left[\frac{n^k-1}{n-1}, \frac{n^{k+1}-n}{n-1} \right] \right\} \\ I_{k+1} &= \left\{ n \middle| n \in \left[\frac{n^{k+1}-1}{n-1}, \frac{n^{k+2}-n}{n-1} \right] \right\}. \end{aligned}$$

Since

$$\frac{n^{k+1}-n}{n-1} < \frac{n^{k+1}-1}{n-1},$$

it follows that $I_k \cap I_{k+1} = \emptyset$ for $k \in \mathbb{N}$.

Note that:

- for any word $x \in A^k$ we have $\phi_A(x) \in I_k$;
- no two distinct words in A^k can be mapped into the same number in I_k, so ϕ_A defines an injection of A^k into I_k;
- for every number $m \in I_k$ there is a word $x \in A^k$ such that $\phi_A(x) = m$.

Therefore, ϕ_A is a bijection between A^* and $\mathbb{N}$.