Languages

ian.mcloughlin@gmit.ie

Alphabets and strings

In some contexts:

- Sets are called alphabets.
- Tuples over alphabets are called strings or words.
- We omit the brackets and commas when we can.
- The empty tuple is called the empty string and denoted ϵ.

Example

Let A be the alphabet $\{0, 1\}$. Each of the following is a string over that alphabet:

$$\epsilon, 0, 1, 00, 01, 10, 11, 001, 010, 011, 100, \ldots$$

Concatenation of strings

We can take two strings over the same alphabet and concatenate them.

Example

- Let $s_1 = 00$ and $s_2 = 101$ be two strings over $\{0, 1\}$.
- Concatenating 00 and 101 gives $s_1 \circ s_2 = 00101$.
- Technically, $(0, 0) \circ (1, 0, 1) = (0, 0, 1, 0, 1)$.
- When we can, we omit the notation: $s_1 s_2 = 00101$.
- Also, $s_2 s_1 = 10100$ which is a different string.
- Concatenation is not commutative.

Kleene star of an alphabet

$$A = \{0, 1\}$$

$A^1 = \{0, 1\}$, the strings of length one over A.

$A^2 = \{00, 01, 10, 11\}$, the strings of length two over A.

$A^i =$ the strings of length i over A, $i \in \mathbb{N}_0$.

$A^0 = \{\epsilon\}$, the strings of length zero over A.

Definition

The Kleene star of A is the union of all the A^i:

$$A^* = \bigcup_i A^i = \{\epsilon, 0, 1, 00, 01, 10, \ldots\}$$

Note this applies to any alphabet, not just $\{0, 1\}$*.*

Languages

$$L \subseteq A^*$$

A subset of the star of an alphabet is a **language** over it.

Example

- Let $A = \{a, b, c\}$.
- Then $A^* = \{\epsilon, a, b, c, aa, ab, ac, ba, bb, bc, \ldots\}$.
- $L_1 = \{aa, bbb, ccc\}$ is a language.
- $L_2 = \{s \mid s \text{ contains an } a\}$ is also a language.
- Read this as "all strings s where s contains an a".

We can create new languages from smaller ones using union, concatenate and star.

Union of languages

$$L_1 = \{00, 11\} \quad L_2 = \{11, 111, 1111\}$$

$$L_1 \cup L_2 = \{00, 11, 111, 1111\}$$

- The union of languages is the set of all strings in any of them.
- We could consider two languages over different alphabets, if we took the union of the alphabets – we usually don't.

Concatenation of languages

$$L_1 = \{00, 11\} \quad L_2 = \{11, 111, 1111\}$$

$$L_1 \circ L_2 = \{0011, 00111, 00111111, 1111, 11111, 11111111\}$$

- The concatenation of two languages is the set of concatenations of each string from the first language with each string from the second language.

- Note, usually $L_1 \circ L_2 \neq L_2 \circ L_1$. When are they equal?

- We usually omit the circle: $L_1 L_2$.

Star of languages

$$L = \{00, 11\}$$

$$L^* = \{\epsilon, 00, 11, 0000, 0011, 1100, 1111, \ldots\}$$

- The star of a language is the same idea as the star of an alphabet.
- L^1 is the language itself.
- L^2 is the language concatenated to itself.
- L^3 is the language concatenated to itself twice.
- L^0 is the language containing only ϵ.
- The star is the union of L^i for all $i \in \mathbb{N}_0$.
- We also define L^+ as the union of L^i for all $i \in \mathbb{N}$.

File types as languages

- The set of all valid pdf files is a language over the alphabet $A = \{0, 1\}$.
- So, the set of valid pdf's is a subset of A^*.
- As is the set of valid docx files.
- A computer program that converts pdf's to docx's maps one subset of A^* to another.
- Remember too, that executable files themselves are stored as files in 0's and 1's.
- In fact, all files are in $\{0, 1\}^*$.