Théorie des Langages Rationnels

A Framework for Words

Adrien Pommellet, LRE

January 25, 2023

What is a Language?

A few examples

Spoken languages. Made of meaningful sequences of sounds.
The spoken French language.

Written languages. Made of meaningful sequences of characters.
The written French language.

Computer languages. Made of sequences of keywords interpreted as a program.
Valid C source codes.

As computer scientists, we will focus on **the last case**.

What is a Language?

A preliminary definition

A **language** is a set of sequences of elementary objects to which we ascribe meaning.

What is a Language?

A curriculum

Théorie des langages rationnels. **Describe** somewhat simple languages and **recognize** their elements.
This is a valid C identifier.

Théorie des langages. Describe more complex languages, recognize their elements, and **analyze their structure**.
This is a syntactically correct C program.

TIGER project. Recognize, analyze, and **interpret** complex languages in order to build a compiler.
I can compile this valid C program.

Formalizing Languages

Every language depends on a limited set of **elementary symbols**.

Alphabet

It is a **finite set** Σ of symbols. We call the elements of Σ **letters**.

Remember that a **set** is a collection of **distinct** elements: $\{a, b, c\}$ is a set but $\{0, 0, 1, 2, 3, 3\}$ is not.

As an example, we can consider the following alphabets:

Binary. $\Sigma = \{0, 1\}$.

Digits. $\Sigma = \{0, 1, 2, \ldots, 9\}$.

Latin letters. $\Sigma = \{a, \ldots, z, A, \ldots, Z\}$.

Formalizing Languages

Words

Word

A word w over an alphabet Σ is a (possibly empty) **finite sequence** of letters. The **empty word** is written ε.

The set of all words

The set Σ^* is the set of **all** words over Σ, and the set $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$, the set of all **non-empty** words over Σ.

As an example, if $\Sigma = \{0, 1\}$, then $011101 \in \Sigma^*$, as it is a word over Σ.

Language

A language L over an alphabet Σ is a set of words over Σ.

Note that L may be finite or infinite. Moreover, $L \subseteq \Sigma^*$, that is, L is a **subset** of the set of all words.

Given $\Sigma = \{0, 1\}$, the set $L = \{w \in \Sigma^* \mid w \text{ has an even number of } 1\}$ is an (infinite) language.

Length of a word

Given a word w over an alphabet Σ, its length $|w|$ is equal to its **total** number of letters. In particular, $|\varepsilon| = 0$.

Occurrences of a letter

Given a word w over an alphabet Σ and a letter $a \in \Sigma$, $|w|_a$ stands for the number of occurrences of a in w.

If $\Sigma = \{0, 1\}$ and $w = 011101$, then $|w| = 6$ and $|w|_1 = 4$.

Practical Application

Exercise 1. What is the length of the word $w = CatCatDog$?

Answer

We can't answer this question **without knowing the alphabet** Σ:

- If it is the Latin alphabet, $|w| = 9$.
- If $\Sigma = \{Cat, Dog\}$, $|w| = 3$.
- if $\Sigma = \{0, 1\}$, w is not even a word.

Letters are **arbitrary symbols**: explicit the alphabet if there is any ambiguity.

Concatenation

Given two words $w_1 = a_1 \ldots a_n, w_2 = b_1 \ldots b_m$ over the alphabet Σ, we define their concatenation $w_1 \cdot w_2 = a_1 \ldots a_n b_1 \ldots b_m$.

If $\Sigma = \{a, \ldots, z\}$, $w_1 = abc$, and $w_2 = def$, then $w_1 \cdot w_2 = abcdef$.

Operations on Words

Properties of concatenation

For all words w_1, w_2, w_3 over the alphabet Σ, the following properties hold:

- $|w_1 \cdot w_2| = |w_1| + |w_2|$.
- $\varepsilon \cdot w_1 = w_1 \cdot \varepsilon = w_1$. The empty word ε is said to be a **neutral element** for concatenation.
- $w_1 \cdot (w_2 \cdot w_3) = (w_1 \cdot w_2) \cdot w_3$. Concatenation is said to be **associative**: we may either compute $w_1 \cdot w_2$ or $w_2 \cdot w_3$ first, it doesn't make a difference.
- It is possible that $w_1 \cdot w_2 \neq w_2 \cdot w_1$. Indeed, consider $\Sigma = \{0, 1\}$, $w_1 = 0$, and $w_2 = 1$. Concatenation is not **commutative**.

Operations on Words

Powers

Exponentiation of a word

Given an integer $k \in \mathbb{N}$ and a word w over an alphabet Σ, we define the words:

- $w^0 = \varepsilon$.
- $w^k = \underbrace{w \cdot \ \cdots \ \cdot w}_{k \text{ times}}$.

Obviously, $\forall k_1, k_2 \in \mathbb{N}$, $w^{k_1} \cdot w^{k_2} = w^{k_1 + k_2}$, hence the **power notation**.

Prefixes, suffixes, and factors

Given four words w, x, y, z over the alphabet Σ, we say that:

- x is a prefix of w if $w = x \cdot y$. We then write $x \in \mathrm{Pref}(w)$.
- z is a suffix of w if $w = y \cdot z$. We then write $z \in \mathrm{Suff}(w)$.
- y is a factor of w if $w = x \cdot y \cdot z$. We then write $y \in \mathrm{Fact}(w)$.

As an example, if $w = abcde$, then abc is a prefix, $bcde$ is a suffix, and cd is a factor.

Practical Application

Exercise 2. Consider the word $w = SUP$ on the Latin alphabet. Compute $\mathrm{Pref}(w)$, $\mathrm{Suff}(w)$, and $\mathrm{Fact}(w)$.

Answer

$\text{Pref}(w) = \{\varepsilon, S, SU, SUP\}$, $\text{Suff}(w) = \{\varepsilon, P, UP, SUP\}$, and $\text{Fact}(w) = \{\varepsilon, S, U, P, SU, UP, SUP\}$. Don't forget ε!

For any word w over the alphabet Σ, the following properties hold:

- $\varepsilon \in \mathit{Pref}(w)$, $\varepsilon \in \mathrm{Suff}(w)$, $\varepsilon \in \mathit{Fact}(w)$.
- $w \in \mathrm{Pref}(w)$, $w \in \mathrm{Suff}(w)$, $w \in \mathrm{Fact}(w)$.
- $\mathrm{Pref}(w) \subseteq \mathrm{Fact}(w)$ and $\mathrm{Suff}(w) \subseteq \mathrm{Fact}(w)$.
- $\mathrm{Fact}(w) = \mathrm{Pref}(\mathrm{Suff}(w)) = \mathrm{Suff}(\mathrm{Pref}(w))$. We either first remove the tail then the head to create a factor, or the other way round.

Mirror

Let $w = w_1 \dots w_n$ be a word of length n over an alphabet Σ. We then define its mirror $w^R = w_n \dots w_1$.

If a word w is its own mirror, that is, $w = w^R$, then it is called a **palindrome**. Consider as an example *radar*, *madam*, or *rotator*.

Is it possible to measure **how much** two words differ?

We will generalize the mathematical notion of distance.

Distance

Let E be a set. A function $d : E^2 \to \mathbb{R}_+$ is said to be a distance if it verifies the following properties $\forall x, y, z \in E$:

- Separation. $d(x, y) = 0 \iff x = y$.
- Symmetry. $d(x, y) = d(y, x)$.
- Triangle inequality. $d(x, y) + d(y, z) \geq d(x, z)$.

As an example, consider the usual distance between two points of $\mathbb{R}^2$.

We introduce a distance on Σ^*.

Edit distance

The edit distance $d_e(w_1, w_2)$ between two words w_1 and w_2 in Σ^* is equal to the **minimal** number of single letter insertions and deletions needed to turn w_1 into w_2

As an example, $d_e(\text{dog}, \text{bugs}) = 5$.

$$\text{dog} \overset{-o}{\rightarrow} \text{dg} \overset{+u}{\rightarrow} \text{dug} \overset{-d}{\rightarrow} \text{ug} \overset{+b}{\rightarrow} \text{bug} \overset{+s}{\rightarrow} \text{bugs}$$

Practical Application

Exercise 3. Compute $d_e(\mathrm{EPITA}, \mathrm{EPUISE})$.

Answer

$$\text{EPITA} \xrightarrow{-T} \text{EPIA} \xrightarrow{-A} \text{EPI} \xrightarrow{+U} \text{EPUI} \xrightarrow{+S} \text{EPUIS} \xrightarrow{+E} \text{EPUISE}$$

$$d_e(\text{EPITA}, \text{EPUISE}) = 5$$