Einführung in die Computerlinguistik – reguläre Sprachen und endliche Automaten

Dozentin: Wiebke Petersen

Foliensatz 3
Describing formal languages by enumerating all words

- Peter says that Mary has fallen off the tree.
- Oskar says that Peter says that Mary has fallen off the tree.
- Lisa says that Oskar says that Peter says that Mary has fallen off the tree.
- ...
Describing formal languages by enumerating all words

- Peter says that Mary has fallen off the tree.
- Oskar says that Peter says that Mary has fallen off the tree.
- Lisa says that Oskar says that Peter says that Mary has fallen off the tree.
- . . .

The set of strings of a natural language is infinite.
The enumeration does not gather generalizations.
Describing formal languages by grammars

Grammar

- A formal grammar is a generating device which can generate (and analyze) strings/words.
- Grammars are finite rule systems.
- The set of all strings generated by a grammar is the formal language generated by the grammar.

\[
S \rightarrow NP \ VP \\
VP \rightarrow V \\
NP \rightarrow D \ N \\
D \rightarrow \text{the} \\
N \rightarrow \text{cat} \\
V \rightarrow \text{sleeps}
\]

Generates: the cat sleeps
Describing formal languages by automata

Automaton

- An automaton is a **recognizing device** which accepts strings/words.
- The set of all strings accepted by an automaton is the formal language accepted by the automaton.
Zusammenhang nach Klabunde 1998


Vorsicht: per Definition besitzen formale Sprachen keine strukturellen Eigenschaften; uns interessieren aber nur solche mit strukturellen Eigenschaften, die von einer Grammatik erzeugt werden können. Außerdem können Grammatiken auch für die Analyse (Erkennung) formaler Sprachen und endliche Automaten für ihre Erzeugung genutzt werden.
The set of regular expressions $RE_\Sigma$ over an alphabet $\Sigma = \{a_1, \ldots, a_n\}$ is defined by:

- $\emptyset$ is a regular expression.
- $\varepsilon$ is a regular expression.
- $a_1, \ldots, a_n$ are regular expressions.
- If $a$ and $b$ are regular expressions over $\Sigma$ then
  - $(a + b)$
  - $(a \cdot b)$
  - $(a^*)$

are regular expressions too.

(The brackets are frequently omitted w.r.t. the following dominance scheme: $\star$ dominates $\bullet$ dominates $+$)
Regular expressions

**RE: semantics**

Each regular expression $r$ over an alphabet $\Sigma$ describes a formal language $L(r) \subseteq \Sigma^*$. Regular languages are those formal languages which can be described by a regular expression. The function $L$ is defined inductively:

- $L(\emptyset) = \emptyset$, $L(\varepsilon) = \{\varepsilon\}$, $L(a_i) = \{a_i\}$
- $L(a + b) = L(a) \cup L(b)$
- $L(a \cdot b) = L(a) \circ L(b)$
- $L(a^*) = L(a)^*$
Exercise 1

Find a regular expression which describes the regular language $L$ (be careful: at least one language is not regular!)

- $L$ is the language over the alphabet $\{a, b\}$ with
  $L = \{aba, \varepsilon, aa, bbb\}$.
- $L$ is the language over the alphabet $\{a, b\}$ which consists of all words which start with a nonempty string of $b$’s followed by at least one $a$ followed by any number of $b$’s.
- $L$ is the language over the alphabet $\{a, b\}$ such that every $a$ has a $b$ immediately to its left.
- $L$ is the language over the alphabet $\{a, b\}$ which consists of all words which contain an uneven number of $a$’s.
- $L$ is the language of all palindromes over the alphabet $\{a, b\}$.
Deterministic finite-state automaton (DFSA)

Definition

A deterministic finite-state automaton is a tuple \( \langle Q, \Sigma, \delta, q_0, F \rangle \) with:

1. a finite, non-empty set of states \( Q \)
2. an alphabet \( \Sigma \) with \( Q \cap \Sigma = \emptyset \)
3. a partial transition function \( \delta : Q \times \Sigma \rightarrow Q \)
4. an initial state \( q_0 \in Q \) and
5. a set of final/accept states \( F \subseteq Q \).

accepts: \( L(a^*ba^*) \)
partial/total transition function

FSA with partial transition function

accepts $ab^*a$

transition table
partial/total transition function

FSA with partial transition function

FSA with complete transition function

accepts $ab^*a$

transition table
Example DFSA / NFSA

The language $L(ab^* + ac^*)$ is accepted by
Definition

A **nondeterministic finite-state automaton** is a tuple $\langle Q, \Sigma, \Delta, q_0, F \rangle$ with:

1. a finite non-empty set of states $Q$
2. an alphabet $\Sigma$ with $Q \cap \Sigma = \emptyset$
3. a **transition relation** $\Delta \subseteq Q \times \Sigma \times Q$
4. an initial state $q_0 \in Q$ and
5. a set of **final states** $F \subseteq Q$.

Theorem

A language $L$ can be accepted by a DFSA iff $L$ can be accepted by a NFSA.

Note: Even automatons with $\varepsilon$-transitions accept the same languages like DFSA’s.
Automaton with $\varepsilon$-transition
Exercise 2

Give an FSA for each of the following languages over the alphabet \{a, b\} (and try to make it deterministic):

- \(L = \{w \mid \text{between each two } 'a's \text{ in } w \text{ there are at least three } 'b's\}\)
- \(L = \{w \mid w \text{ is any word except } "bab"\}\)
- \(L = \{w \mid w \text{ does not contain the infix } "ab"\}\)
- \(L = \{w \mid w \text{ contains at most two } 'a's\}\)
- \(L = \{w \mid w \text{ contains an uneven number of } 'a's\}\)
- \(L((b^*a)^*ab^*)\)
- \(L(a^*(ab)^*)\)
- \(L(aa^*b).\)
- \(L((bb^* + ab^*a))\)