Automatentheorie und formale Sprachen
reguläre Ausdrücke

Dozentin: Wiebke Petersen

6.5.2009

Formal language

Definition

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

Formal language

Definition

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

Examples:

- language L_{pal} of the palindromes in English
  $L_{pal} = \{\text{mum, madam, } \ldots\}$
- the empty set
- the set of words of length 13 over the alphabet $\{a, b, c\}$

Formal language

Definition

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

Examples:

- language L_{pal} of the palindromes in English
  $L_{pal} = \{\text{mum, madam, } \dots\}$
- the empty set
- the set of words of length 13 over the alphabet $\{a, b, c\}$
- English?

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.

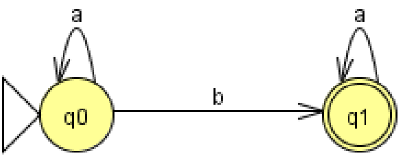
$$\begin{array}{rcl@{\qquad}rcl@{\qquad}rcl} \mathrm{S} & \rightarrow & \mathrm{NP\ VP} & \mathrm{VP} & \rightarrow & \mathrm{V} & \mathrm{NP} & \rightarrow & \mathrm{D\ N} \\ \mathrm{D} & \rightarrow & \mathrm{the} & \mathrm{N} & \rightarrow & \mathrm{cat} & \mathrm{V} & \rightarrow & \mathrm{sleeps} \end{array}$$

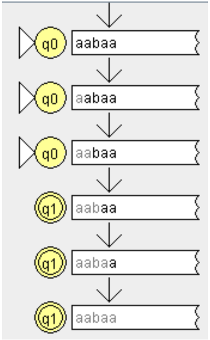
Generates: the cat sleeps

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.





Sprachbeschreibung

Zusammenhang nach Klabunde 1998

"Formale Sprachen besitzen strukturelle Eigenschaften.

Grammatiken sind Erzeugungssysteme für formale Sprachen.

Automaten sind Erkennungssysteme für formale Sprachen."

Vorsicht: per Definition besitzen formale Sprachen keine strukturellen Eigenschaften; uns interessieren aber nur solche mit strukturellen Eigenschaften, die von einer Grammatik erzeugt werden können.

Regular expressions

RE: syntax

The set of regular expressions RE_Σ over an alphabet $\Sigma = \{a_1, \ldots, a_n\}$ is defined by:

- $\underline{\emptyset}$ is a regular expression.
- ϵ is a regular expression.
- $a_1, \ldots, a_n$ are regular expressions
- If a and b are regular expressions over Σ then
  - $(a + b)$
  - $(a \bullet b)$
  - $(a^\star)$

  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 Σ 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:

- $L(\underline{\emptyset}) = \emptyset$, $L(\epsilon) = \{\epsilon\}$, $L(a_i) = \{a_i\}$
- $L(a + b) = L(a) \cup L(b)$
- $L(a \bullet b) = L(a) \circ L(b)$
- $L(a^\star) = L(a)^*$

Exercise: regular expressions

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* $L = \{aa, \epsilon, ab, bb\}$.
- *L is the language over the alphabet* $\{a, b\}$ *which consists of all words which start with a nonempty string of a's 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 the right.*
- *L is the language over the alphabet* $\{a, b\}$ *which consists of all words which contain an even number of a's.*
- *L is the language of all palindromes over the alphabet* $\{a, b\}$.

What we know so far about formal languages

- Formal languages are sets of words (NL: sets of sentences) which are strings of symbols (NL: words).
- Everything in the set is a "grammatical word", everything else isn't.
- Some formal languages, namely the regular ones, can be described by regular expressions
  Example: $(a^\star \bullet b \bullet a^\star \bullet b \bullet a^\star)^\star$ is the regular language consisting of all words over the alphabet $\{a, b\}$ which contain an even number of b's.
- Not all formal languages are regular (We have not proven this yet!).
  Example: The formal language of all palindromes over the alphabet $\{a, b\}$ is not regular.

Exercise 2

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

1. $L = \{w|$ *between each two 'b's in w there are at least two 'a's*$\}$
2. $L = \{w|w$ *is any word except "ab"*$\}$
3. $L = \{w|w$ *does not contain the infix "ba"*$\}$
4. $L = \{w|w$ *contains at most three 'b's*$\}$
5. $L = \{w|w$ *contains an even number of 'a's*$\}$
6. $L((a^\star b)^\star ab^\star)$
7. $L(a^\star(bb)^\star)$
8. $L(ab^\star b)$.
9. $L((ab^\star + ba^\star a))$

Solve at least 4 tasks (2 out of 1-5 and 2 out of 6-9)