

Basic format for describing operations

Abstract

This document describes the syntax for three different judges. The first one is a judge that allows to construct regular languages by means of regularity-preserving operations applied on automata. The second one extends the previous judge by allowing also the introduction of context-free grammars, thus being able to define context-free languages. The third one is focused on reducing undecidable problems on context-free grammars.

1 Regular languages

Before formally detailing the format we introduce an illustrative example. The program

```
main {
  a = "a";
  b = "b";
  ab = a | b;
  even_a = "    a b
            q0 q1 q0 +
            q1 q0 q1 ";
  output ab* a ab ab & even_a;
}
```

defines a variable a whose language is $\{a\}$, a variable b whose language is $\{b\}$, a variable ab whose language is the union of the two previous variables, i.e., $\{a, b\}$, and a variable $even_a$ that is assigned an automaton recognizing the language $\{w \in \{a, b\}^* \mid |w|_a \in \hat{2}\}$. Finally, it outputs the language obtained by: the Kleene's closure of the variable ab , concatenated with the variable a , concatenated twice with the variable ab , and everything intersected with the variable $even_a$, i.e., the output is $\{w_1aw_2 \mid w_1, w_2 \in \{a, b\}^* \wedge |w_2| = 2 \wedge |w_1aw_2|_a \in \hat{2}\}$.

Now we precise the syntax of the programs. In order to describe a regular language it suffices to write a program of the following form:

```
main {
  <ident1> = <expr1>;
  ...
  <identN> = <exprN>;
  output <expr>;
}
```

where the $\langle ident1 \rangle, \dots, \langle identN \rangle$ are variable identifiers, and the $\langle expr1 \rangle, \dots, \langle exprN \rangle, \langle expr \rangle$ are expressions over regular languages using operators that preserve the regularity. More concretely, a basic expression is one of the following:

- A variable identifier (see **IDENTIFIER** below).
- A word (representing the language that contains exactly that word; see **WORD** below).
- A deterministic finite automaton (representing its recognized language; see **DFA** below).

Expressions can be combined using the following regularity-preserving operators:

- Language intersection, denoted by the binary operator $\&$.
- Language union, denoted by the binary operator $|$.
- Language subtraction, denoted by the binary operator $-$.

- Language concatenation, a binary operation denoted without any symbol.
- Kleene's closure of the language, denoted by the unary postfix operator `*`.
- Language reverse, denoted by the function `reverse(...)`.
- Language substitution, denoted by the function `substitution(...)`. This function has as a first parameter the expression defining the language where the substitution must be performed. The actual substitution is defined by the following parameters, each being of the form `WORD -> expr`, denoting that the image of the terminal symbol `WORD` is the language defined by the expression `expr` (the `WORD` must have one single symbol).

As another example, the following two programs define the language of words over the alphabet $\{a, b\}$ having the subword *abbba*:

```
main {
    output ("a"|"b")* "abbba" ("a"|"b")*;
}

main {
    // automaton recognizing words with at least one 'x'
    aux = "      a  b  x
          ini ini ini acc
          acc acc acc acc + ";
    output substitution(aux, "x" -> "abbba");
}
```

As a final remark, we provide the grammar describing the syntax of the programs:

```
program: 'main' '{' instruction* '}'

instruction: IDENTIFIER '=' expr ';'
           | 'output' expr ';'

expr: subtraction (('|'|'&') subtraction)*

subtraction: concatenation ('-' concatenation)*

concatenation: starred starred*

starred: basic ('*' | )

basic: IDENTIFIER
      | WORD
      | DFA
      | '(' expr ')'
      | 'reverse' '(' expr ')'
      | 'substitution' '(' expr (',' WORD '->' expr)* ')'
```

where:

- `IDENTIFIER` is a string over alphanumeric characters and underscore,
- `WORD` is a string delimited by quotation marks `"` and composed over lower-case letters, digits, and the special characters `+ - * / () [] > <`, and
- `DFA` is a string delimited by quotation marks `"` that describes a deterministic finite automaton.

