Теория автоматов и формальных языков

Введение

Лектор: Екатерина Вербицкая

НИУ-ВШЭ

5 сентября 2022

О чем этот курс?

Теория автоматов и формальных языков изучает:

- Математические модели для описания языков
- Абстрактные машины для работы с языками

Также рассматриваются:

- Подходы к описанию синтаксиса языков
- Подходы к описанию “смысла” программ и предложений
- Принципиальные ограничения механизмов для работы с языками

Какие бывают языки?

Какие бывают языки?

- Естественные
  - Русский, английский...

Какие бывают языки?

- Естественные
  - Русский, английский...
- Искусственные
  - Эсперанто, ложбан...
  - Клингонский, эльфийский...

Какие бывают языки?

- Естественные
  - Русский, английский...
- Искусственные
  - Эсперанто, ложбан...
  - Клингонский, эльфийский...
  - C++, Python, Java, C#, Haskell, OCaml, Perl, Coq, Agda...

Где можно встретить языки?

В повседневной жизни:

- при разговоре, в переписке
- на заборах, на стенах гробниц
- в собственной голове при формулировке мыслей...

Где можно встретить языки?

В повседневной жизни:

- при разговоре, в переписке
- на заборах, на стенах гробниц
- в собственной голове при формулировке мыслей...

При работе с различными **языковыми процессорами**:

- текстовыми редакторами
- компиляторами, интерпретаторами, трансляторами
- средами разработки...

Где можно встретить языки?

В повседневной жизни:

- при разговоре, в переписке
- на заборах, на стенах гробниц
- в собственной голове при формулировке мыслей...

При работе с различными **языковыми процессорами**:

- текстовыми редакторами
- компиляторами, интерпретаторами, трансляторами
- средами разработки...

Все нуждаются в **формализованном представлении** языка

Два аспекта спецификации языка программирования

- Синтаксис — правила построения программ из символов
  - Форма
- Семантика — правила истолкования программ
  - Смысл

Пример: русский язык

вы продоёте рыбов

- Синтаксис
  - ...
  - Порядок слов в предложении: подлежащее, сказуемое, дополнение
  - В конце вопросительного предложения ставится вопросительный знак
  - Дополнение выражается существительным в косвенном падеже без предлога
  - ...
- Семантика
  - Говорящий спрашивает, продаются ли рыбины

Пример: язык арифметических выражений

$$1 * (2 + 3)/4 - 5$$

- Синтаксис
  - **Терм**: последовательность цифр или любое **выражение** в скобках
  - **Слагаемое**: последовательность **термов**, соединненых знаками умножения и деления
  - **Выражение**: последовательность **слагаемых**, соединенных знаками сложения и вычитания (перед первым **слагаемым** может стоять минус)
- Семантика
  - Значение арифметического выражения

Пример: язык арифметических выражений

$$1 * (2 + 3)/4 - 5$$

- Синтаксис
  - **Терм**: последовательность цифр или любое **выражение** в скобках
  - **Слагаемое**: последовательность **термов**, соединненых знаками умножения и деления
  - **Выражение**: последовательность **слагаемых**, соединенных знаками сложения и вычитания (перед первым **слагаемым** может стоять минус)
- Семантика
  - Значение арифметического выражения
    - -3.75
    - -4

Пример: синтаксис if-выражений

```python
if temperature > 23:
  print('Wear shorts.')
else:
  print('Wear long pants.')
```

```cpp
if ( temperature > 23 ) {
  cout<<"Wear shorts.\n";
}
else
  cout<<"Wear long pants.\n";
}
```

```haskell
if temperature > 23
then print "Wear shorts."
else print "Wear long pants."
```

```lisp
(if (> temperature 23)
  (print "Wear shorts.")
  (print "Wear long pants."))
```

Что такое язык?

Что такое язык?

Язык — множество строк

Что такое множество?

Что такое множество?

Множество — набор уникальных элементов

Что такое множество?

Множество — набор уникальных элементов

- $x \in X$: x — элемент множества X (x принадлежит X)
- $x \notin X$: x не является элементом множества X (x не принадлежит X)
- Уникальность, неупорядоченность: $\{13, 42\} = \{42, 13\} = \{42, 13, 42\}$
- Универсальное множество (универсум $\mathcal{U}$): множество всех мыслимых объектов
    - $\mathbb{N} = \{1, 2, 3, \dots\}$
    - $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$
    - $\mathbb{Q} = \{m/n \mid m, n \in \mathbb{Z}; n \neq 0\}$

Подмножества

A является **подмножеством** B тогда и только тогда, когда все элементы A являются элементами B

$$A \subseteq B \iff \forall x : x \in A \Rightarrow x \in B$$

Подмножества

A является **подмножеством** B тогда и только тогда, когда все элементы A являются элементами B

$$A \subseteq B \iff \forall x : x \in A \Rightarrow x \in B$$

- $\{13, 42\} \subseteq \{7, 13, 37, 42, 99\}$
- $\{1, 3, 5, ...\} \subseteq \mathbb{N}$
- $\mathbb{N} \subseteq \mathbb{Z}$
- $\forall A : A \subseteq A$
- Пустое множество ($\varnothing$): множество без элементов
  - $\forall x : x \notin \varnothing$
  - $\forall A : \varnothing \subseteq A$

Подмножества

Множества A и B **равны** тогда и только тогда, когда A является подмножеством B и B является подмножеством A

$$A = B \iff A \subseteq B \text{ и } B \subseteq A$$

Подмножества

Множества A и B **равны** тогда и только тогда, когда A является подмножеством B и B является подмножеством A

$$A = B \iff A \subseteq B \text{ и } B \subseteq A$$

A является **строгим подмножеством** B тогда и только тогда, когда A является подмножеством B, но они не равны друг другу

$$A \subset B \iff A \subseteq B \text{ и } A \neq B$$

Подмножества

Множества A и B **равны** тогда и только тогда, когда A является подмножеством B и B является подмножеством A

$$A = B \iff A \subseteq B \text{ и } B \subseteq A$$

A является **строгим подмножеством** B тогда и только тогда, когда A является подмножеством B, но они не равны друг другу

$$A \subset B \iff A \subseteq B \text{ и } A \neq B$$

- $\forall x : \varnothing \subset \{x\}$
- $\mathbb{N} \subset \mathbb{Z}$
- $\mathbb{Z} \not\subset \mathbb{N}$
- $\forall A : A = A$ и $A \not\subset A$

Множество всех подмножеств (powerset)

Множество всех подмножеств множества A состоит из всех подмножеств A

$$2^A = \{B \mid B \subseteq A\}$$

- $\forall A : \varnothing \in 2^A$
- $\forall A : A \in 2^A$
- $A = \{0, 1\} \Rightarrow \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$

Множество всех подмножеств (powerset)

Множество всех подмножеств множества A состоит из всех подмножеств A

$$2^A = \{B \mid B \subseteq A\}$$

- $\forall A : \varnothing \in 2^A$
- $\forall A : A \in 2^A$
- $A = \{0, 1\} \Rightarrow \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$

Сколько элементов может быть в множестве всех подмножеств?

Операции над множествами

$$
\begin{aligned}
\text{Объединение:} \quad & A \cup B & = \quad & \{x \mid x \in A \text{ или } x \in B\} \\
\text{Пересечение:} \quad & A \cap B & = \quad & \{x \mid x \in A \text{ и } x \in B\} \\
\text{Разность:} \quad & A \setminus B & = \quad & \{x \mid x \in A \text{ и } x \notin B\} \\
\text{Дополнение:} \quad & \overline{A} & = \quad & \{x \mid x \in \mathcal{U} \text{ и } x \notin A\} = \mathcal{U} \setminus A
\end{aligned}
$$

Строки: неформально

Строки: неформально

Строка — последовательность символов

Алфавит

- **Алфавит** (Σ) — конечное множество (атомарных, неделимых) *символов*
  - $\{a, b, c, \ldots, z\}$
  - $\{\alpha, \beta, \gamma, \ldots, \omega\}$
  - $\{0, 1\}$
  - $\{$<u>include</u>, <u>for</u>, <u>if</u>, ...$\}$
  - $\{$<u>let</u>, <u>in</u>, <u>where</u>, ...$\}$

Цепочка

- **Цепочка (предложение, слово, *строка*)** — любая конечная последовательность символов алфавита
  - cat
  - $\kappa\alpha\tau$
  - 011000110110000101110100
  - `main = putStrLn . show . inc 2 where inc = \x -> x + 1`
- **Пустая цепочка** ε — цепочка, не содержащая ни одного символа
  - ε не является символом алфавита

Конкатенация строк

- **Конкатенация строк** α **и** β $(\alpha \cdot \beta = \alpha\beta)$ — результат приписывания строки β в конец строки α
  - $\forall \alpha\beta\gamma : (\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$
  - $\forall \alpha : \alpha \cdot \varepsilon = \varepsilon \cdot \alpha = \alpha$

Пример: арифметические выражения

- Алфавит $\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, *, /, (,)\}$
- $1 * (2 + 3)/4 - 5 =$
  $1 \cdot *(2 + 3)/4 - 5 =$
  $1 * (2 + 3) \cdot /4 - 5 =$
  $1 \cdot * \cdot (\cdot 2 \cdot + \cdot 3 \cdot) \cdot / \cdot 4 \cdot - \cdot 5 =$
  $1 * (2 + 3)/4 - 5 \cdot \varepsilon$
- Является ли ε арифметическим выражением?

Операции над строками

- **Обращение (реверс) цепочки** a^R — цепочка, символы которой записаны в обратном порядке
  - Если $x = abc$, то $x^R = cba$
  - $\varepsilon^R = \varepsilon$
- $n-$**я степень цепочки** a^n — конкатенация n повторений цепочки
  - $a^0 = \varepsilon$
  - $a^n = a \cdot a^{n-1} = a^{n-1} \cdot a$
- **Длина цепочки** $|a|$ — количество составляющих ее символов
  - $|babb| = 4$
  - $|babb|_a = 1, |babb|_b = 3, |babb|_c = 0$
  - $|\varepsilon| = 0$

Формальный язык

- Σ — алфавит
  - $\Sigma = \{0, 1\}$
- Σ^* — множество, содержащее все цепочки в алфавите Σ, включая пустую цепочку
  - $\Sigma^* = \{\varepsilon, 0, 1, 00, 11, 01, 10, 000, 001, 011, ...\}$
  - Сколько может быть элементов в Σ^*?

Формальный язык

- Σ — алфавит
  - $\Sigma = \{0, 1\}$
- Σ^* — множество, содержащее все цепочки в алфавите Σ, включая пустую цепочку
  - $\Sigma^* = \{\varepsilon, 0, 1, 00, 11, 01, 10, 000, 001, 011, ...\}$
  - Сколько может быть элементов в Σ^*?
- $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$
  - $\Sigma^+ = \{0, 1, 00, 11, 01, 10, 000, 001, 011, \ldots\}$
  - Сколько может быть элементов в Σ^+?

Формальный язык

- Σ — алфавит
  - $\Sigma = \{0, 1\}$
- Σ^* — множество, содержащее все цепочки в алфавите Σ, включая пустую цепочку
  - $\Sigma^* = \{\varepsilon, 0, 1, 00, 11, 01, 10, 000, 001, 011, ...\}$
  - Сколько может быть элементов в Σ^*?
- $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$
  - $\Sigma^+ = \{0, 1, 00, 11, 01, 10, 000, 001, 011, \dots\}$
  - Сколько может быть элементов в Σ^+?
- *Формальный язык* в алфавите Σ — подмножество множества всех цепочек в этом алфавите.
  - Для любого языка L (в алфавите Σ) справедливо $L \subseteq \Sigma^*$
  - $L = \{0, 00, 000, \dots\} \subset \{0, 1\}^*$
  - $L = \{0, 0101, 011011011, \dots\} \subset \{0, 1\}^*$

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - Язык металингвистических формул Бэкуса (БНФ)

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - Язык металингвистических формул Бэкуса (БНФ)
  - Синтаксические диаграммы

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - Язык металингвистических формул Бэкуса (БНФ)
  - Синтаксические диаграммы
  - Грамматики
  - ...

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - **Язык металингвистических формул Бэкуса (БНФ)**
  - Синтаксические диаграммы
  - Грамматики
  - ...

БНФ — Бэкуса-Наура форма

- **Символ** — элементарное понятие языка
  - $+$ означает сложение в языке арифметических выражений
- **Метапеременная** — сложное понятие языка
  - Переменной <выражение> можно обозначить выражение
- **Формула**
  - $<$определяемый символ$>::=<$посл.1 $>|\ldots|<$посл.n $>$
  - В правой части формулы — альтернатива конкатенаций строк, составленных из символов и метапеременных
- Пример: число
  - $<$число$>::=<$цифра$>|<$цифра$><$число$>$
  - $<$цифра$>::= 0 \mid 1 \mid \cdots \mid 9$

Расширенная форма Бэкуса Наура (EBNF)

- Более емкие операции
- **Итерация**
  - $<x> ::= \{<y>\}$ эквивалентно: $<x> ::= \varepsilon \mid <y><x>$
- **Условное вхождение**
  - $<x> ::= [<y>]$ эквивалентно: $<x> ::= \varepsilon \mid <y>$
- Скобки для группировки
  - $(<x>|<y>)\ <z>$ эквивалентно: $<x><z>|<y><z>$

Пример: арифметические выражения

- **Терм**: последовательность цифр или любое **выражение** в скобках
- **Слагаемое**: последовательность **термов**, соединненых знаками умножения и деления
- **Выражение**: последовательность **слагаемых**, соединенных знаками сложения и вычитания (перед первым **слагаемым** может стоять минус)

$$\begin{aligned} < expr > \quad &::= \quad [-] < factor > \{ (+ \mid -) < factor > \} \\ < factor > \quad &::= \quad < term > \{ (* \mid /) < term > \} \\ < term > \quad &::= \quad < number > \mid \ '(' < expr > ')' \end{aligned}$$

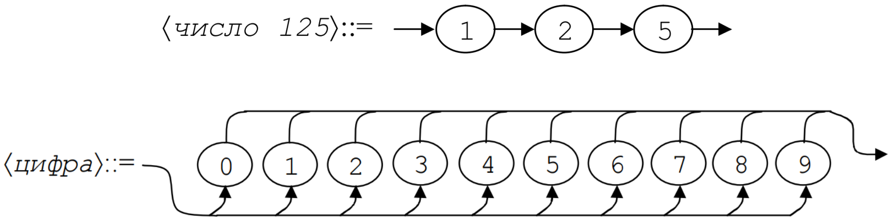
Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - Язык металингвистических формул Бэкуса (БНФ)
  - **Синтаксические диаграммы**
  - Грамматики
  - . . .

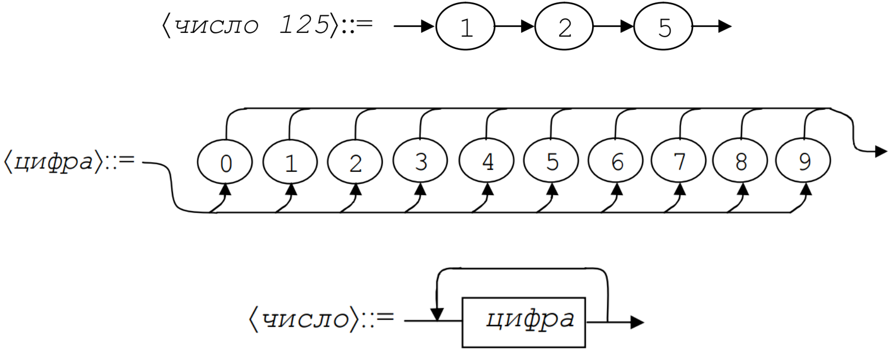
Синтаксические диаграммы Вирта

⟨*число 125*⟩::=

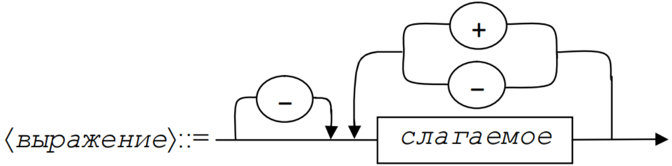
Синтаксические диаграммы Вирта



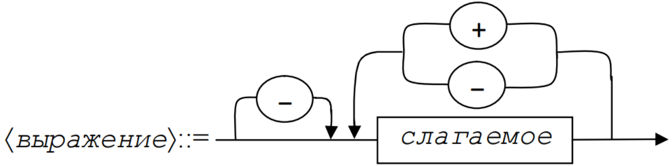
Синтаксические диаграммы Вирта

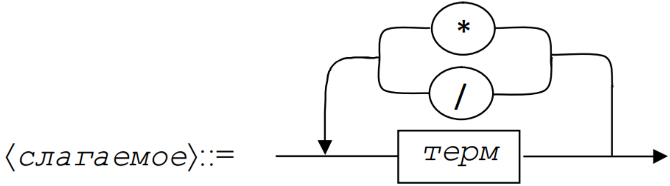


Синтаксические диаграммы Вирта



Синтаксические диаграммы Вирта





Синтаксические диаграммы Вирта