Tokenizing

Tasks of the Tokenizer

- Group the input (which is just a stream/string of characters) into tokens. (Numbers, operators, special words, strings)
- Eliminate comments.
- In *C* : Deal with `#include, #if #else #endif` and the like.

Tokenizers are also called scanners.

Tokens

Definition: A token type is a pair (Λ, T), in which Λ is a finite set of tokens labels, and T is a function s.t. for each $\lambda \in \Lambda$, $T(\lambda)$ is a set.

A token (with attribute) is a pair (λ, x), s.t. $\lambda \in \Lambda$ and $x \in T(\lambda)$.

Example: If one puts $\Lambda = \{\mathbf{int}, \mathbf{real}\}$, with
$T(\mathbf{int}) = \mathcal{Z}$, $T(\mathbf{real}) = \mathcal{R}$, then
$(\mathbf{real}, 3)$, $(\mathbf{real}, 3.141526535)$, $(\mathbf{int}, 2)$, $(\mathbf{int}, -1)$ are tokens.

$(\mathbf{int}, 2.718271828)$ is not a token.

Tokens without Attribute

Not all tokens have an attribute. For example reserved words **while**, **do**, **if**, **then**, **else** usually don't.

For those, one needs a trivial type $\top = \{(\ )\}$. Then
$T(\textbf{while}) = T(\textbf{do}) = T(\textbf{if}) = T(\textbf{then}) = T(\textbf{then}) = \top$.

Implementation Issues

I usually use C^{++}. A token is a struct containing an enum type, and a list for each possible type of attribute. The list has length 1 when the attribute has the type of the list, and 0 otherwise.

In C, one could use a struct containing an enum and a pointer to the heap, or an enum with a union type.

In Java, one could use a struct containing an enum and an Object.

Whatever implemention you choose, you should use an object-oriented approach. Make sure that there is a token class, make sure that it can be printed, that it can be passed to procedures, and put in containers.

Implementation (2)

It is a good idea to add information about where it comes from to a token. This makes it more easy to generate error messages.

A tokenizer is a function with signature
`token readtoken( reader& );` When called, it reads characters from reader until it has enough information to build a token.

The reader has a field `char nextchar;` and a method
`void moveforward( );` which replaces nextchar by the next char.

Building a Tokenizer

There are basically two ways of building a tokenizer.

- Hacking (sometimes also called 'careful coding by hand.') If the tokenizer is not big, you can follow this approach.
- Using a scanner generator. (Lex/Flex)

Writing a Tokenizer by Hand

If the tokens are not too many, you can follow this approach.

Draw an NDFA for each non-trivial token.

Stare long at the NDFAs and at the tokens for which you didn't draw an NDFA and find all overlaps.

Find ways of dealing with the overlaps. (Combine NDFAs with overlaps into one. First read one token, if NDFA gets stuck, read as another token. Do postprocessing of read tokens. Be careful not to become exponential.)

Overlaps

Sometimes, different tokens have shared prefixes.

An example is **int** and **real**. One can decide only at the end that 123453343433433434.5 is a **real**, and not an **int**.

Similarly, identifiers and reserved words overlap, like **while**, **do**, **whilethegrasskeepsgrowing**, **which**.

Operators `+`, `++` and `-`, `->`, `--` overlap.

`-`, `--` overlap with **integer** `-1`

If you write a tokenizer by hand, you have to worry about overlaps. (which means that you loose modularity)

Usage of a scanner generator

Nearly all tokens can be defined by regular expressions. The scanner generator constructs an NDFA, and translates this into an equivalent DFA. The resulting DFA is very efficient (optimal). The DFA reads the input only once. When defining the tokens, the user doesn't need to worry about overlaps.

Disadvantages are that the user has to spend time learning to use the tool, and that the resulting scanner does not give much help when computing the attribute, (DFAs are only good at saying 'yes' or 'no') and that the resulting scanners are not flexible.

Non-flexibility

In general, tokenizers tend to be not as clean as parsers, and sometimes one has to use tricks.

For example in Prolog, it is important whether there is a space between an identifier and a '('.

In some languages, a `.` terminates the input, but inside `( )`, or `[ ]`, it is just a usual operator.

In C^{++}-11, a `>>` can denote the `>>`-operator, or two separate occurrences of `>`, as in `std::list< std::list< int >>`.

Non-Deterministic Finite Automata

Definition: An NDFA is a structure of form $(\Sigma, Q, Q_s, Q_a, \delta)$, in which

- Σ is the alphabet,
- Q is the set of states (finite),
- $Q_s \subseteq Q$ is the set of starting states,
- $Q_a \subseteq Q$ is the set of accepting states,
- $\delta \subseteq Q \times \Sigma^* \times Q$ is the transition relation.

We have been drawing NDFAs in the lecture and in the exercises. In a drawing, states are anonymous. If you want to represent an NDFA in a computer, you need some set Q.

NDFAs (2)

An NDFA accepts a word w iff there exist a finite sequence of words $w_1, \ldots, w_n$, and a sequence of states $q_1, q_2, \ldots, q_{n+1}$, s.t.

- $w = w_1 \cdot \ldots \cdot w_n$,
- $q_1 \in Q_s, \quad Q_{n+1} \in Q_a$,
- Each $(q_i, w_i, q_{i+1}) \in \delta$.

Non-Determinism

It would be nice if one could use use a program of form

```
state = Qs;
nextstate = delta( state, r. lookahead );
while( nextstate != undefined )
{
   r. moveforward( );  // Reads new r. lookahead
   state = nextstate;
   nextstate = delta( state, r. lookahead );
}

// Determine the type of token, based on
// the state in which we got stuck.
```

Non-Determinism (2)

Unfortunately, this is not possible, because (1) δ is not a function, but a relation, and (2) Q_s is not a single state, but a set of states.

In practice, (1) is rarely a problem, but (2) usually is. Problem (2) is caused by the fact that in the beginning one does not know what token will come, so one has to start with the initial states for all of them.

Remarks

NDFAs can be programmed by hand using gotos, or by keeping an explicit state variable or a set of state variables.

Tokenizers are usually greedy. This means that they try to read the longest possible token. Doing something else would be problematic.

Regular Expressions (1)

('Regular' means 'according to rules', which is actually a quite meaningless word.)

Let Σ be an alphabet.

- Every word $s \in \Sigma^*$ is a regular expression.
- If e is a regular expression then e^* is also a regular expression.
- If e_1, e_2 are regular expressions then $e_1 \cdot e_2$ is a regular expression.
- If e_1, e_2 are regular expressions then $e_1 \mid e_2$ is a regular expression.

Regular Expressions (2)

Other constructs can be added as well:

- If e is a regular expressions, $n \geq 0$, then e^n is a regular expression.
- If e is a regular expression, then $e?$ is a regular expression.
- If e is a regular expression, then e^+ is a regular expression.
- If the alphabet Σ is ordered by a total order $<$, and $\sigma_1, \sigma_2 \in \Sigma$ $\sigma_1 \leq \sigma_2$, then $\sigma_1 \cdots \sigma_2$ is a regular expression.

All these constructions are definable, but the definitions can be quite long. For example, $e^{20} = e \cdot \ldots \cdot e$.

$$\sigma_i \cdots \sigma_k = \sigma_i \mid \sigma_{i+1} \mid \sigma_{i+2} \mid \cdots \mid \sigma_{k-1} \mid \sigma_k.$$

(This is only possible if Σ is finite, and not too big.)

Regular Expressions (3)

Examples:

```
digit := '0' .. '9'
letter := 'a' .. 'z' | 'A' .. 'Z'
ident := letter ( letter | digit | '_' ) +
float := ( "" | "+" | "-" )
         digit +
         ( '.' digit + ) ?
         (( 'e' | 'E' ) ( '-' | '+' | "" ) digit + ) ?
```

What do you find more readable? NDFAs or regular expressions?

Regular Expressions (4)

We define recursively when a word w in Σ^* satisfies a regular expression:

- If s is a regular expression consisting of a single word, and $w = s$, then w satisfies s.
- If w is a word, then w satisfies e^* if there exist $n \geq 0$ words $w_1, \ldots, w_n$, s.t. $w = w_1 \cdot w_2 \cdot \ldots \cdot w_n$, and each w_i $(1 \leq i \leq n)$ satisfies e.
- If w is a word, then w satisfies $e_1 \cdot e_2$ if there exist w_1, w_2, s.t. $w = w_1 \cdot w_2$, w_1 satisfies e_1, and w_2 satisfies e_2.
- If w is a word, then w satisfies $e_1 \mid e_2$ if either w satisfies e_1 or w satisfies e_2.

Structure of a Scanner Generator

A scanner generator proceeds as follows:

1. Translate the regular expressions belonging to the tokens into NDFAs.
2. Combine the NDFAs for the tokens into a single NDFA.
3. Translate the NDFA into a DFA.
4. Minimize the DFA.
5. Compress the DFA and generate tables.

Translating a Regular Expression into an NDFA

Translating regular expressions into NDFAs is surprisingly easy.

For a regular expression e, the translation $\mathcal{A}(e) = (\Sigma, Q, Q_s, Q_a, \delta)$ will be defined on the following slides.

Σ will be always the same.

$\mathcal{A}$ is defined by recursion on the structure of e.

Translating a Regular Expression into an NDFA

Assume that e is built from a single word s. The translation $\mathcal{A}(e)$ is the automaton $(\Sigma, \{q_s, q_a\}, \{q_s\}, \{q_a\}, \{(q_s, s, q_a)\})$.

Translating a Regular Expression into an NDFA

Asume that e has form $e = e_1 \cdot e_2$. Let

$$A_1 = \mathcal{A}(a_1) = (\Sigma, Q_1, Q_{1,s}, Q_{1,a}, \delta_1),$$

and let

$$A_2 = \mathcal{A}(a_2) = (\Sigma, Q_2, Q_{2,s}, Q_{2,a}, \delta_2).$$

Assume that Q_1 and Q_2 have no states in common. Otherwise, rename the states in A_1.

$\mathcal{A}(e_1 \cdot e_2) = (\Sigma, Q, Q_s, Q_a, \delta)$ is obtained as follows:

- $Q = Q_1 \cup Q_2$,
- $Q_s = Q_{1,s}, \quad Q_a = Q_{2,a}$,
- $\delta = \delta_1 \cup \delta_2 \cup \{(q, \epsilon, q') \mid q \in Q_{1,a},\ q' \in Q_{2,s}\}$.

Translating a Regular Expression into an NDFA

For a regular expression e of form $e = e_1 \mid e_2$, let

$$A_1 = \mathcal{A}(e_1) = (\Sigma, Q_1, Q_{1,s}, Q_{1,a}, \delta_1),$$

and let

$$A_2 = \mathcal{A}(e_2) = (\Sigma, Q_2, Q_{2,s}, Q_{2,a}, \delta_2).$$

Assume that A_1 and A_2 have no states in common. If they have, then rename the states in A_1. Then $\mathcal{A}(e_1 \mid e_2) = (\Sigma, Q, Q_s, Q_a, \delta)$ is obtained as follows:

- $Q = Q_1 \cup Q_2$,
- $Q_s = Q_{1,s} \cup Q_{2,s}, \quad Q_a = Q_{1,a} \cup Q_{2,a}$.
- $\delta = \delta_1 \cup \delta_2$.

Translating a Regular Expression into an NDFA

For a regular expression e of form $e = e_1^*$, let

$$A_1 = \mathcal{A}(e_1) = (\Sigma, Q_1, Q_{1,s}, Q_{1,a}, \delta_1).$$

Then $\mathcal{A}(e_1^*) = (\Sigma, Q, Q_s, Q_a, \delta)$ is obtained as follows:

- $Q = Q_1 \cup \{\hat{q}\}$
- $Q_s = \{\hat{q}\}, \quad Q_a = \{\hat{q}\},$
- $\delta = \delta_1 \cup \{(\hat{q}, \epsilon, q) \mid q \in Q_{1,s}\} \cup \{(q, \epsilon, \hat{q}) \mid q \in Q_{1,a}\}.$

Translating a Regular Expression into an NDFA

Theorem Let e be regular expression. Then w satisfies e iff $\mathcal{A}(e)$ accepts w.

Deterministic Finite Automata

Definition: An NDFA $\mathcal{A} = (\Sigma, Q, Q_s, Q_a, \delta)$ is called deterministic if

1. Q_s contains at most one element,
2. $(q, s, q') \in \delta \Rightarrow s$ has length 1.
3. $(q, s, q_1), (q, s, q_2) \in \delta \Rightarrow q_1 = q_2$.

In summary, a DFA always knows which transition to make when it sees the next token.

Determinization

In the slides that follow, we present a procedure that transforms an NDFA into an equivalent DFA.

Simplification of δ

In our definition of NDFA, it is allowed to have transitions of form (q, w, q') in δ, where $|w| \geq 2$.

The first step is to eliminate these transitions. Let $\mathcal{A} = (\Sigma, Q, Q_s, Q_a, \delta)$ be an NDFA.

- As long as δ contains a transition (q, w, q') with $|w| \geq 2$, do the following: Write $n = |w|$. Let $q_1, \ldots, q_{n-1}$ a sequence of new states, not in Q. Put

$$Q := Q \cup \{q_1, \ldots, q_{n-1}\},$$

and put

$$\delta := \delta \backslash \{(q, w, q')\} \cup \{(q, w_1, q_1), (q_1, w_2, q_2), \ldots, (q_{n-1}, w_n, q')\}.$$

Outline (1)

If you have a non-deterministic automaton $\mathcal{A} = (\Sigma, Q, Q_s, Q_a, \delta)$, then for every word $w \in \Sigma^*$, there exists a set of reachable states $Q' \subseteq Q$, which is obtained as follows:

A state q is reachable under w if there exists a finite sequence of words $w_1, \ldots, w_n$, and a sequence of states $q_1, q_2, \ldots, q_{n+1}$, s.t.

- $w = w_1 \cdot \ldots \cdot w_n$,
- $q_1 \in Q_s, \quad q_{n+1} = q$,
- Each $(q_i, w_i, q_{i+1}) \in \delta$.

(Intuitively, the state q is reachable under w if the automaton can start in a starting state, eat the word w, and end up in state q)

Outline (2)

The algorithm explores all sets of reachable states $R \subseteq Q$ and constructs the graph of them.

Since there are only finitely many subsets of Q, this exploration will eventually end, and the resulting graph will be a deterministic finite automaton.

Epsilon Closure

Let $S \subseteq Q$ a set of states belonging to an NDFA $\mathcal{A} = (\Sigma, Q, Q_s, Q_a, \delta)$. The ϵ-closure of S is the smallest set S', s.t.

- $S \subseteq S'$,
- If $q \in S'$ and $(q, \epsilon, q') \in \delta$, then $q' \in S'$.

CLOS(S) can be computed as follows:

- $S' := S$,
- As long as there exist $q \in S'$ and $(q, \epsilon, q') \in \delta$, s.t. $q' \notin S'$ do $S' := S' \cup \{q'\}$,
- Now $S' = \text{CLOS}(S)$.

Step Function

Let $S \subseteq Q$ be a set of states belonging to an NDFA $\mathcal{A} = (\Sigma, Q, Q_s, Q_a, \delta)$. Let $\sigma \in \Sigma$. Then $\text{STEP}(S, \sigma)$ is defined as the set

$$\{q' \mid \text{ there is a } q \in S, \text{ s.t. } (q, \sigma, q') \in \delta\}.$$

Let $\mathcal{A} = (\Sigma, Q, Q_s, Q_a, \delta)$ be an NDFA. The determinization of $\mathcal{A}$ is the automaton $\mathcal{A}' = (\Sigma, Q', Q'_s, Q'_a, \delta')$, which is the result of the following algorithm:

- Start with $\mathcal{A}' := (\Sigma,\ \{\mathrm{CLOS}(Q_s)\},\ \mathrm{CLOS}(Q_s), \emptyset, \emptyset)$.
- As long as there exist an $S \in Q'$, and a $\sigma \in \Sigma$, s.t. $S' = \mathrm{CLOS}(\mathrm{STEP}(S, \sigma)) \notin Q'$, put

$$Q' := Q' \cup \{S'\}, \quad \delta' := \delta' \cup \{(S, \sigma, S')\}.$$

- As long as there exist $S, S' \in Q'$, and a $\sigma \in \Sigma$, such that $S' = \mathrm{CLOS}(\mathrm{STEP}(S, \sigma))$ and $(S, \sigma, S') \notin \delta'$, put

$$\delta' := \delta' \cup \{(S, \sigma, S')\}.$$

- At the end, put

$$Q'_a := \{S \in Q' \mid S \cap Q_a \neq \emptyset\}.$$

Minimalization of a DFA

It can happen that the DFA that was obtained by the previous construction, is not minimal. Such a DFA will appear if one determinizes the NDFA resulting from the following regular expression: $(ab|(ab)^*)^*$.

On the following slides we will give a procedure for detecting states with the same observational behaviour. Once these states are found, they can be unified, which results in a minimal automaton.

Definition: Let $(\Sigma, Q, Q_s, Q_a, \delta)$ be a DFA. A state partition Π is a set of sets of states with the following properties:

- For every q in Q, there is an $S \in \Pi$, s.t. $q \in S$.
- For every $q \in Q$, if there are $S_1, S_2 \in \Pi$, s.t. $q \in S_1$ and $q \in S_2$, then $S_1 = S_2$.

So Π separates Q into different groups. Each $q \in Q$ occurs in exacty one group.

The goal is to construct Π in such a way that states that 'behave in the same way' go into the same group as much as possible.

We start with partitioning the states into two groups: Accepting and non-accepting. After that, all groups are inspected for states that behave different. If such states are found, the group is separated into different groups. The procedure stops when no more separations are possible.

A group (of states) must be separated if it contains two states q_1, q_2, for which there exists a symbol $s \in \Sigma$, s.t. $\delta(q_1, s)$ and $\delta(q_2, s)$ are in different groups.

Minimalization Algorithm (Initial Partition)

- The algorithm starts with the partition

$$\Pi := \{Q \backslash Q_a, Q_a\}.$$

If different elements in Q_a accept different tokens, one has to further partition Q_a according to the tokens that are being accepted.

For example if Q_a consists of three states q_1, q_2, q_3, where q_1 accepts **real**, and q_2, q_3 accept **int**, then one has to start with the partition

$$\Pi := \{\ Q \backslash \{q_1, q_2, q_3\}, \{q_1\}, \{q_2, q_3\}\ \}.$$

Minimalization Algorithm (Refining the Partition)

As long as there exist an $S \in \Pi$ and a $\sigma \in \Sigma$, s.t. there are $q_1, q_2 \in S$, and $q_1' \in S'$, $q_2' \in S''$ with $S' \neq S''$, $(q_1, \sigma, q_1') \in \delta$, and $(q_2, \sigma, q_2') \in \delta$, separate S as follows:

- Write $\Pi = \{S_1, \ldots, S_n\}$. For each S_i, construct the set

$$N_i' = \{\ q \in S \mid \exists q' \in S_i \text{ s.t. } (q, \sigma, q') \in \delta\ \}.$$

- Replace Π by

$$(\ \Pi \backslash \{S\}\ ) \cup \{\ N_i' \mid 1 \leq i \leq n \text{ and } N_i' \neq \emptyset\ \}.$$

Continue until no further partitions have to be split.

Minimalization Algorithm (Reading Of the Result)

Let $\mathcal{A} = (\Sigma, Q, Q_s, Q_a, \delta)$ be a DFA. Let Π be the partition constructed by the determinization algorithm. Then the simplified automaton $\mathcal{A}' = (\Sigma, Q', Q'_s, Q'_a, \delta')$ can be constructed as follows:

- $Q' = \Pi$,
- $Q'_s = \{S \in \Pi \mid Q_s \in S\}$,
- $Q'_a = \{S \in \Pi \mid Q_a \in S\}$,
- $\delta' = \{(S, s, S') \mid$ there are $q, q' \in S, S',$ s.t. $(q, s, q') \in \delta\}$.

Pruning the DFA

Let $(\Sigma, Q, Q_s, Q_a, \delta)$ a DFA. If Q contains a state from which there exists no path to an accepting state, then this state can be removed from the DFA.

DFAs obtained from the subset construction contain such a state.

Large Character Types

Originally, (up to 1995?), characters were 8 bits long.

Since there were at most 256 characters, it was possible to implement δ as an array of type `unsigned int [ ] [ 256 ]`.

With modern character sets (unicode), this is not possible anymore.

Also, other representations (for example using `std::map< state, uchar >`) are not practical.

Large Character Types (2)

We assume that:

1. Σ is totally ordered by a relation $<$.
2. Every $\sigma \in \Sigma$ has a next character, with property: $\sigma < \text{next}(\sigma)$ and there is no character σ', s.t. $\sigma < \sigma' < \text{next}(\sigma)$.

(So the set of reals $\mathcal{R}$ wouldn't work.)

A letter σ is in the interval $[\sigma_1; \sigma_2)$ if $\sigma_1 \leq \sigma$ and $\sigma < \sigma_2$.

The main idea is to partition Σ into intervals $[b_1; b_2)$, $[b_2; b_3)$, $\ldots$, $[b_{n-1}; b_n)$, whose elements are not distinguished by the regular expressions. After that, the intervals can be used to construct the NDFA and the DFA in the same way as before. In practical cases, $n \ll |\Sigma|$.

Partitioning the Character Set

We start with a set of regular expressions $e_1, \ldots, e_m$, which possibly contains intervals.

1. Replace all subexpressions of form $w = (w_1, w_2, \ldots, w_m)$ with $m > 1$ by $w_1 \cdot w_2 \cdot \ldots \cdot w_m$.
2. Replace all subexpressions of form a where a is a single letter, by $[\ a; \mathrm{next}(a)\ )$.
3. If there is a subexpression consisting of a closed interval $[\sigma_1; \sigma_2]$, then replace it by $[\ \sigma_1; \mathrm{next}(\sigma_2)\ )$.

At this point, the regular expression is completely built-up from intervals of form $[\sigma_1; \sigma_2)$.

Partitioning the Character Set

Let $b_1, \ldots, b_n$ be all the interval borders occurring in $e_1, \ldots, e_m$, sorted by $<$.

As long as there is a subexpression of form $[\sigma_1; \sigma_2)$, where $\sigma_1 = b_i$ and $\sigma_2 = b_j$ with $j > i + 1$, partition $[\sigma_1; \sigma_2)$ into

$$[\sigma_1; b_{i+1}) \mid [b_{i+1}; b_{i+2}) \mid \cdots \mid [b_{j-1}; \sigma_2).$$

After this, proceed as before, treating the intervals $[b_i, b_{i+1})$ as single letters.

FLEX

The FLEX tool reads a list of regular expressions and associated actions.

It constructs the minimal DFA as described on the previous slides. It constructs C code that can run the DFA.

It can be installed on Ubuntu through `apt-get`.

Usage of FLEX tool (my impression)

- It is really very easy to write a complex scanner with FLEX.
- FLEX gives no support in the computation of the attributes. One still has to use **atoi**, **atof**. This means that one still uses an NDFA that somebody wrote by hand. (But you have seen in the exercises that the main problem is in the combination of different tokens. At least this problem was solved by FLEX)
- The C^{++} interface is not good. C^{++} is more than a few plusses on C. Flex runs under C^{++}, because $C \subseteq C^{++}$, but I think that this is not enough.

In my experience, the question whether a scanner generator should be used is always borderline.

I made some implementations of logic systems (theorem provers), and I never used a scanner generator tool. I do use parser generators.

There are always additional conditions that interfere with the scanner. (Like the >> in C^{++}-11, or the no-space condition in Prolog.)

C^{++} also has a problem with typenames, because declarations can be mixed with statements.

Summary

The tokenizer must:

Deal with the input source.

Group the input into tokens, and compute the attributes.

Possibly look up identifiers.

take context into account, (unfortunately often) because not all tokens can occur in all contexts.

attach information about the origin to the token. (File, line number, position) This information must be preserved through the compilation process, until is certain that no errors will occur. (GCC does not follow this rule, because the linker can still generate errors.)