Formal Languages

Wednesday, September 19, 2007
Reading: Stoughton 2.1, 2.2, end of 2.3, beginning of 3.1

CS235 Languages and Automata

Department of Computer Science
Wellesley College

Alphabets, Strings, and Languages

An **alphabet** is a set of symbols.
E.g.: $\Sigma_1 = \{0,1\}$; $\Sigma_2 = \{-,0,+\}$ $\Sigma_3 = \{a,b, \ldots, y, z\}$; $\Sigma_4 = \{☺, \Rightarrow, a, aa\}$

A **string over Σ** is a sequence of symbols from Σ.
The empty string is traditionally written ε, but Stoughton uses %.
Σ^* denotes all strings over Σ. E.g.:

- Σ_1^* contains %, 0, 1, 00, 01, 10, 11, 000, …
- Σ_2^* contains %, -, 0, +, --, -0, -+, 0-, 00, 0+, +-, +0, ++, ---, …
- Σ_3^* contains %, a, b, …, aa, ab, …, bar, baz, foo, wellesley, …
- Σ_4^* contains %, ☺, ⇒, a, aa, …, a ⇒ ☺, …, a aa, aa a,…

A **language over Σ** (Stoughton's Σ-language) is any subset of Σ*.
I.e., it's a (possibly countably infinite) set of strings over Σ. E.g.:

- L_1 over Σ_1 is all sequences of 1s and all sequences of 10s.
- L_2 over Σ_2 is all strings with equal numbers of -, 0, and +.
- L_3 over Σ_3 is all lowercase words in the OED.
- L_4 over Σ_4 is {☺, ☺ ⇒ ☺, a aa}.

String Operations

Length: $|s|$ is the length of a string s. E.g.:
$|\%| = 0$, $|\text{foo}| = 3$, $|\text{☺} \Rightarrow \boxed{a}\boxed{aa}| = 4$

Concatenation: If x, y in Σ^*, then $x@y$ in Σ^* is the string consisting of all symbols in x followed by all symbols in y. (Concatenation is traditionally written xy, and Stoughton uses this as an abbreviation.)
E.g. baz@quux = bazquux

Concatenation Properties:

- $(x@y)@z = xyz = x@(y@z)$ (*Associativity*)
- $x@\% = x = \%@x$ (*Identity*)
- $|x@y| = |x| + |y|$

Monoid

Other Definitions:

- x is a **prefix** of y iff $y = xv$ for some v
- x is a **suffix** of y iff $y = ux$ for some u
- x is a **substring** of y iff $y = uxv$ for some u and v

There are **proper** versions of these, too.

What are all prefixes, suffixes, substrings of bar?

More String Operations

String Powers: Suppose x is a string.

- $x^0 = \%$
- $x^n = x@x^{n-1}$, abbreviated $xx^{n-1} = x(x^{n-1})$

Power Properties:

- $x^{a+b} = x^a x^b$
- $|x^n| = n \cdot |x|$

String Reversal: Suppose a is a symbol and x is a string.

- $\%^R = \%$
- $(a@x)^R = x^R@a$

Reversal Properties:

- $(x@y)^R = y^R@x^R$
- $(x^R)^R = x$
- $|x^R| = |x|$

Alphabet: Suppose a is a symbol and x is a string.

- alphabet(%) = { }
- alphabet(ax) = $\{a\} \cup$ alphabet(x)

What are some alphabet properties?

String Induction

Suppose P(w) is a property of strings w in Σ^*. Can prove P(w) by *natural induction (or strong induction) on |w|* . Equivalently:

Right String Induction:

Suppose that

the inductive hypthesis (IH)

1. (basis step) P(%) holds.
2. (inductive step) For all $a \in \Sigma$ and x in Σ^*, $P(x) \Rightarrow P(ax)$.

Then P(w) holds for all $w \in \Sigma^*$.

Left String Induction:

(inductive step) For all $a \in \Sigma$ and x in Σ^*, $P(x) \Rightarrow P(xa)$.

Strong String Induction:

(inductive step) For all $w \in \Sigma^*$, $(\bigwedge_{x \in \Sigma^* \text{ s.t. } |x| < |w|} P(x)) \Rightarrow P(w)$.

String Induction Example: Reversal

Prove that $(x@y)^R = y^R@x^R$

Which form of induction to use?

(basis step)

(inductive step)

What is I.H.?

Language Operations

Suppose L_1 and L_2 are Σ-languages.

Set Operations: The following are all Σ-languages:

$L_1 \cup L_2$, $L_1 \cap L_2$, $L_1 - L_2$, $\overline{L_1}$ $(= \Sigma^* - L_1)$

Language Concatenation:

$L_1 @ L_2 = \{x @ y \mid x \text{ in } L_1 \text{ and } y \text{ in } L_2\}$ (abbreviated L_1L_2)

E.g. {CS, PHYS} @ {110, 111, 115} =

Concatenation Properties:

- $(L_1 @ L_2) @ L_3 = L_1 @ (L_2 @ L_3)$ (Associativity)
- $\{\%\} @ L = L = L @ \{\%\}$ (Identity)
- $\varnothing @ L = \varnothing = L @ \varnothing$ (Zero)
- $|L_1 @ L_2| = |L_1| \cdot |L_2|$ for finite L_1, L_2

More Language Operations

Language Powers:

- $L^0 = \{\%\}$
- $L^n = L @ L^{n-1}$

E.g., $\{0,1\}^2$ =

Power Properties:

- $L^{a+b} = L^a @ L^b$
- $|L^n| = |L|^n$ for finite L
- $\{x\}^n = \{x^n\}$
- $\{\%\}^n = \{\%\}$

Kleene Closure (Kleene Star): (pronounced "clay knee")

$L^* = \{L^n \mid n \text{ in Nat}\}$

E.g. {0,1}* =

(This is consistent with notation Σ^*.)

Stoughton's Conventions

He fixes the set of symbols to be **Sym**, which contains:

- digits 0, 1, …, 9
- upper case letters A, B, …, Z
- lower case letters a, b, …, z
- < and > delimiting any sequence of ASCII characters in which < and > are properly nested. E.g., <plus>, <<a>, <b, c>>

Symbols are ordered first by length, and then lexicographically:

0 < … < 9 < A < … < Z < a … < z < <0> < … <z> < <00> … < <zz> < …

He fixes the universe of strings to be **Str** = all strings over **Sym**.
He uses % for the empty string.

Strings are ordered first by length, and then lexicographically. E.g.:

% < a < <ab> < ab < ac < a<bbb> < <bb><aa> < <bbb>a < aa<bb> < <bb>aa

Can't write (a + b)*c. Instead, write <LPAREN>a<PLUS>b<RPAREN><TIMES>c