Linguistics 696f
Hammond
Spring '03

Formal language theory

A. Overview

(1)
a. What is Statistical NLP?
b. Why do this?
c. Finite state machines
d. Context-free languages
e. Chomsky-Normal Form
f. Linear/Regular Grammars

B. Why do this?

(2) Which book did you read t without buying t?

(3)
$H_0 = \{a, aa, aaa, \ldots\}$
$H_1 = \{a\}$
$H_2 = \{a, aa\}$
$H_3 = \{a, aa, aaa\}$
$H_i = \{a, aa, \ldots, a_i\}$

C. Formal language

(4) A formal language is a possibly infinite set of words constructed from some finite alphabet.

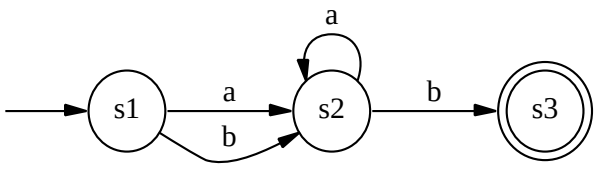
(5) A regular language is a language that can be described using only three operations: *union*, *concatenation*, and *Kleene star*, e.g. $a(b|c)d*$.

(6) $a_n b_n$: *ab*, *aabb*, *aaabbb*, etc. is not regular.

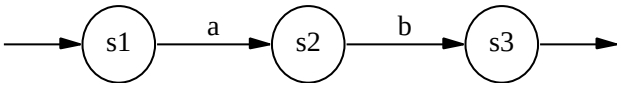
(7) $(\mathrm{CV}|\mathrm{CVC})(\mathrm{CV}|\mathrm{CVC})^*$

D. Finite State Machines

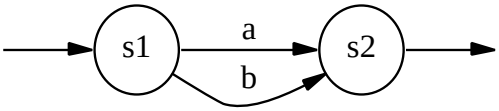
(8) A simple FSA



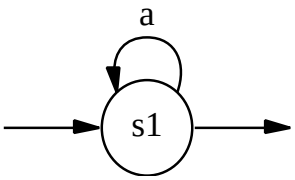
(9) Concatenation



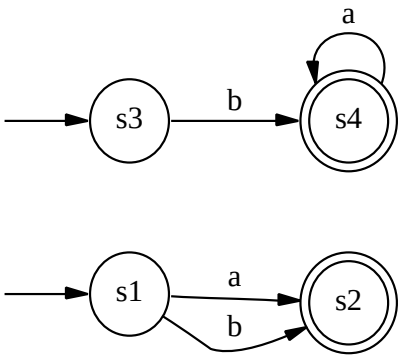
(10) Union



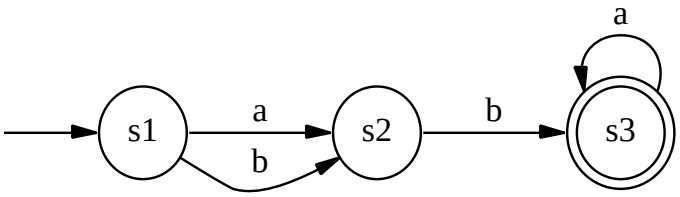
(11) Kleene star



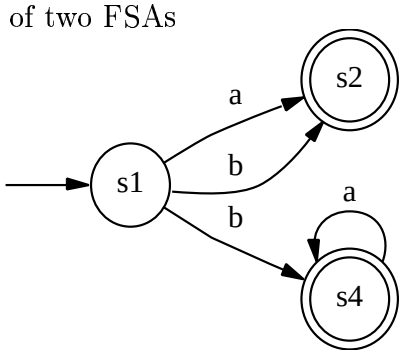
(12) Two FSAs



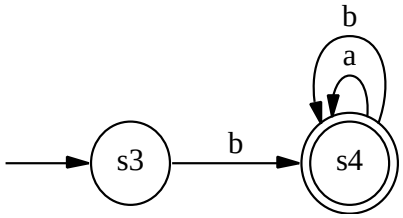
(13) Concatenated FSAs



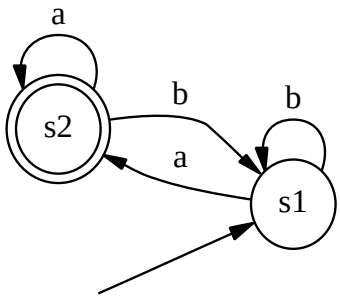
(14) union of two FSAs



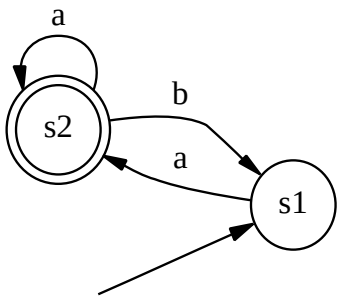
(15) Kleene star



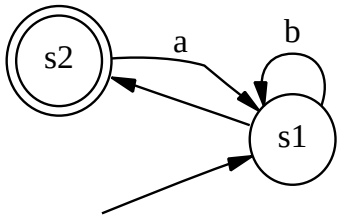
(16) Deterministic FSA



(17) Nondeterministic FSA



(18) FSA with null transition



E. Re-write rules

(19) Context-free languages

terminals:	a, b
non-terminals:	A, B
starting node:	A
production rules:	$A \rightarrow a\ B\ b$
	$B \rightarrow a\ b$
	$B \rightarrow \emptyset$

(20) Chomsky-Normal Form

a. $A \rightarrow BC$, where A, B, and C are non-terminal symbols.

b. $A \rightarrow a$, where A is a non-terminal and a is a single terminal.

(21) Linear grammars

a. $A \rightarrow a_1 \ldots a_n$, where a is a terminal element.

b. $A \rightarrow a_1 \ldots a_n B$, where B is a single non-terminal element.

References

CHARNIAK, EUGENE. 1993. *Statistical Language Learning*. Cambridge: MIT Press.

HOPCROFT, J.E., & J.D. ULLMAN. 1979. *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley.