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) \( a_nb_n \): \( ab, aabb, aaabbb \), etc. is not regular.

(7) \( (CV|CVC)(CV|CVC)^* \)

D. Finite State Machines

(8) A simple FSA

![Diagram of a simple FSA with states S1, S2, and S3 and transitions labeled 'a' and 'b'.]
(9) Concatenation

\[ \begin{array}{c}
  s1 \xrightarrow{a} s2 \xrightarrow{b} s3 \\
\end{array} \]

(10) Union

\[ \begin{array}{c}
  s1 \xrightarrow{a} s2 \\
  s1 \xrightarrow{b} s2 \\
\end{array} \]

(11) Kleene star

\[ \begin{array}{c}
  s1 \xrightarrow{a} s1 \\
\end{array} \]

(12) Two FSAs

\[ \begin{array}{c}
  s3 \xrightarrow{b} s4 \\
  s1 \xrightarrow{a} s2 \\
\end{array} \]

(13) Concatenated FSAs

\[ \begin{array}{c}
  s1 \xrightarrow{a} s2 \xrightarrow{b} s3 \\
  s1 \xrightarrow{b} s2 \\
\end{array} \]
(14) union of two FSAs

(15) Kleene star

(16) Deterministic FSA

(17) Nondeterministic FSA
(18) FSA with null transition

![Diagram](attachment:image.png)

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
