Computational Linguistics

Lecture 4

Dr. Dina Khattab

dina.khattab@cis.asu.edu.eg

REGULAR EXPRESSION (REGEX)

Regular Expressions

- Easy way to generate a language that is accepted by FSA
- Rules:
  - ε is a regular expression
  - Any symbol in Σ is a regular expression

  If *r* and *s* are any regular expressions then so is:
  - *r* | *s* denotes union e.g. "*r* or *s*"
  - *rs* denotes *r* followed by *s (concatination)*
  - (*r*)* denotes concatination of r with itself zero or more times (Kleene closure)
  - (*r*)+ denotes concatination of r with itself one or more times (positive closure)
  - () used for controlling order of operations

Example Regular Expressions

Regular Expression	Corresponding Language
ε	ε
a	a
abc	abc
a|b|c	a, b, c
ab*	a, ab, abb, abbb, ….
ab+	ab, abb, abbb, …
(ab)*	ε, ab, abab, ababab, ….
(a|b|c)*	ε, a, b, c, aa, ab, ac, aaa, …
a|b*	ε, a, b, bb, bbb, ….
a|b|c|…|z|A|B|…|Z	Any letter
0|1|2|…|9	Any digit

Precedence in Regular Expressions

- * has highest precedence, left associative
- Concatenation has second highest precedence, left associative
- | has lowest associative, left associative

More Regular Expression Examples

Regular Expression	Corresponding Language
ε|a|b|ab*	ε, a, b, ab, abb, abbb, …
ab*c	ac, abc, abbc,…
ab*|a*	ε, a, ab, aa, aaa, abb,…
a(b*|a*)	a, ab, aa, abb, aaa, …
a(b|a)*	a, ab, aa, aaa, aab, aba,…

ε because a*

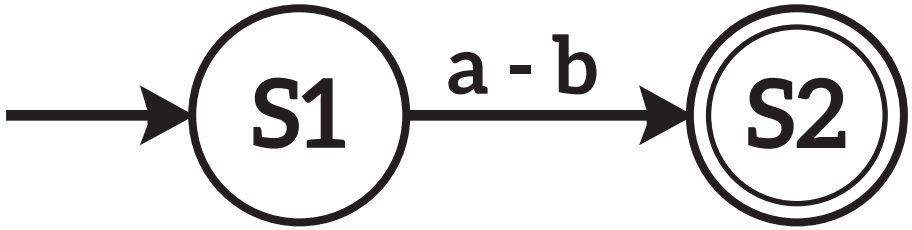
Examples

- a*b*
- a+b*a
- bac*b+
- cba*b*
- c(ba)+c
- ab*(c|ε)

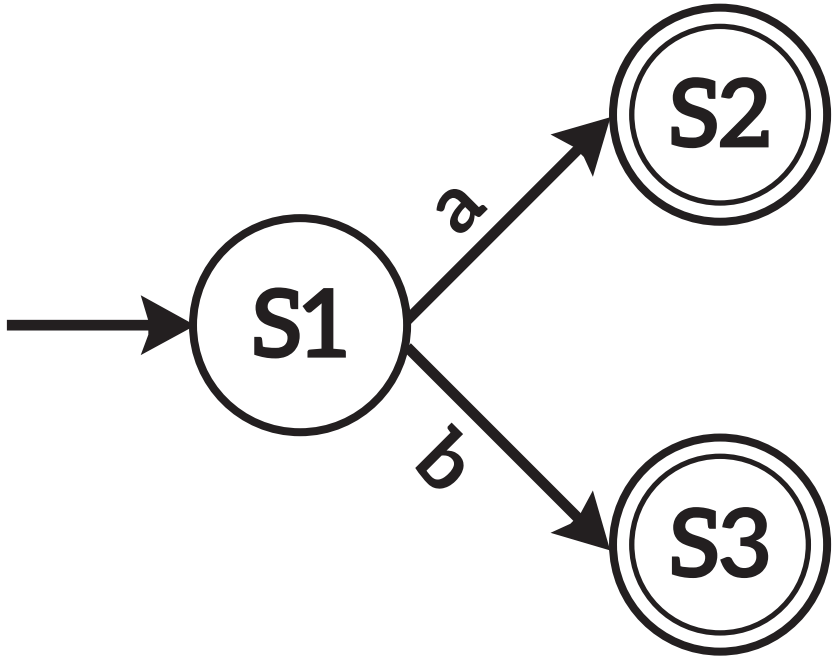
Draw a graph for FSA that represents the following regular expressions:

- a | b
- b *
- $(ab)^+$
- ab*c
- a(b*|a*)
- (a | b)*
- a+b*a

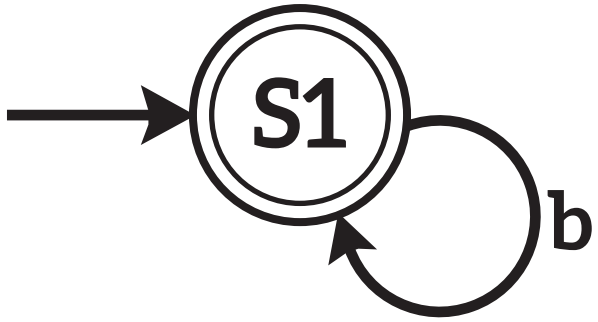
a|b = a , b



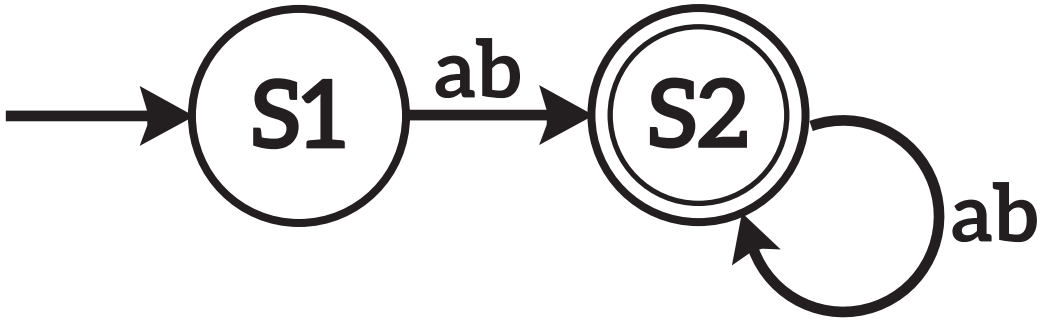
ANOTHER SOLUTION



$b^* = \varepsilon , b , bb , bbb , ...$

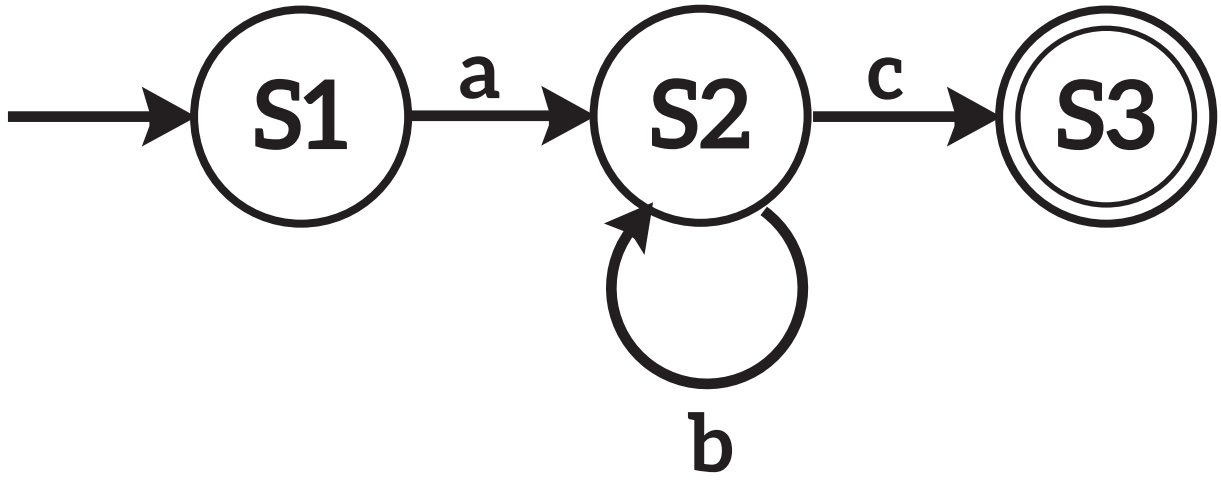


$(ab)^+$ = ab , abab , ababab , ...

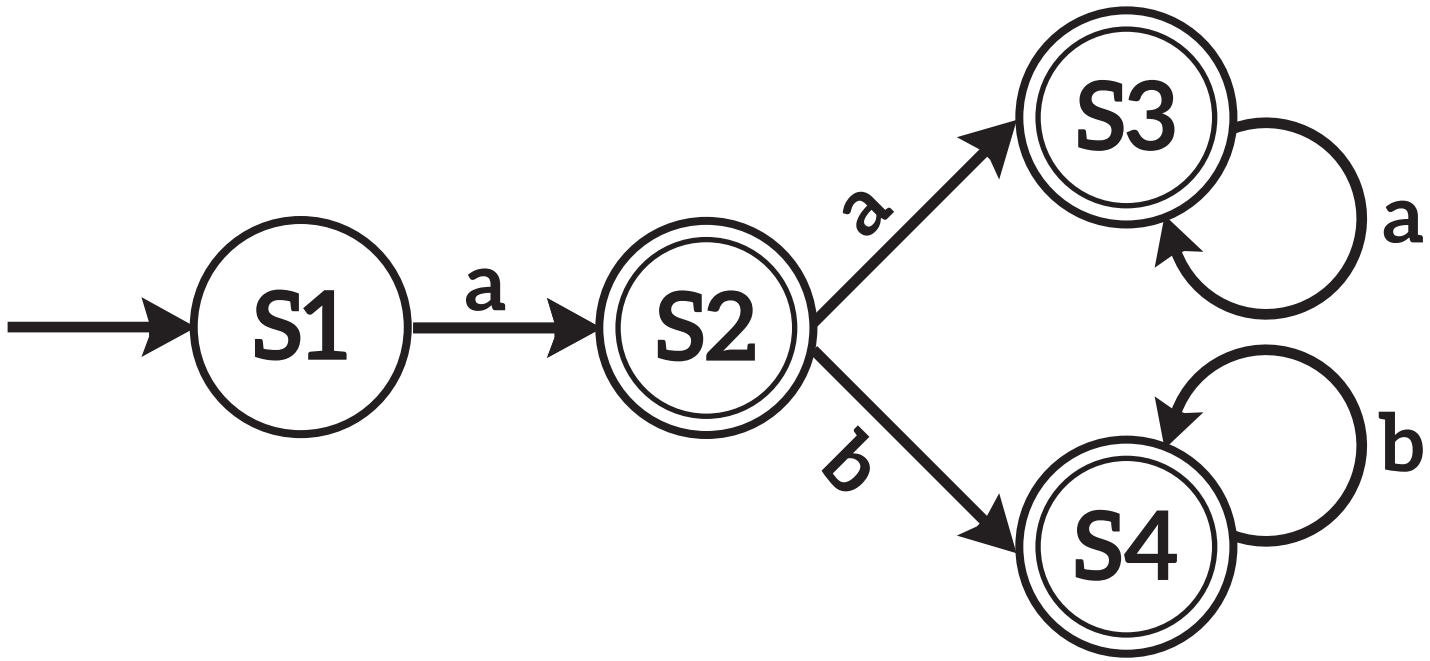


-NOTE-

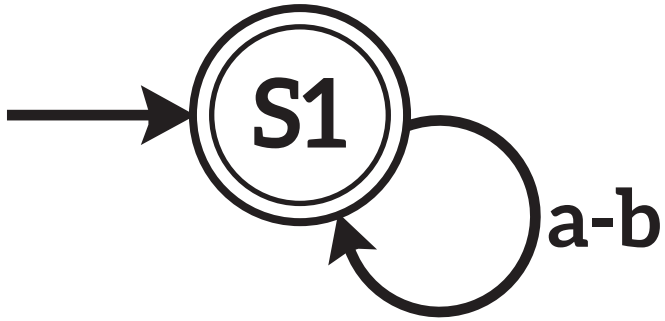
ab*c = ac , abc , abbc , abbbc , ...



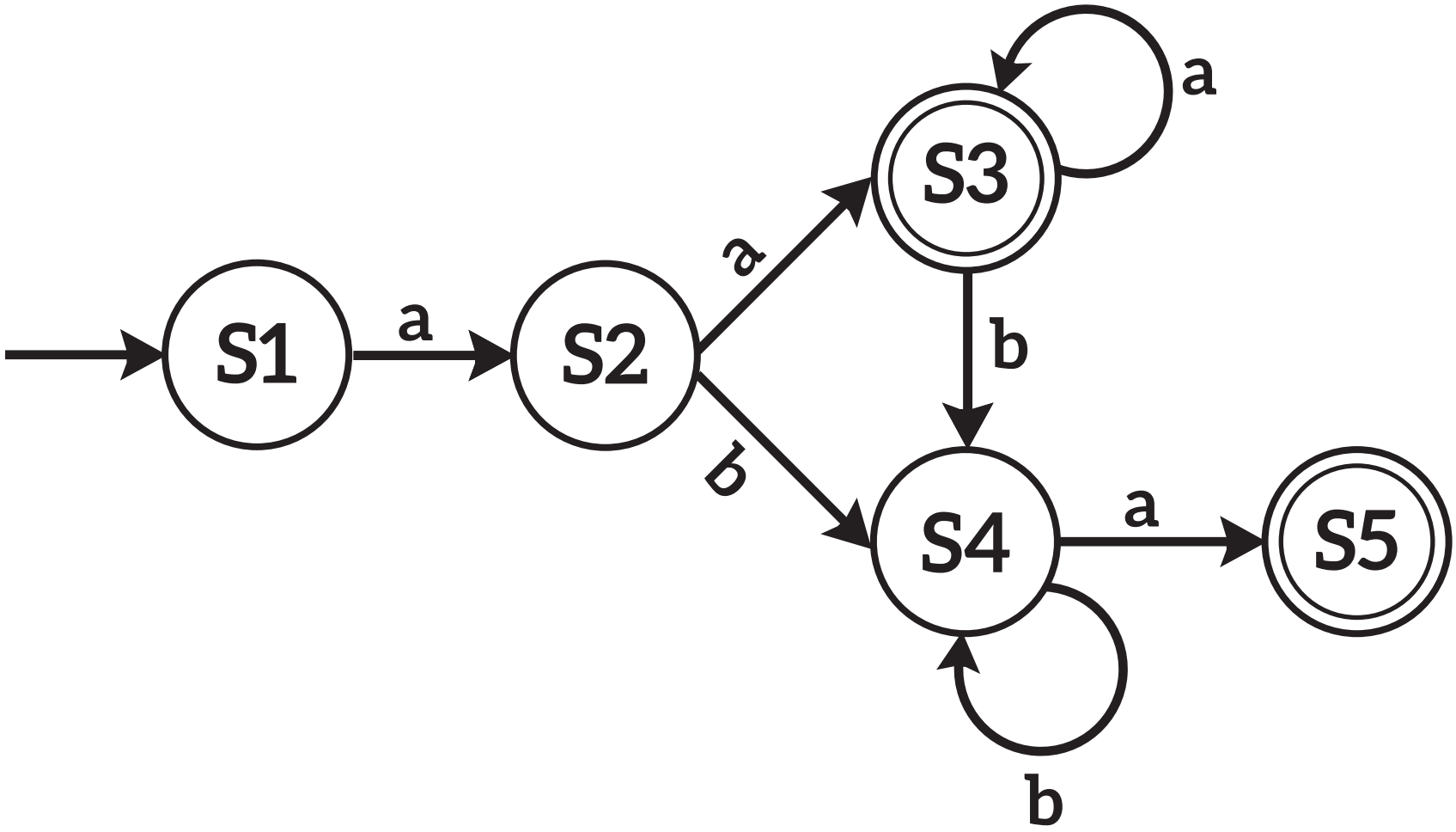
a(b*|a*) = a , ab , abb , … , aa , aaa , …



(a|b)* = ε , a , b , aa , ab , ba , bb , aaa , abb , aba , aab , bbb , baa , bab , baa , ...



a^+b^*a = aa , aaa , aaaa , ... ,
aba , abba , aaba , aabba , ...



-NOTE-

String must start and end with an "a"