Mina Lee, Sunbeom So and Hakjoo Oh
Korea University

Automatic Synthesis of Regular Expressions from Examples

1. Motivation

To help the students who learn REs

Strings have exactly one pair of consecutive 0s. (Σ={0,1})

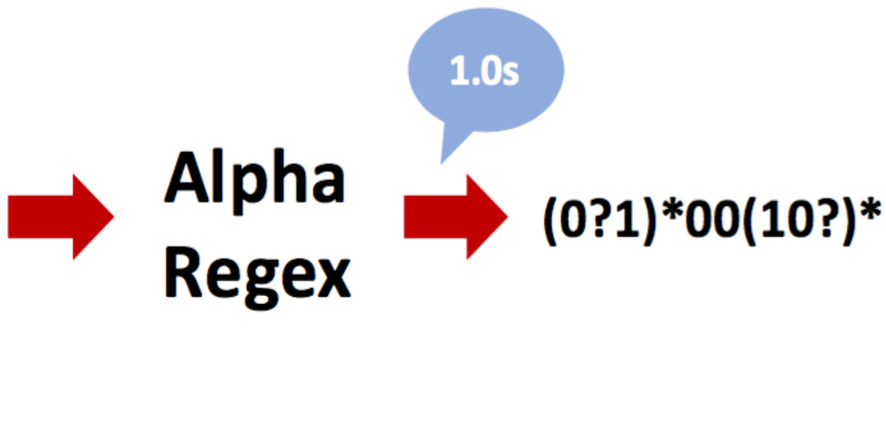
Positive	Negative
00	01
1001	11
010010	000
1111001111	00100

2. Synthesis Goal

Find a regex consistent with positive and negative examples

Positive Example
00
1001
010010
1011001110

Negative Example
01
11
000
00100

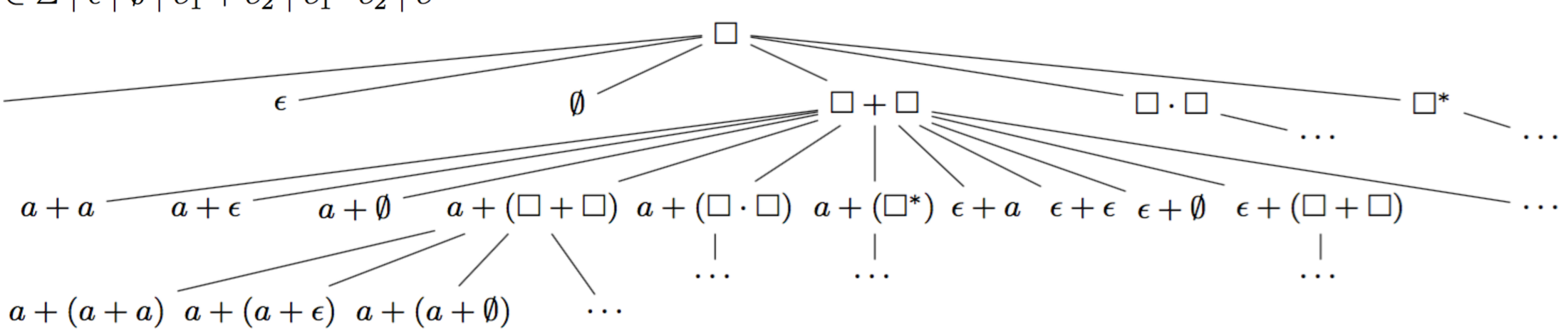


3. Algorithm

Basic Algorithm Exhaustive search in increasing size

Challenge $O(7^{2^d-1})$: Maximum number of states at depth d

$e \to a \in \Sigma \mid \epsilon \mid \emptyset \mid e_1 + e_2 \mid e_1 \cdot e_2 \mid e^*$



Pruning techniques

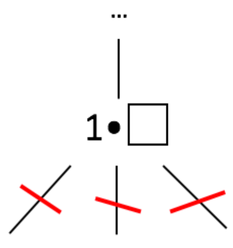
1. Normalization

$$[\![s^* s^*]\!] = [\![s^*]\!]$$
$$[\![(s + s)]\!] = [\![s]\!]$$
$$[\![(s \cdot s^*)^*]\!] = [\![s^*]\!]$$

Do not re-visit semantically-equivalent states.

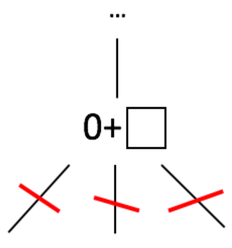
2. Over approximations

0 is in the positive examples



3. Under approximations

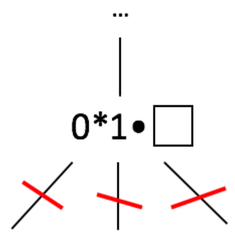
0 is in the negative examples



Further search does not produce any solutions.

4. Identify redundant state

1 is a prefix of positive examples



Redundant symbols exist.

Completeness **If there exists a solution, the solution will eventually be found.**

4. Evaluation

- ✓ 25 problems from textbooks
- ✓ 3 levels of difficulty
  - Easy, Normal, Hard
- ✓ 6.7 seconds on average
- ✓ Worst case : < 1min

Problem	Examples		Output	Time (sec)		
	P	N		Full	No Apr	No Rd
The number of 0s is divisible by 3.	8	7	(1+01*01*0)*	**9.5**	238.4	36.9
The 5th symbol from the right end is 1.	3	3	(0+1)*1(0+1)(0+1)(0+1)(0+1)	**9.0**	56.3	14.3
0 and 1 alternate.	9	8	1?(01)*0?	**1.7**	19.7	3.5
Have at most two 0s.	8	7	1*0?1*0?1*	**1.4**	9.5	2.3
Start with 0 and have odd length or start with 1 and have even length.	5	5	(0+1(0+1))((0+1)(0+1))*	**16.5**	771.9	14.7
Contain at least one 0 and at most one 1.	6	9	0*(01?+100*)	**2.7**	33.5	7.7
At most one pair of consecutive 1s.	5	5	(1?0)*1?1?(01?)*	**28.8**	863.6	211.2

Conducted on an Ubuntu machine with Intel Xeon CPU E5-2630 (2.40GHz).

5. Conclusion

- ✓ Simple, but efficient algorithms
- ✓ Quickly synthesizes the solutions from small examples
- ✓ Demo : http://prl.korea.ac.kr/AlphaRegex