Content and Context: Two-pronged Bootstrapped Learning for Regex-formatted Entity Extraction

Stanley Simoes[†], Deepak P[♯], Munu Sairamesh[†], Deepak Khemani[†], Sameep Mehta[♮]

[†]Indian Institute of Technology Madras [♯]Queen's University Belfast [♮]IBM Research - India

1. Regexes for Entity Extraction

- *Regular Expressions* (*regexes*) can be used to elegantly characterize entities having an underlying syntactical pattern
- Easy to develop, interpret, maintain, and incorporate domain knowledge

ENTITY	REGEX	MATCHED INSTANCE
CS course	`CS\d{3}`	*CS376*
Intel CPU	`i\d-\d{3}`	*i5-750*

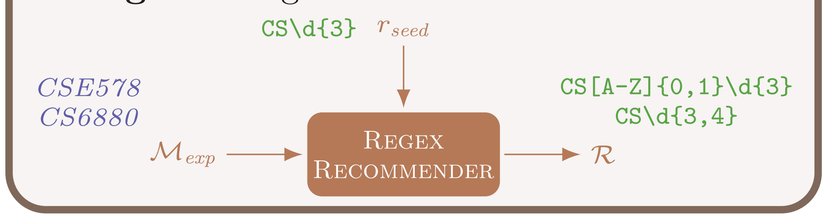
- High quality regexes have: ▷ **high precision:** all matches are instances
  ▷ **high recall:** all instances are matches
- Difficult to manually design a high quality regex for entity extraction
  ▷ A human would generalize some instances to a high precision regex, but would miss out on some variants
  ▷ eg: `CS\d{3}` does not match CS courses *CS6880* and *CSE578* (high precision, low recall)
- Need a system that automatically learns and suggests: a) overlooked instances
  b) high quality regexes
- Use cases: ▷ design a regex for a new entity extraction system
  ▷ enhance regex in an existing entity extraction system

2. Problem Statement

- **Given:**
  ▷ a high precision seed regex r_{seed}
  ▷ a document corpus $\mathcal{D}$
  i.e., a set of seed matches $\mathcal{M}_{seed}$
- **Goal**: To identify
  a) $\mathcal{M}_{exp}$: instances in $\mathcal{D}$ *not* matched by r_{seed}
  b) $\mathcal{R}$: set of diverse, high quality regexes
- **Proposed Approach:** (Two stages)

Stage I: Match Set Expansion

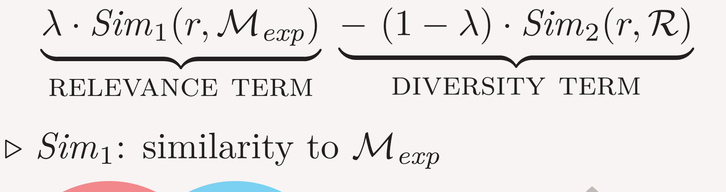
Stage II: Regex Recommendation



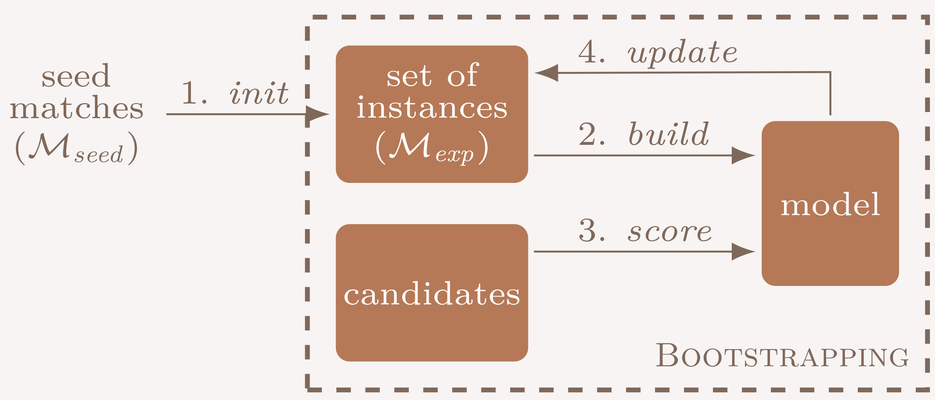
3. High Quality Regexes

- Search the *generalization space* of r_{seed}
- Regex units: character class & quantifier
  ▷ eg: `[A-Z]{2}\d{3,4}`
- Generalizing a regex:
  a) allow more characters in character class
  b) decrease quantifier's lower bound
  c) increase quantifier's upper bound

▷ eg:



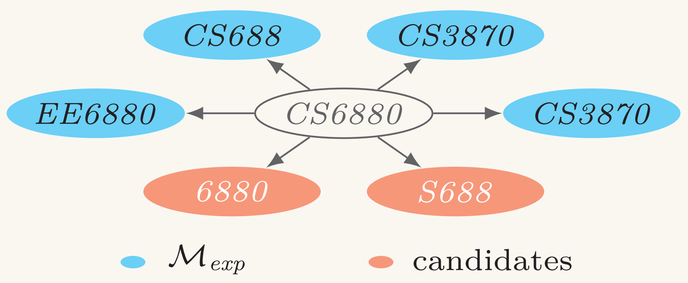
4. Match Set Expansion



- Expand using **Bootstrapping**
  ▷ repeat $\{build \rightarrow score \rightarrow update\}$
- **candidates:**
  ▷ matches of r_{seed}'s generalizations
- **model: Logistic Regression**
  ▷ $\mathcal{M}_{exp}$ as 1
  ▷ candidates as 0
- Matches represented using: ▷ content
  ▷ context

Content Features

- Matches in $\mathcal{M}_{exp}$ within k edit distance



$$Content\text{-}Score_k(CS6880) = \frac{4}{6} = 0.67$$

- **F1:** $Content\text{-}Score_2$ (2 edit distance)
  F2: $Content\text{-}Score_3$ (3 edit distance)

Context Features

- Similarity to contexts of matches in $\mathcal{M}_{exp}$

- Language model $L_{\mathcal{P}}$: $\mathcal{M}_{exp}$'s left contexts

$$Context\text{-}Score_L(CS6680) = \sum_i \ln L_{\mathcal{P}}(w_i)$$

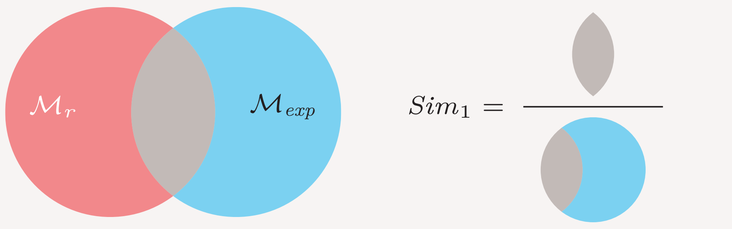
- **F3:** $Context\text{-}Score_L$ (left contexts)
  F4: $Context\text{-}Score_R$ (right contexts)

5. Regex Recommendation

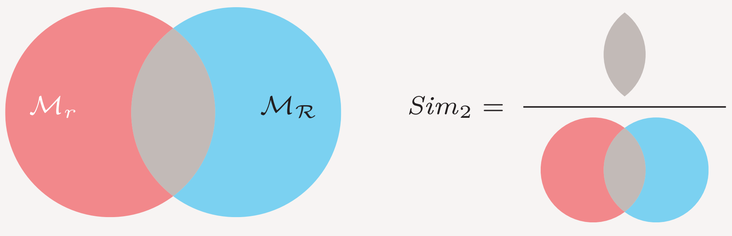
- Generate a set of regexes $\mathcal{R}$ with
  ▷ high recall:
  $\forall r \in \mathcal{R}, r$ is a generalization of r_{seed}
  ▷ high precision:
  $\forall r \in \mathcal{R}$, r's matches are in $\mathcal{M}_{exp}$
  ▷ high diversity:
  $\forall r_1, r_2 \in \mathcal{R}$, r_1's matches are not in r_2's
- Using **Maximal Marginal Relevance**
  ▷ add generalization r to $\mathcal{R}$ that maximizes

$$\underbrace{\lambda \cdot Sim_1(r, \mathcal{M}_{exp})}_{\text{RELEVANCE TERM}} \underbrace{- (1-\lambda) \cdot Sim_2(r, \mathcal{R})}_{\text{DIVERSITY TERM}}$$

▷ Sim_1: similarity to $\mathcal{M}_{exp}$



▷ Sim_2: similarity to $\mathcal{R}$



*$\mathcal{M}_r$ = r's matches in $\mathcal{D}$
$\mathcal{M}_{\mathcal{R}}$ = matches in $\mathcal{D}$ for all $r \in \mathcal{R}$

✉ stanleysimoes@gmail.com

6. Experiments

- Two extraction tasks: ▷ DateWebKB – `\d{2}/\d{2}/\d{2}`
  ▷ CourseWebKB – `CS\d{3}`
- Parameters: ▷ ≤ 4 regex units generalized ▷ 150 iterations
  ▷ 1% expansion per iteration ▷ $\lambda = 0.7$

Comparison of our Match Set Expander's Accuracy with Baselines

TASK	PRECISION				RECALL				F-SCORE				
	FREQ	GM_{10}	$GM_{1\%}$	OURS	$\mathcal{M}_{seed}$	GM_{10}	$GM_{1\%}$	OURS	$\mathcal{M}_{seed}$	FREQ	GM_{10}	$GM_{1\%}$	OURS
DateWebKB	0.032	0.186	0.479	**1.000**	0.251	0.436	0.330	**0.857**	0.402	0.063	0.261	0.391	**0.923**
CourseWebKB	0.070	0.672	0.633	**0.994**	0.342	0.348	0.349	**0.855**	0.509	0.131	0.459	0.450	**0.919**

▷ PRECISION of $\mathcal{M}_{seed} = 1$ ▷ RECALL of FREQ = 1 *best values **boldfaced**

Recommended Regexes

TASK	DateWebKB		CourseWebKB	
r_{seed}	`\d{2}/\d{2}/\d{2}`	eg: *12/13/01*	`CS\d{3}`	eg: *CS376*
$\mathcal{R}$	`\d{1,2}/\d{1,2}/\d{2}`	eg: *2/3/01*	`C?[a-zA-Z]{1,2}\d{3}`	eg: *CSE578*
	`\d{2}/\d{1,2}/\d{2}`	eg: *12/3/01*	`[a-zA-Z]{1,2}S?\d{3}`	eg: *cs615*
	`\d{1,2}/\d{2}/\d{2}`	eg: *2/13/01*	`CS\w{1,3}\d\w`	eg: *CS382M*

*generalized units in **red and underlined**