Human-in-the-Loop Entity Extraction

Lihong He

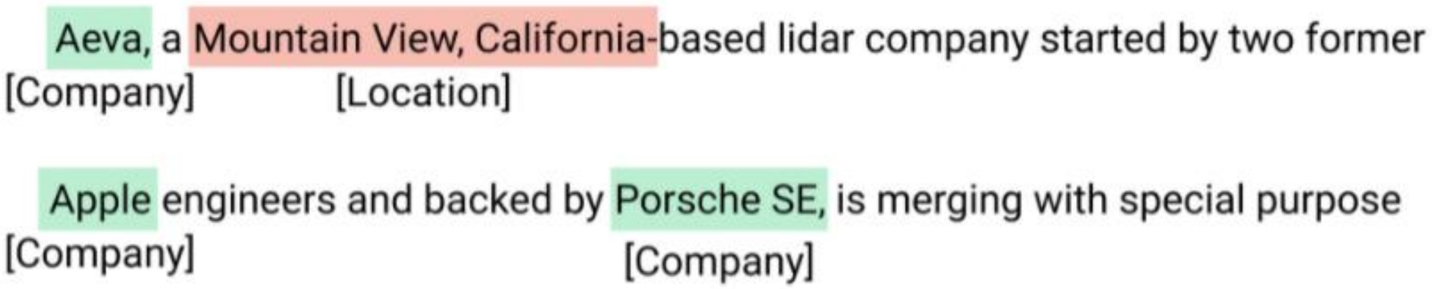
May, 2021

Temple University

Computer and Information Science

Entity Extraction

- Named Entity Extraction

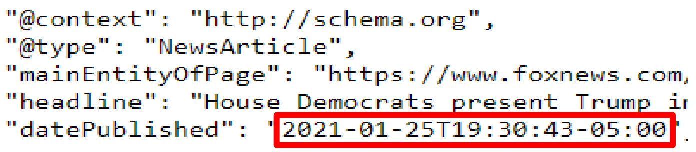


Aeva, a Mountain View, California-based lidar company started by two former
[Company] [Location]

Apple engineers and backed by Porsche SE, is merging with special purpose
[Company] [Company]

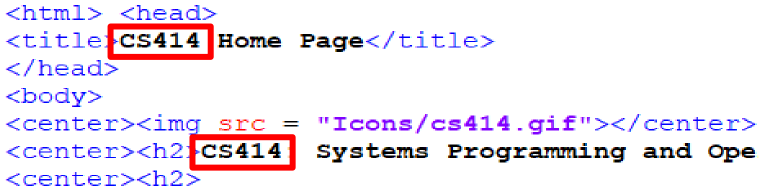
- Entity not “named”

(1) Date Time



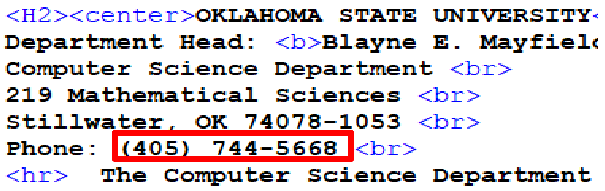
```
"@context": "http://schema.org",
"@type": "NewsArticle",
"mainEntityOfPage": "https://www.foxnews.com,
"headline": "House Democrats present Trump ir
"datePublished": "2021-01-25T19:30:43-05:00".
```

(2) Course Number



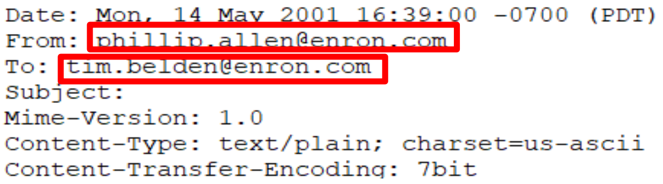
```
<html> <head>
<title>CS414 Home Page</title>
</head>
<body>
<center><img src = "Icons/cs414.gif"></center>
<center><h2>CS414 Systems Programming and Ope
<center><h2>
```

(3) Phone Number



```
<H2><center>OKLAHOMA STATE UNIVERSITY<
Department Head: <b>Blayne E. Mayfiel
Computer Science Department <br>
219 Mathematical Sciences <br>
Stillwater, OK 74078-1053 <br>
Phone: (405) 744-5668 <br>
<hr>  The Computer Science Department
```

(4) Email Address



```
Date: Mon, 14 May 2001 16:39:00 -0700 (PDT)
From: phillip.allen@enron.com
To: tim.belden@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
```

Previous Solutions

- Rule-based matching: Regular Expression (RE)
  - Pre-defined
    - RE1: \d{4,4}-\d{2,2}-\d{2,2}T\d{2,2}:\d{2,2}:\d{2,2}Z
    - RE2: \d{8,8} \d{2,2}:\d{2,2}:\d{2,2}Z
    - RE3: \d{14,14}
    - ……

2021-01-27T06:37:36Z

20210127 06:37:36

20210127063736

Jan 27 06:37:36 2021

06:37 Jan 27, 2021

……

Cannot cover all possible formats!

Previous Solutions

- Deep Learning

Data & Labels

“datePublished”: “2021-01-27T06:37:36Z”

<meta time=“20210127 06:37:36” />

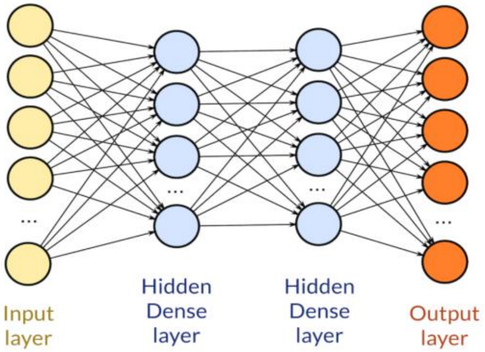
<Article timestamp=“20210127063736”

<span> Jan 27 06:37:36 2021 </span>

Update at 06:37 Jan 27, 2021 by Andrew

……

Deep Model



Test String

Article published at 01/27/2021 20:54

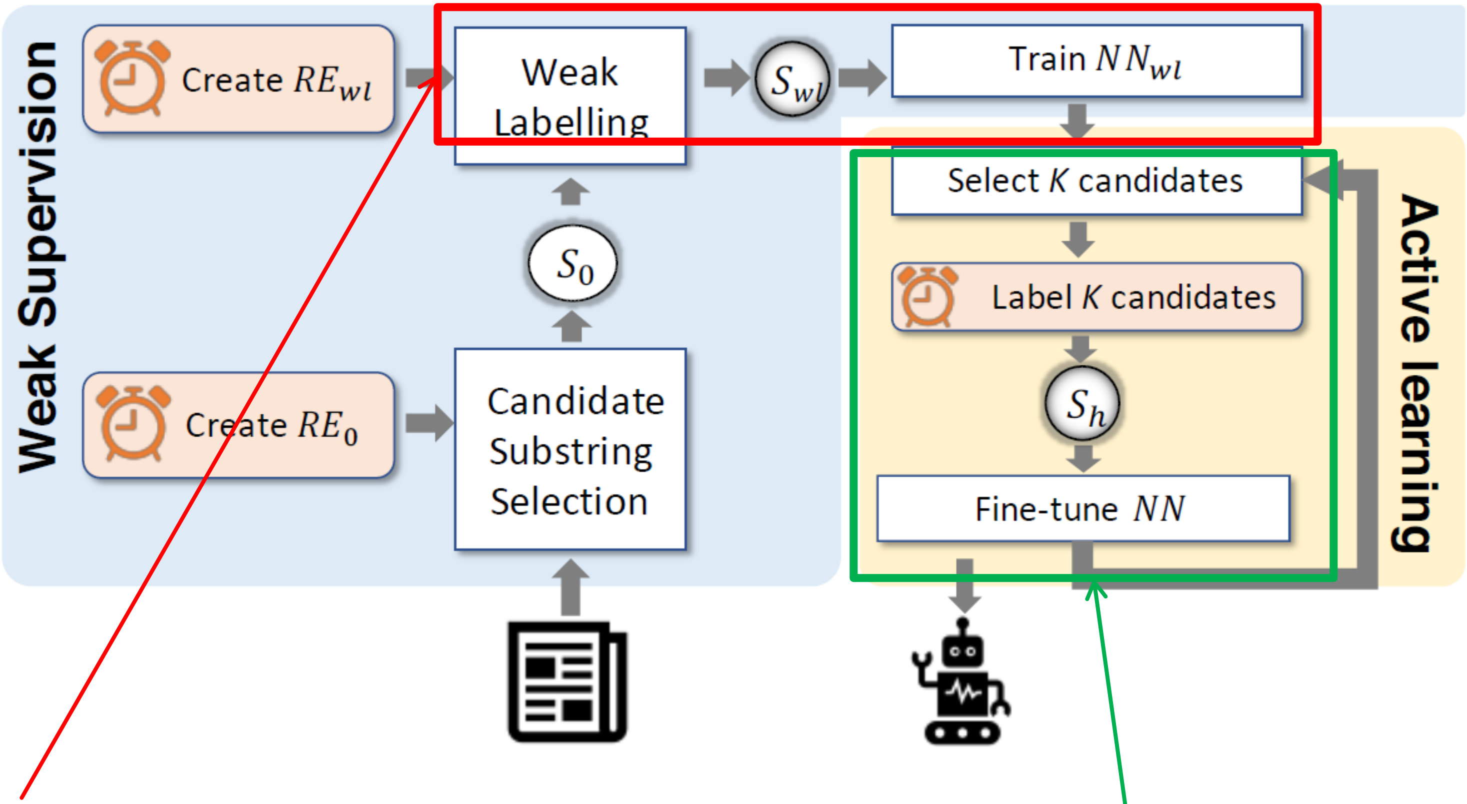
01/27/2021 20:54

Output

Require a lot of human efforts in labeling!

Our Solution

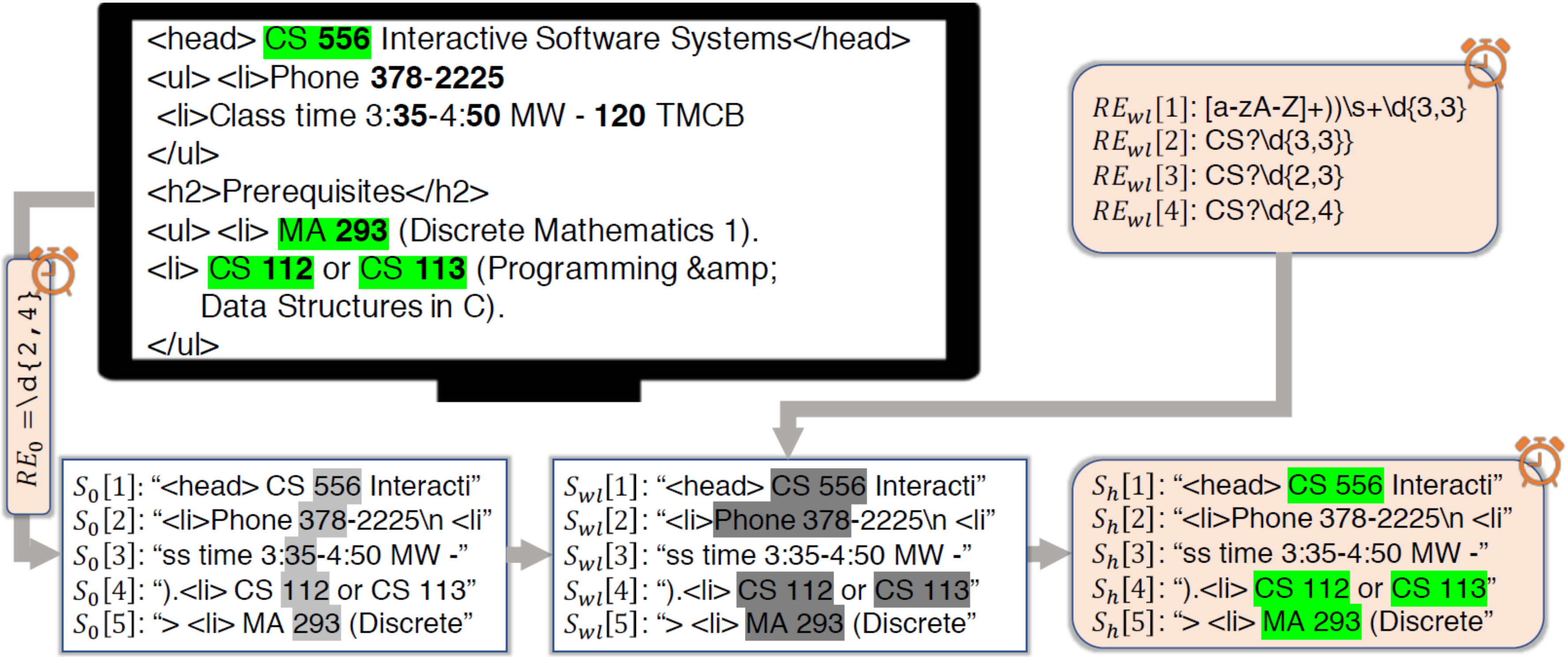
- Weak Supervision + Deep Learning + Active Learning



- Pre-training: RE -> Weak Labels -> NN_{wl}
- Fine-tuning: $|S_h| << |S_0|$, active learning based on entropy.

Our Solution

- Example of Course Number extraction



```
<head> CS 556 Interactive Software Systems</head>
<ul> <li>Phone 378-2225
 <li>Class time 3:35-4:50 MW - 120 TMCB
</ul>
<h2>Prerequisites</h2>
<ul> <li> MA 293 (Discrete Mathematics 1).
<li> CS 112 or CS 113 (Programming &amp;
     Data Structures in C).
</ul>
```

RE_0 = \d{2,4}

$RE_{wl}[1]$: [a-zA-Z]+))\s+\d{3,3}
$RE_{wl}[2]$: CS?\d{3,3}}
$RE_{wl}[3]$: CS?\d{2,3}
$RE_{wl}[4]$: CS?\d{2,4}

$S_0[1]$: "<head> CS 556 Interacti"
$S_0[2]$: "<li>Phone 378-2225\n <li"
$S_0[3]$: "ss time 3:35-4:50 MW -"
$S_0[4]$: ").<li> CS 112 or CS 113"
$S_0[5]$: "> <li> MA 293 (Discrete"

Candidate substrings S_0

$S_{wl}[1]$: "<head> CS 556 Interacti"
$S_{wl}[2]$: "<li>Phone 378-2225\n <li"
$S_{wl}[3]$: "ss time 3:35-4:50 MW -"
$S_{wl}[4]$: ").<li> CS 112 or CS 113"
$S_{wl}[5]$: "> <li> MA 293 (Discrete"

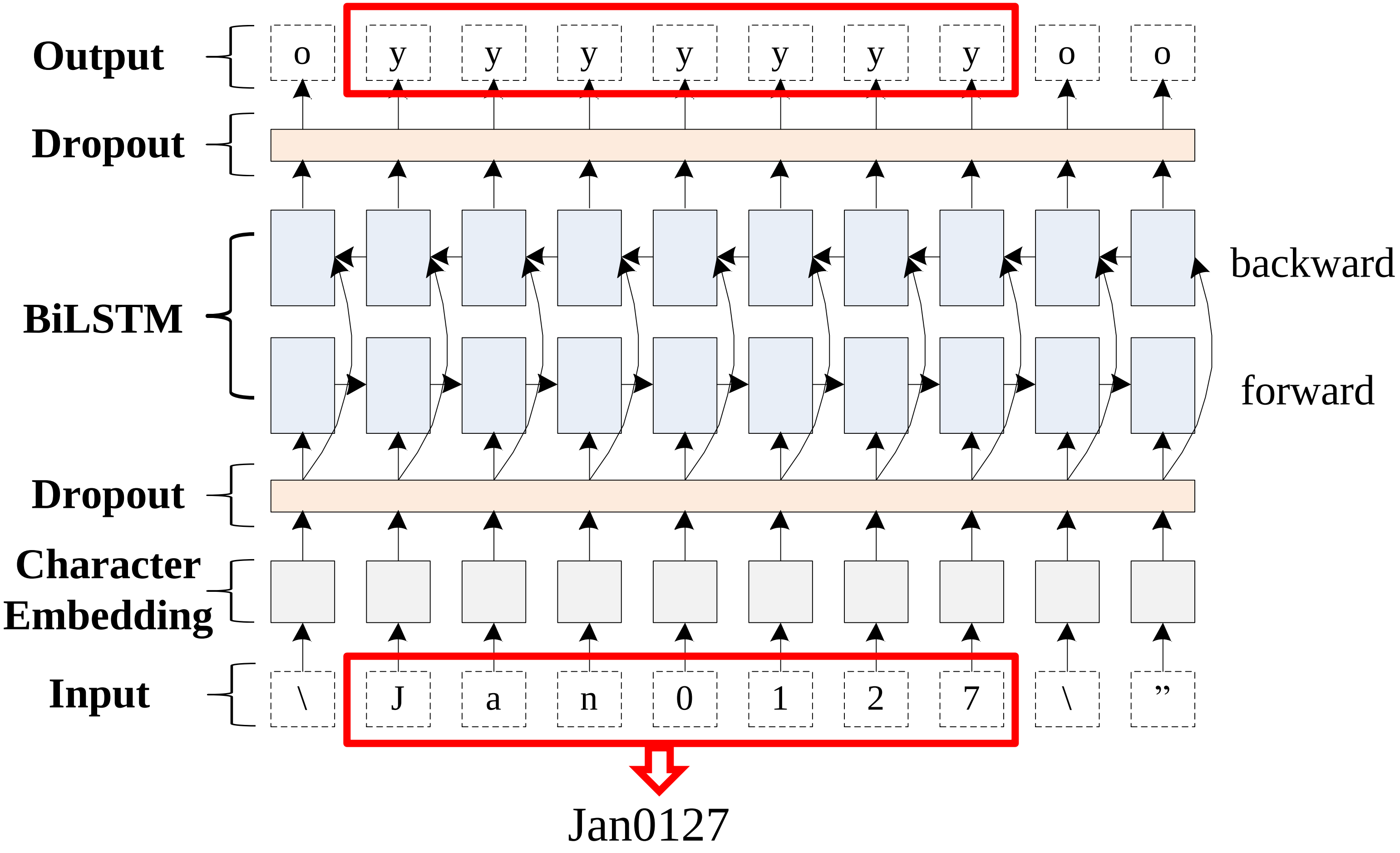
Weak labels S_{wl}

$S_h[1]$: "<head> CS 556 Interacti"
$S_h[2]$: "<li>Phone 378-2225\n <li"
$S_h[3]$: "ss time 3:35-4:50 MW -"
$S_h[4]$: ").<li> CS 112 or CS 113"
$S_h[5]$: "> <li> MA 293 (Discrete"

Human annotations S_h

Our Solution

- Deep Model



Data of Entity Extraction

- 5 tasks

	$\|D\|$	Doc avg length (*chars*)	#entities in D	$\|S_0\|$
Date Time	6,000	137.4K	1,399	761.0K
Course Number	600	4.6K	4,588	43.6K
Phone Number	3,149	2.7K	2,018	25.1K
Email Address	602	1.3K	2,206	5.5K
Bill Date	600	27.5K	3,085	72.2K

Evaluation Metrics

- Character level

$$PosPrec = \frac{\sum_{i=1}^{n} 1(y_i == 1 \cap \hat{y}_i == 1)}{\sum_{i=1}^{n} 1(\hat{y}_i == 1)}$$

$$PosRecall = \frac{\sum_{i=1}^{n} 1(y_i == 1 \cap \hat{y}_i == 1)}{\sum_{i=1}^{n} 1(y_i == 1)}$$

$$PosF1 = \frac{2 \times PosPrec \times PosRecall}{PosPrec + PosRecall}$$

- Entity level

$$EntPrec = {|E_{true} \cap E_{pred}|}\Big/{|E_{pred}|}$$

$$EntRecall = {|E_{true} \cap E_{pred}|}\Big/{|E_{true}|}$$

$$EntF1 = {2 \times EntPrec \times EntRecall}\Big/{EntPrec + EntRecall}$$

Entity Extraction Results

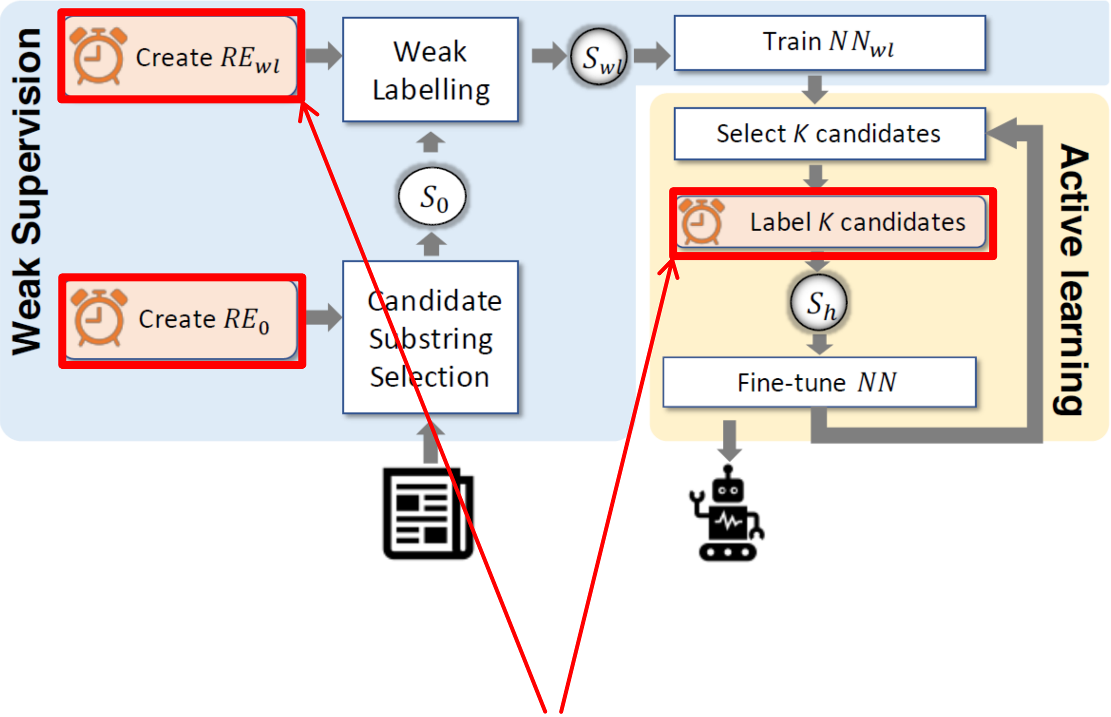
- EntF1 results

Model	Date Time	Course Number	Phone Number	Emaill Address	Bill Date
RE_{wl}	.434	.393	.318	.881	.283
NN_{wl}	.441	.408	.314	.882	.283
NN w/o (100)	.045	.531	.142	.694	.285
NN w (100)	.506	.687	.601	.962	.868
NN w/o (1000)	.837	.841	.797	.990	.934
NN w (1000)	**.888**	**.924**	**.896**	**.995**	**.956**
NN w (300)	.879	.901	.882	.991	.948

User Study

Users Involvement

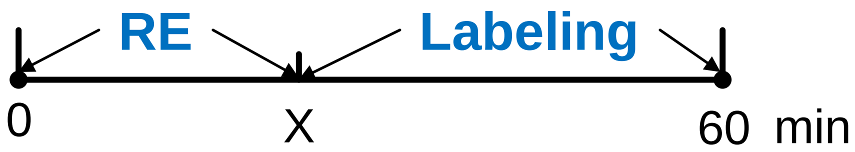
- Create RE_0
  - Ex. DateTime: *\d{4,4}*
  - Small effort
- Create RE_{wl}
- Label Candidates



➢ Study trade-offs between spending time to create a good RE and to manually label the candidate substrings.

User Study

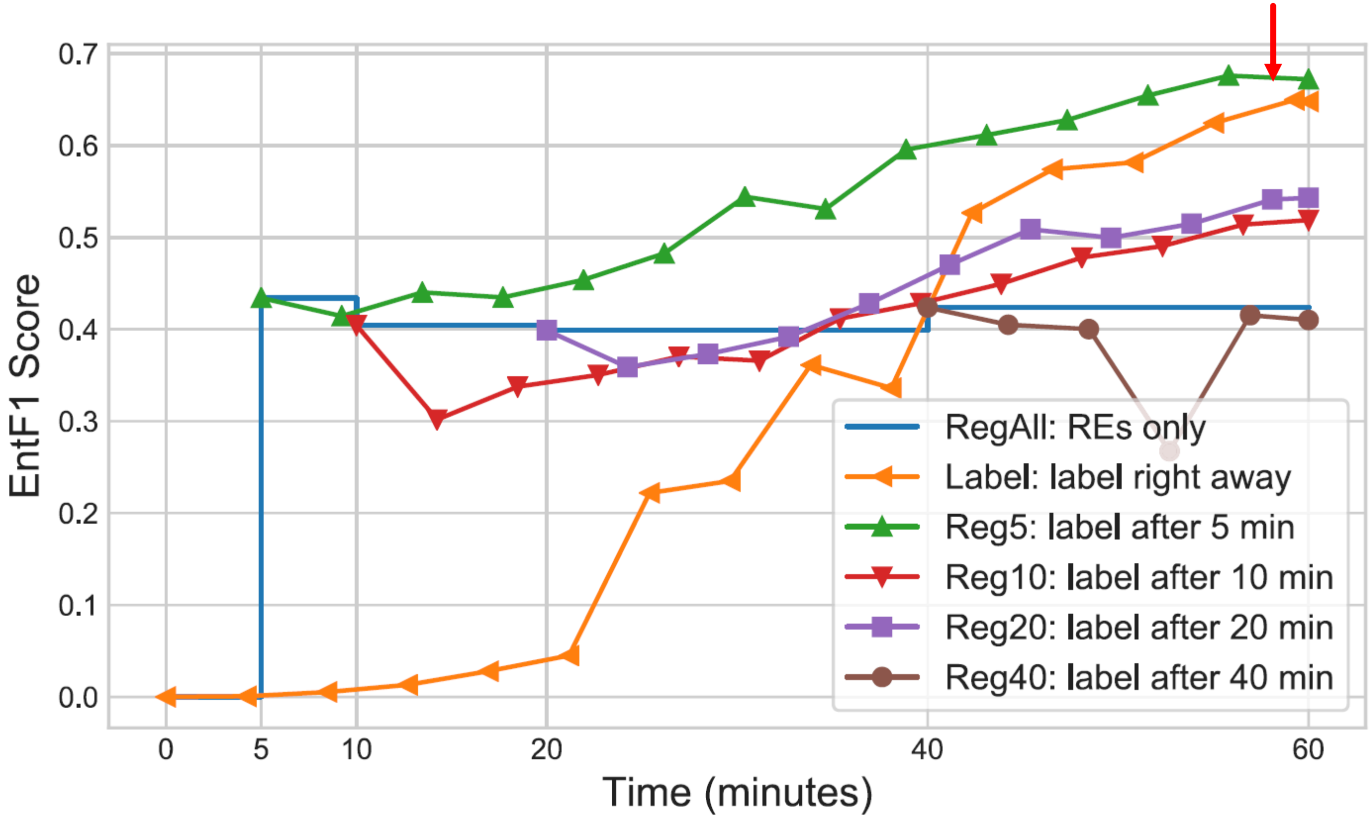
- Experimental Design
  - 4 volunteers, familiar with RE
  - 1k strings in S_0 → construct RE_{wl}
    - * 20210127 06:37:36 * → \d{8,8} \d{2,2}:\d{2,2}:\d{2,2}
    - * 20210127063736 * → \d{14,14}
  - Time Budget



  - Strategies:
    - RegAll: all time on constructing RE_{wl}
    - Label: all time on labeling
    - RegX: X=*5, 10, 20, 40*

User Study

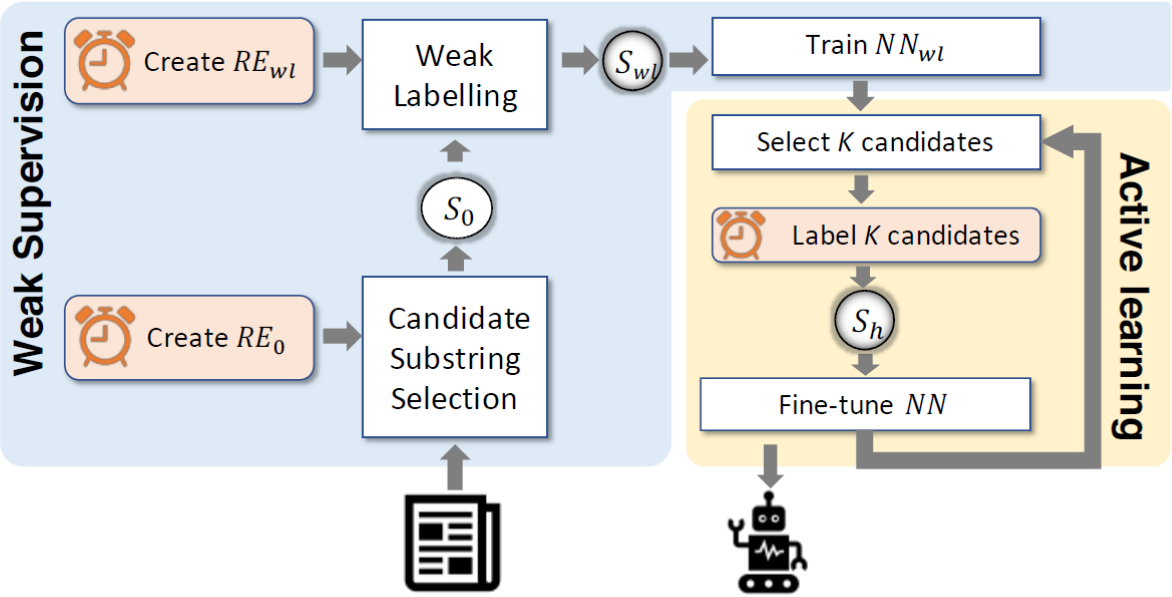
- Study of spending time on RE or labeling, DateTime



➢ Fewer efforts on constructing RE, more on labeling!

Summary

- Entity extraction with few human efforts:
  - Framework
  - User Study



- Publications at EMNLP'18 and KDD'19.