Exploiting Structure in Representation of Named Entities using Active Learning

Nikita Bhutani*
Kun Qian†
Yunyao Li†
H. V. Jagadish*
Mauricio A. Hernandez†
Mitesh Vasa†

* University of Michigan, Ann Arbor
† IBM Research, Almaden

Entities lack unique representation

Barclays

GE Corporation

IBM UK Ltd.

IBM - United Kingdom

Company

Kumagai Professor of Engineering

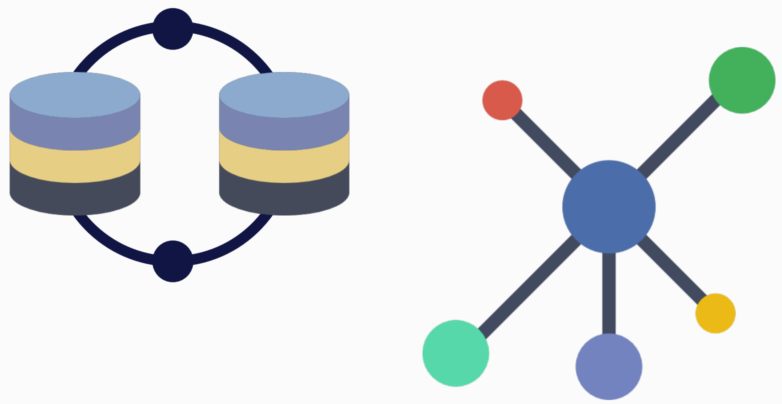
The Helen L. Crocker Faculty Scholar

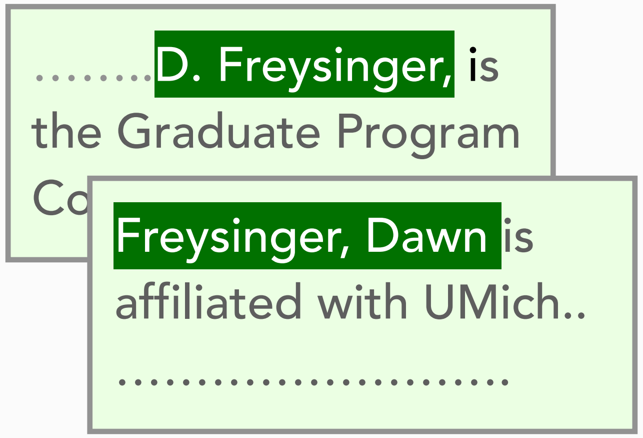
Professor of Public Policy

Kumagai Prof. of Engg.

Academic Title

Entity Linking/Resolution/De-duplication





Entities have an internal structured representation

Company	Academic Title
Barclays	Kumagai Professor of Engineering
GE Corporation	The Helen L. Crocker Faculty Scholar
IBM UK Ltd.	Professor of Public Policy
IBM - United Kingdom	Kumagai Prof. of Engg.

Company

⟨name⟩
⟨loc⟩
⟨suffix⟩

⟨name⟩⟨loc⟩⟨suffix⟩
⟨name⟩⟨suffix⟩
⟨name⟩
....

Academic Title

⟨prefix⟩
⟨position⟩
⟨specialty⟩

⟨prefix⟩⟨position⟩⟨specialty⟩
⟨prefix⟩⟨position⟩
⟨position⟩⟨specialty⟩
....

Structural similarity is more reliable than textual similarity

General Electric Corporation → General Electric | Corporation

⟨name⟩ → *abbreviate* → GE

⟨suffix⟩ → *abbreviate* → Corp.

reasoning over structured representations is more robust!

General Electric Corporation

General Electric China Corporation

GE Corp.

textual similarity can be misleading!

How do we obtain these structured representations?

⟨name⟩⟨suffix⟩
⟨name⟩
⟨name⟩⟨subsidiary⟩⟨suffix⟩...

structured representations

+

"General Electric Corp."
⟨name⟩ *⟨suffix⟩*

programs

Manually[1]

- incorporate domain knowledge (e.g. ⟨suffix⟩ lexicon)
- error-prone, specialized skills, expensive tuning

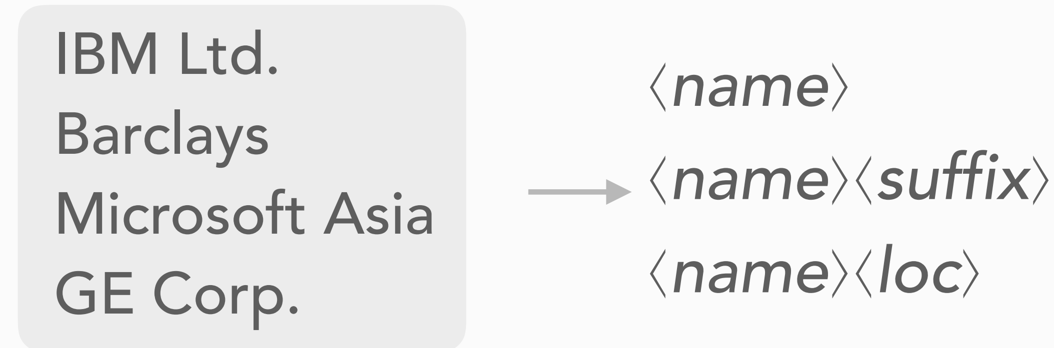
Programmable Framework[2]

- directly manipulate representation of entities
- user has to define a program of grammar rules to parse each mention

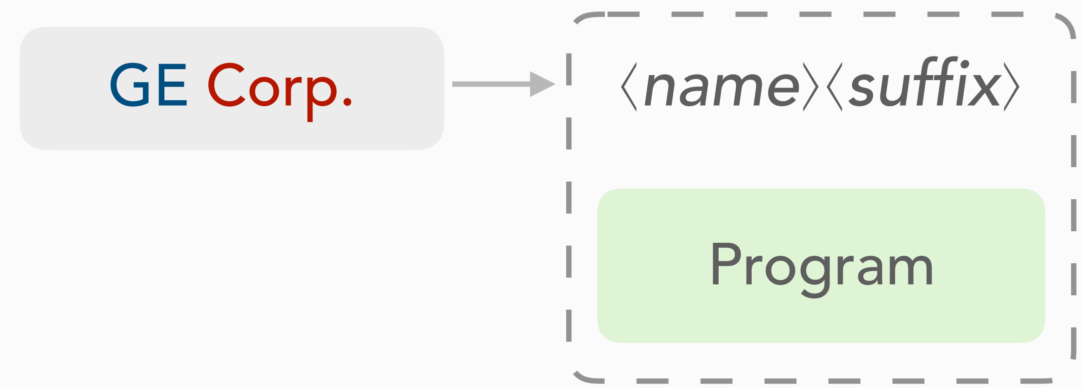
[1] [Campos et al., 2015], [2] [Arasu and Kaushik, 2009]

Reducing user-effort

1. help discover structured representations



2. reduce manual effort in learning structured representations and their programs



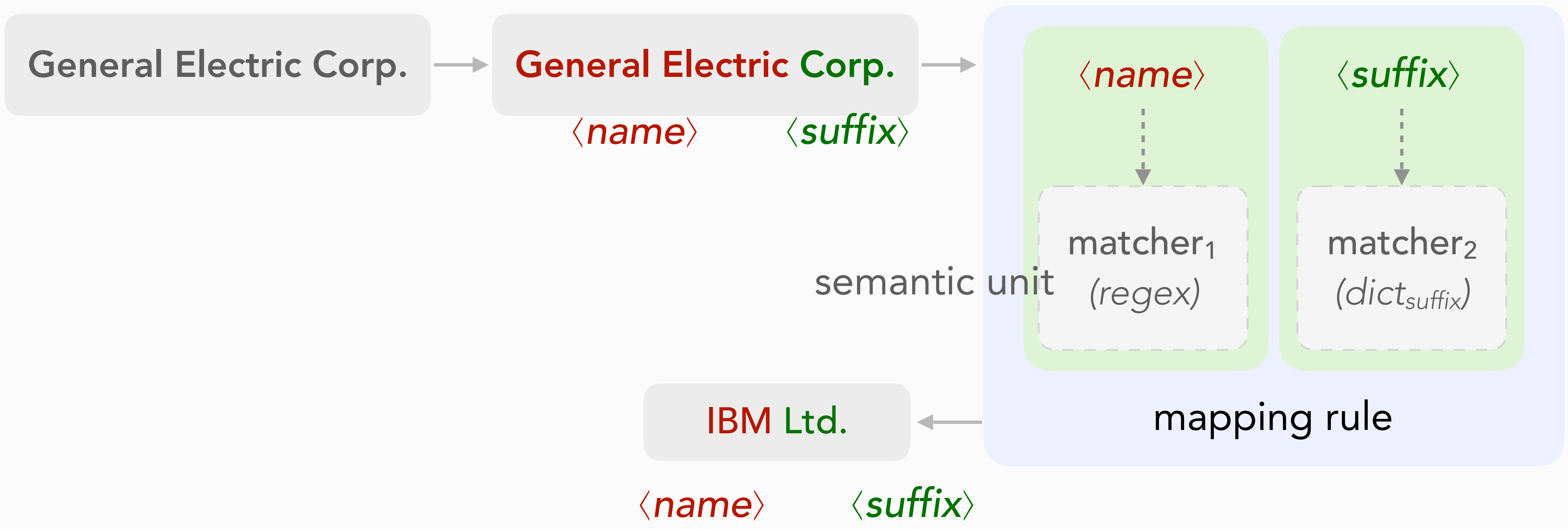
3. incorporate domain knowledge in programs

Key notations and task

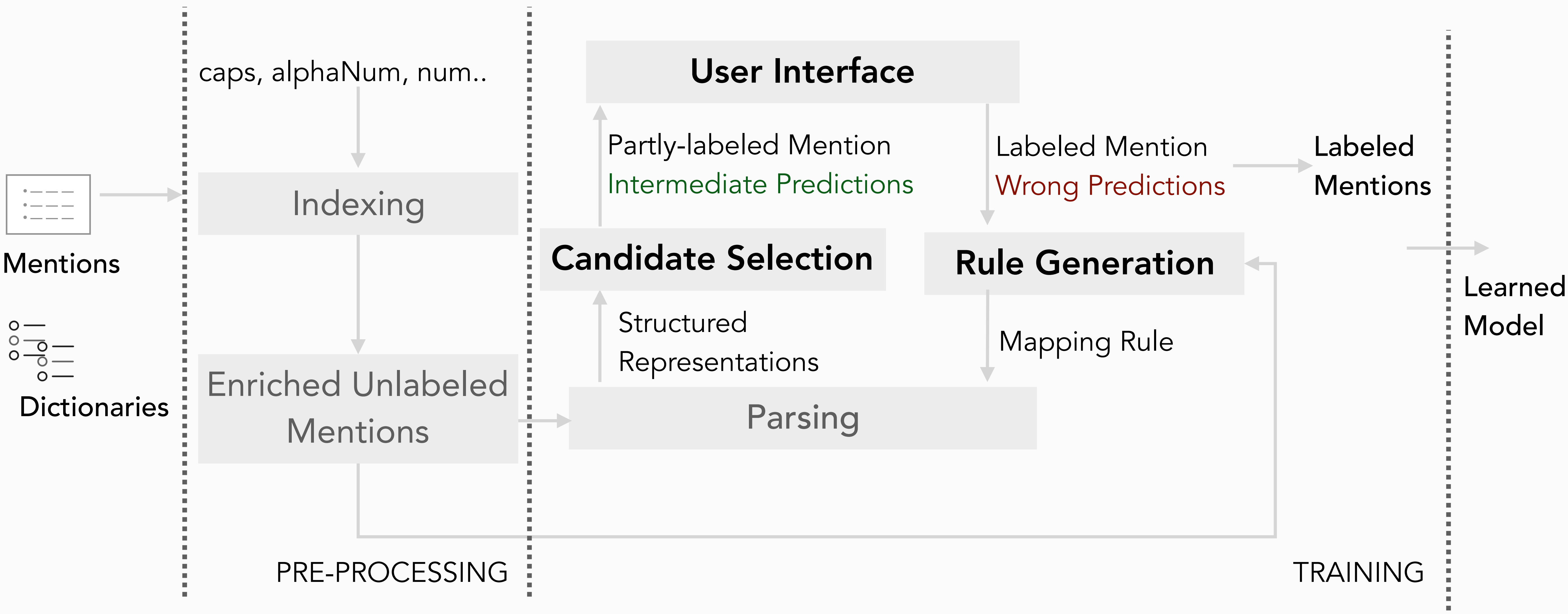
Learn a model of mapping rules with **minimal user effort** by:

- Iteratively seeking labels for **informative** mentions (Active Learning)

- Automatically **infer mapping rules** from user labels (Rule Generation)



LUSTRE System



Inputs and Pre-processing

Inputs

Unlabeled Mentions

Domain Dictionaries

caps, alphanum,num,special..

Built-in regex matchers

matchers

Preprocessing

Evaluate **matchers** against **unlabeled** mentions for candidate selection and rule generation

General	Electric	China	Corporation
caps *alphanum*	*caps* *alphanum*	$d_{country}$ *caps* *alphanum*	d_{suffix} *caps* *alphanum*

Rank matchers to resolve ties: $d_{concept}$ > *caps* > *alphanum* > *num* > *special* > *wild*

Selecting Informative Mention - Query Strategy

Informative Mention

Similar structure as unlabeled mentions

e.g. IBM Ltd. ~ Apple Inc., GE Corp.

Correlation Score:

$$c(m_i) = \mathbf{g}(sim(s_i, s_u)),\ \text{where } u \in U$$

$$sim(s_i, s_u) = 1 - \frac{\text{edit distance}(s_i, s_u)}{\text{max edit distance}}$$

where s_i is the structure of m_i

edit distance(IBM Ltd., GE Corp.) = 0

edit distance(IBM Ltd., Microsoft Asia Inc.) = 1

Unknown or Uncertain structure

e.g. GE Oil & Gas

Uncertainty Score:

$$f(m_i) = \mathbf{f}(P_{r_s})$$

where r_s is the mapping rule of s

Higher the reliability of a mapping rule, lower the uncertainty of its structure

Utility Score:

$$u(m_i) = c(m_i) \times f(m_i)$$

$$m^* = \underset{m_i}{\operatorname{argmax}}\ u(m_i)$$

Seeking user labels for selected mentions

Partly labeled mention

General	Electric	China	Corporation
		⟨*country*⟩	⟨*suffix*⟩

General	Electric	China	Corporation
⟨*name*⟩		⟨*country*⟩	⟨*suffix*⟩

Additional feedback on intermediate predictions

General Motors — ⟨*name*⟩ ✔

IBM — ⟨*name*⟩, UK — ⟨*suffix*⟩ ✗

Barclays — ⟨*name*⟩ ✔

Generating mapping rule

Non-Trivial: semantic units can span multiple tokens and matchers

General	Electric	China	Corporation
⟨*name*⟩		⟨*country*⟩	⟨*suffix*⟩
caps *alphanum*	*caps* *alphanum*	$d_{country}$ *caps* *alphanum*	d_{suffix} *caps* *alphanum*

Solution: reliable rule as the sequence of most **selective**[3] matchers

where selectivity is expected number of matches of a matcher in a dataset

⟨*name:: caps{1,2}*⟩ ⟨*country::* $d_{country}$⟩ ⟨*suffix::*d_{suffix}⟩

[3] [Li et al., 2008]

Updating model with learned rule

Rule Reliability: for query strategy and for resolving structural ambiguities

$P_{r_s} = 1 - \textit{selectivity}(p^*)$
where $p^* = \underset{i}{\text{argmin}}\ \textit{selectivity}(p_i \mid p_i \in r_s)$

For a new rule, estimate as a function of selectivity of matchers in the rule

$P^j_{r_s} = P^i_{r_s} \times (1 - \lambda \text{ frac. incorrect pred})$

For a learned rule, update based on the fraction of predictions of the rule marked incorrect by user

[3] [Li et al., 2008]

Experiments - Datasets, Baselines and Metrics

Datasets

Type	Train	In-Domain	Out-of-domain
Person	200	200	200
Company	200	100	200
Tournament	50	50	-
Academic Title	175	175	-

ACE 2005, Freebase

Baselines

STG[1]: hand-crafted programs used in production

Linear-Chain CRF: sequence labeling with matchers as features

LUSTRE^t: LUSTRE with tf-idf based query strategy

Evaluation Metric

Precision, P: fraction of predictions that are correct

Recall, R: fraction of correct structures that are predicted

Manual effort, α: $\dfrac{\text{F-score of method X}}{\text{\# labels requested by X}}$, where $X \in \{\text{CRF, LUSTRE}\}$

[1] [Campos et al., 2015]

Experiments - Qualitative Analysis

Type	Method	In-Domain			Out-of-domain		
		P	R	F_1	P	R	F_1
Person	STG	0.92	0.92	0.92	0.85	0.85	0.85
	CRF	0.97	0.97	0.97	0.90	0.90	0.90
	$LUSTRE^t$	0.98	0.95	0.96	0.92	0.90	0.91
	LUSTRE	0.99	0.97	0.98	0.92	0.95	0.93
Company	STG	0.83	0.83	0.83	0.79	0.79	0.79
	CRF	0.87	0.87	0.87	0.85	0.85	0.85
	$LUSTRE^t$	0.84	0.77	0.80	0.78	0.60	0.68
	LUSTRE	0.95	0.86	0.90	0.91	0.85	0.88
Tournament	CRF	0.70	0.70	0.70	-	-	-
	$LUSTRE^t$	0.96	0.68	0.79	-	-	-
	LUSTRE	0.96	0.90	0.93	-	-	-
Academic Title	CRF	0.69	0.69	0.69	-	-	-
	$LUSTRE^t$	0.36	0.23	0.28	-	-	-
	LUSTRE	0.79	0.65	0.72	-	-	-

Learned Models (LUSTRE, CRF) > Manually Crafted (STG)

Experiments - Qualitative Analysis

Type	Method	In-Domain			Out-of-domain		
		P	R	F_1	P	R	F_1
Person	STG	0.92	0.92	0.92	0.85	0.85	0.85
	CRF	0.97	0.97	0.97	0.90	0.90	0.90
	$LUSTRE^t$	0.98	0.95	0.96	0.92	0.90	0.91
	LUSTRE	0.99	0.97	0.98	0.92	0.95	0.93
Company	STG	0.83	0.83	0.83	0.79	0.79	0.79
	CRF	0.87	0.87	0.87	0.85	0.85	0.85
	$LUSTRE^t$	0.84	0.77	0.80	0.78	0.60	0.68
	LUSTRE	**0.95**	**0.86**	**0.90**	**0.91**	**0.85**	**0.88**
Tournament	CRF	0.70	0.70	0.70	-	-	-
	$LUSTRE^t$	0.96	0.68	0.79	-	-	-
	LUSTRE	**0.96**	**0.90**	**0.93**	-	-	-
Academic Title	CRF	0.69	0.69	0.69	-	-	-
	$LUSTRE^t$	0.36	0.23	0.28	-	-	-
	LUSTRE	**0.79**	**0.65**	**0.72**	-	-	-

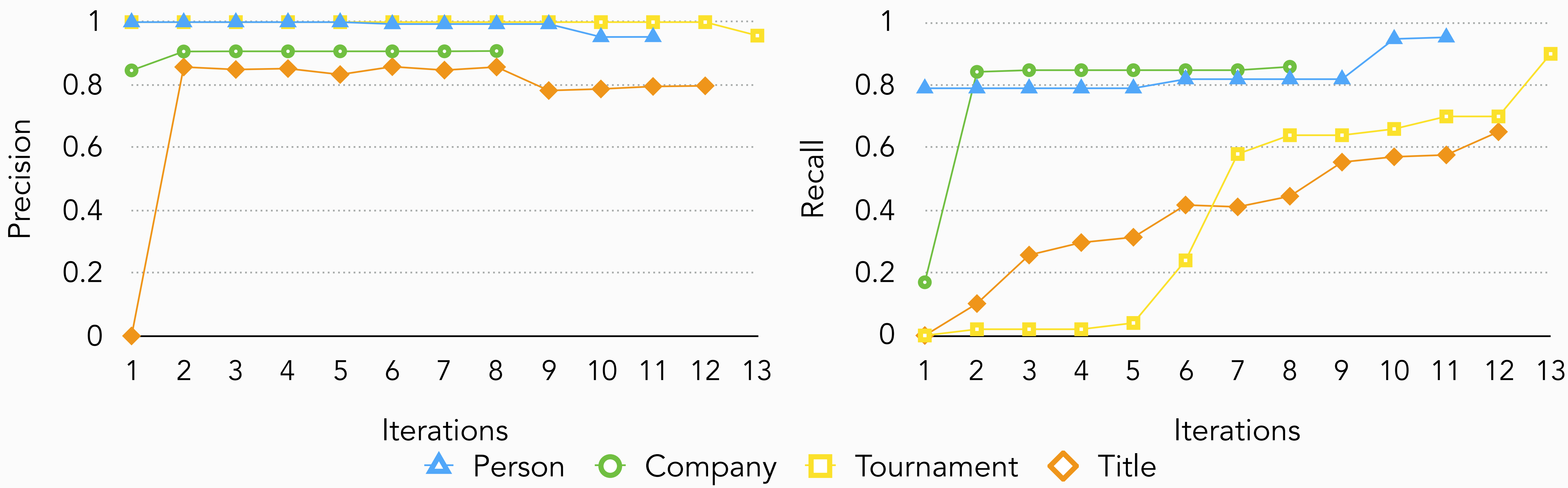
Complex entities have more variations. LUSTRE outperforms other methods for complex types.

Experiments - Qualitative Analysis

Type	Method	In-Domain			Out-of-domain		
		P	R	F_1	P	R	F_1
Person	STG	0.92	0.92	0.92	0.85	0.85	0.85
	CRF	0.97	0.97	0.97	0.90	0.90	0.90
	LUSTRE^t	0.98	0.95	0.96	0.92	0.90	0.91
	LUSTRE	0.99	0.97	0.98	**0.92**	**0.95**	**0.93**
Company	STG	0.83	0.83	0.83	0.79	0.79	0.79
	CRF	0.87	0.87	0.87	0.85	0.85	0.85
	LUSTRE^t	0.84	0.77	0.80	0.78	0.60	0.68
	LUSTRE	0.95	0.86	0.90	**0.91**	**0.85**	**0.88**
Tournament	CRF	0.70	0.70	0.70	-	-	-
	LUSTRE^t	0.96	0.68	0.79	-	-	-
	LUSTRE	0.96	0.90	0.93	-	-	-
Academic Title	CRF	0.69	0.69	0.69	-	-	-
	LUSTRE^t	0.36	0.23	0.28	-	-	-
	LUSTRE	0.79	0.65	0.72	-	-	-

Good out-of-domain performance indicates LUSTRE captures structures regardless of data source.

Experiments - Effectiveness



Constant precision indicates "quality" rules are learned - **effective program synthesis**

Increasing recall indicates new rules are learned - **effective query strategy**

Few iterations (8-13) indicate low manual effort

Type	LUSTRE	CRF
Person	**0.089**	0.005
Company	**0.125**	0.004
Tournament	**0.072**	0.014
Title	**0.060**	0.004

Manual effort, α

Experiments - Usefulness

Relation Extraction

MULTIR[5]: uses weak supervision data created by **exact matching** textual mentions to Freebase entities

matching variations: textual mentions to variations of Freebase entities of type *Person* and *Company*

# of exact matches: 24,882 sentences

# of matches to variations: 34,197 sentences

F1-score: Increased by 3%

Entity Resolution[4]

Refer paper for more details

[4] [Qian et al., 2017], [5] [Hoffmann et al., 2011]

Conclusion

Framework to reason about name variations of entities based on their **structured representations**

An **active-learning** approach to learn structures for an entity type with **minimal human input**

Automatically synthesize generalizable programs from human-understandable labels for structures

Demo Paper: *An Interactive System for Entity Structured Representation and Variant Generation, ICDE 2018*

Video Link: **https://www.youtube.com/watch?v=llaT4Sz6uI4**

Thank You