Coupled Semi-Supervised Learning for Information Extraction

Andrew Carlson, Justin Betteridge,
Richard C. Wang, Estevam R. Hruschka Jr.
and Tom M. Mitchell

Machine Learning Department
Carnegie Mellon University
February 4, 2010

Read the Web

- Project Goal:
  - System that runs 24x7 and continually
    - Extracts knowledge from web text
    - Improves its ability to do so
  - … with limited human effort
  - Learn more at http://rtw.ml.cmu.edu
    - (or search for "read the web cmu")

Problem Statement

- Given initial ontology containing:
  - Dozens of categories and relations
    - (e.g., Company and CompanyHeadquarteredInCity)
  - Relationships between categories and relations
  - 15 seed examples of each
- Task:
  - Learn to extract new instances of categories and relations with high precision
  - Run over 200 million web pages, for a few days

General Approach

- Exploit relationships among categories and relations through *coupled semi-supervised learning*
  - Coupled Textual Pattern Learning
    - e.g., "President of X"
  - Coupled Wrapper Induction
    - Learn to extract from lists and tables
  - Coupling multiple extraction methods
    - Couples the above two methods by combining predictions

Why Is This Worthwhile?

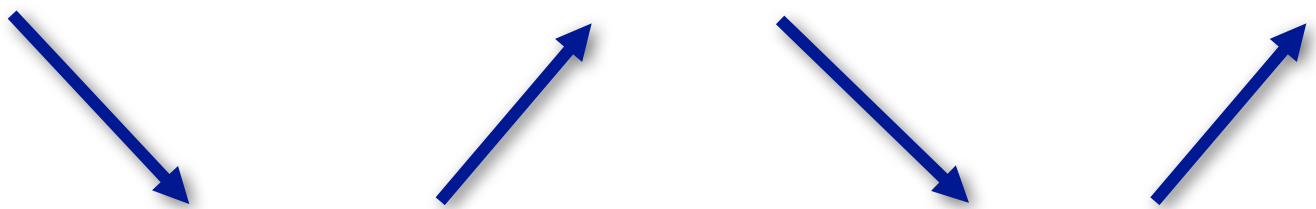
- Semi-supervised methods for information extraction are promising, but suffer from divergence (Riloff and Jones 99, Curran 07)
  - Potential for advances in semi-supervised machine learning
- Extracted knowledge useful for many applications:
  - Computational Advertising
  - Search
  - Question Answering
  - Soumen's vision from this morning's keynote

Bootstrapped Pattern Learning: Countries (Brin 98, Riloff and Jones 99)

Canada
Egypt
France
Germany
Iraq

Pakistan
Sri Lanka
Argentina
Greece
Russia

…



countries except X
X is the only country
home country of X

GDP of X
elected president of X
X has a multi-party system

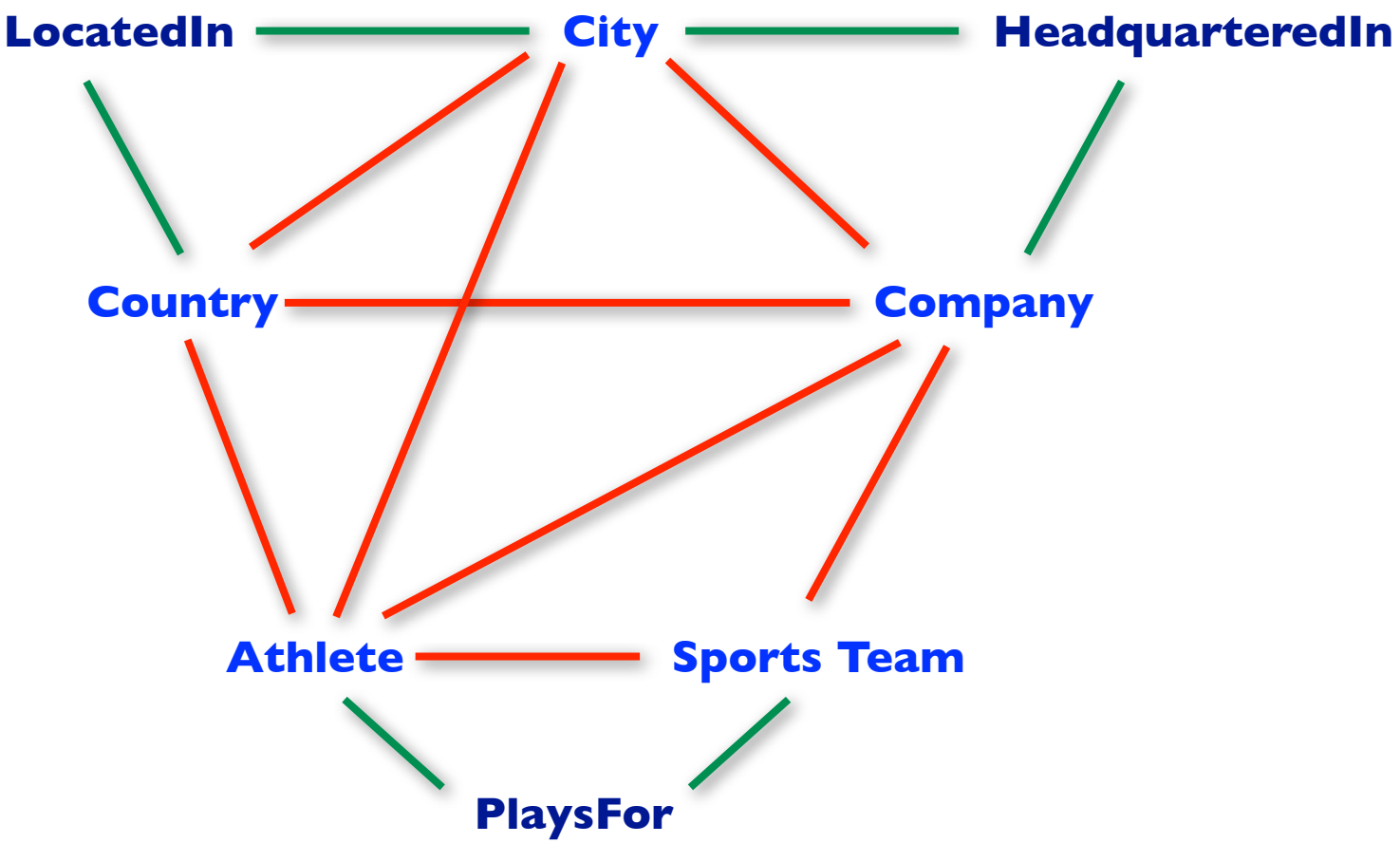
Semantic Drift (Curran 07)

Canada
Egypt
France
Germany
Iraq
....

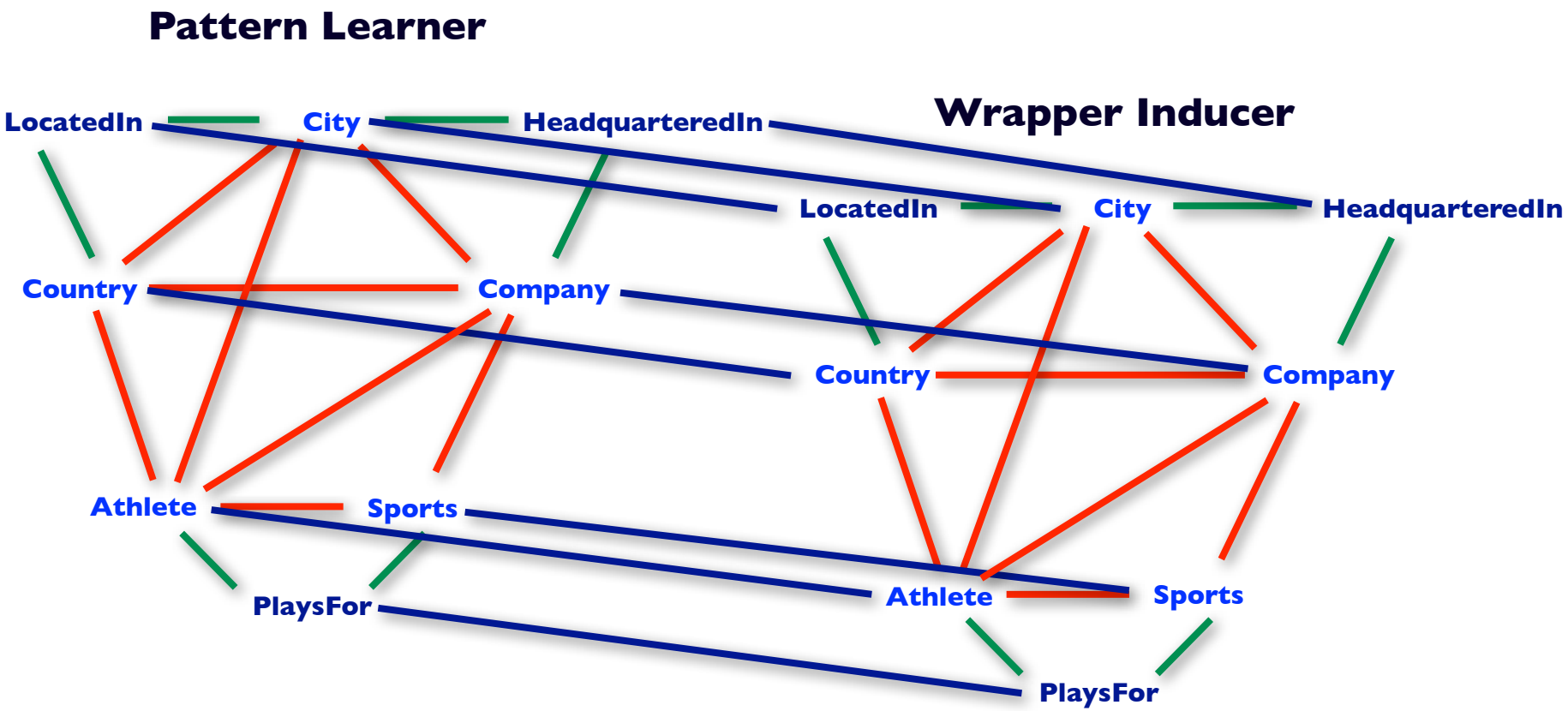
war with X
ambassador to X
war in X
occupation of X
invasion of X

planet Earth
Freetown
North Africa

Coupled Learning of Many Functions



Coupling Different Extraction Techniques

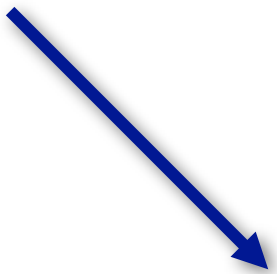


Avoiding Semantic Drift: Mutual Exclusion

Positives:
Canada
Egypt
France
Germany
Iraq
....

war with X
ambassador to X
war in X
occupation of X
invasion of X

planet Earth
Freetown
North Africa

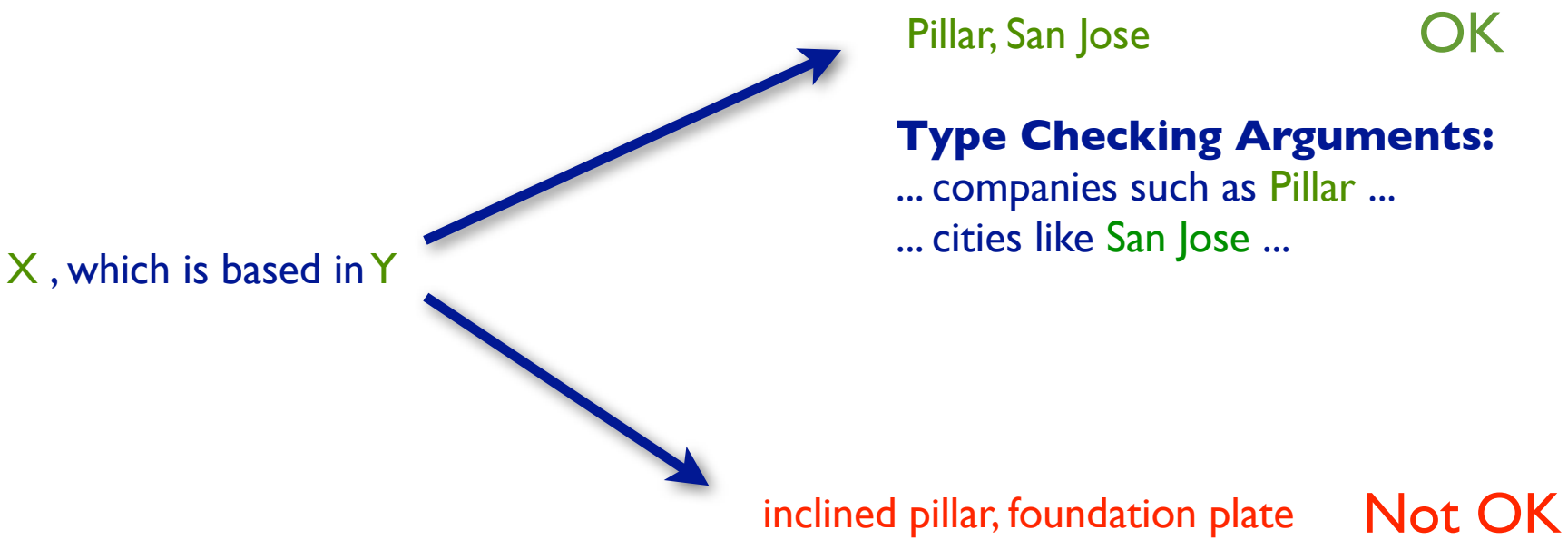


Negatives:
Asia
Europe
London
Florida
Baghdad
...

nations like X
countries other than X
country like X
nations such as X
countries , like X

Pakistan
Sri Lanka
Argentina
Greece
Russia

Avoiding Semantic Drift: Type Checking



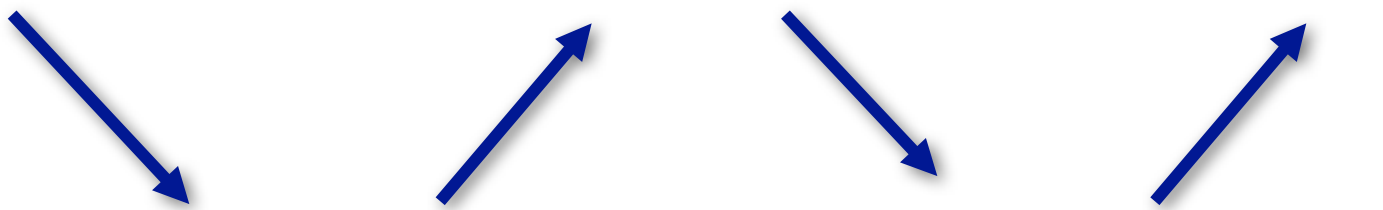
SEAL: Set Expander for Any Language (Wang and Cohen, 2007)

Bootstrapping Wrapper Induction

Canada
Egypt
France
Germany
Iraq

Pakistan
Sri Lanka
Argentina
Greece
Russia

…

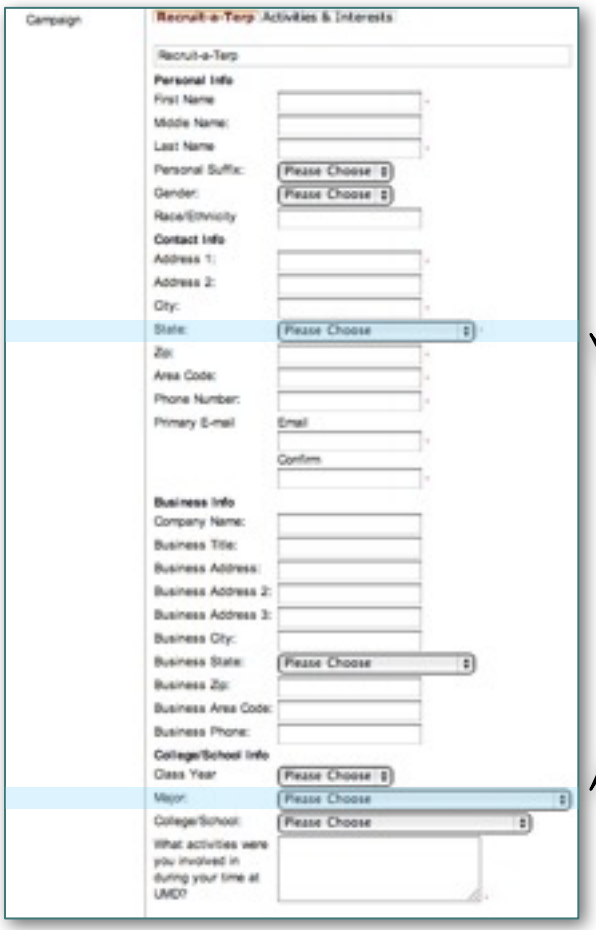


SEAL Wrappers:
(URL, Extraction Template)
(URL, Extraction Template)
(URL, Extraction Template)

More SEAL Wrappers:
(URL, Extraction Template)
(URL, Extraction Template)
(URL, Extraction Template)

Can SEAL benefit from Coupling?

Query: Economics History Biology



Major:

- Audiology
- Aviation Science
- Airway Science
- Bus Admin and Computer Info Systems
- Bacteriology
- Biochemistry
- Business Comm & Liberal Arts
- Building Construction Teaching
- Behavior, Ecology, Evolution,Systematics
- Behavioral & Social Science
- Behavioral Studies

Wrapper:“>[X]</option>

State:

- Alabama
- Military – Pacific
- Arkansas
- American Samoa
- Arizona
- California
- Colorado
- Connecticut

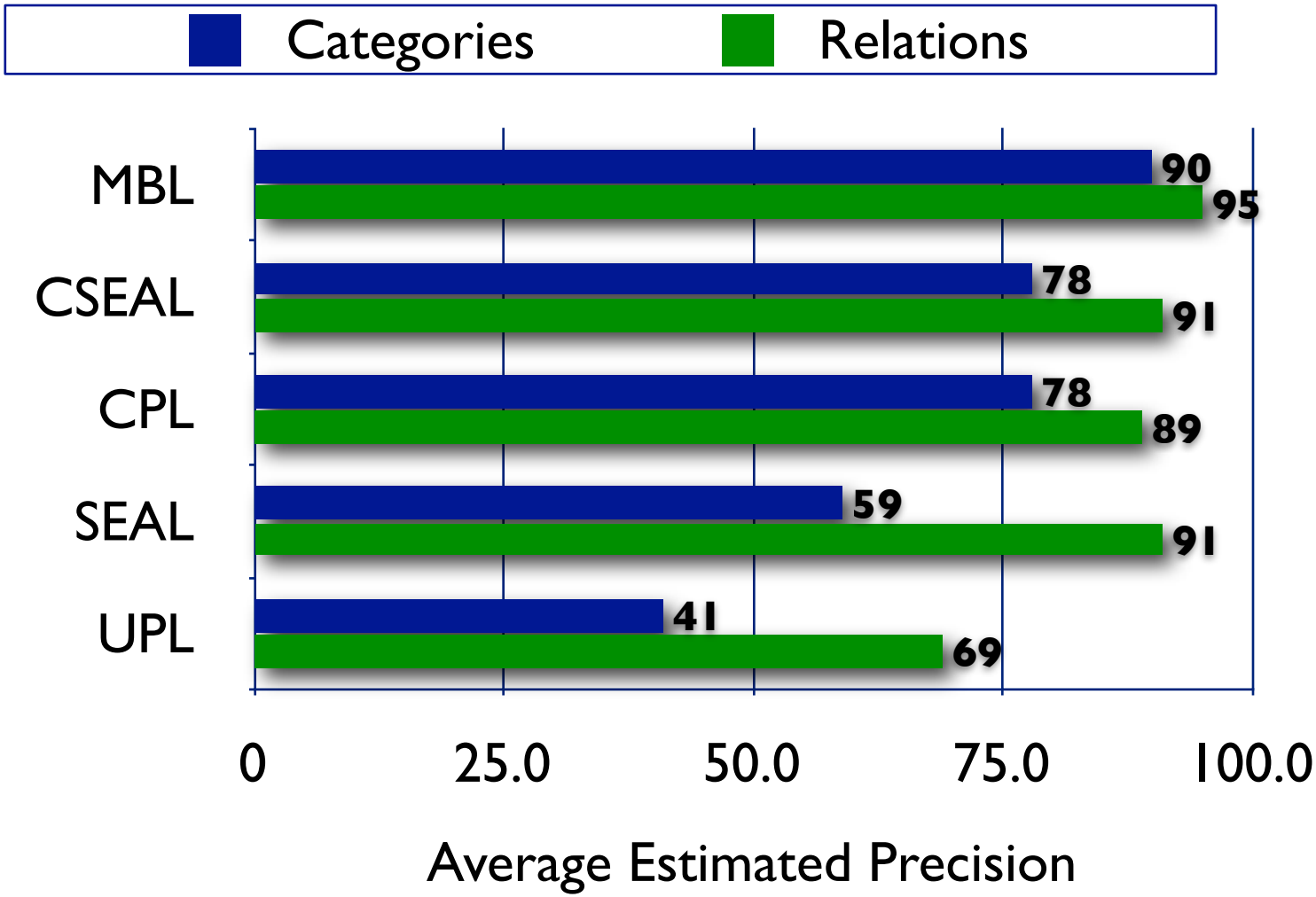
Coupling Multiple Extraction Techniques

- Intuition
  - Different extractors make independent errors
- Strategy (Meta-Bootstrap Learner)
  - Only promote instances recommended by multiple techniques

Experimental Evaluation

- 76 predicates
  - 32 relations, 44 categories
- Run different algorithms for 10 iterations:
  - MBL: Meta-Bootstrap Learner (CPL + CSEAL)
  - CSEAL: Coupled SEAL
  - CPL: Coupled Pattern Learner
  - SEAL: Uncoupled SEAL
  - UPL: Uncoupled Pattern Learner
- Evaluate correctness of instances with Mechanical Turk

Precision of Promoted Instances



Example Promoted Instances

Instance	**Predicate**
solomon islands	country
stuffit	product
marine industry	economicSector
soccer, player	sportUsesEquipment
unocal, oil	companyEconomicSector
final cut pro, software	productInstanceOf

Example Patterns

Pattern	Predicate
blockbuster trade for X	athlete
airlines , including X	company
personal feelings of X	emotion
X announced plans to buy Y	companyAcquiredCompany
X learned to play Y	athletePlaysSport
X dominance in Y	teamPlaysInLeague

Error Analysis

- Worst performers:
  - Sports Equipment
  - Product Type
  - Traits
  - Vehicles
- The good news: More coupling should help!

Conclusions

- Coupling Semi-Supervised Learning of Categories and Relations:
  - Improves free text pattern learning (CPL)
  - Improves semi-structured IE (CSEAL)
  - Improves separate techniques that make independent errors (MBL)

What's Next?

- More components:
  - Morphology Classifier
  - Rule Learner
- More predicates: 100+ categories, 50+ relations
- More iterations: (more efficient code)
- More data: ClueWeb09 (2.5B unique sentences)
- Results from a recent run:
  - 88k facts, 90% precision (vs. 9.5k, 90%)

Acknowledgments

Jamie Callan et al.: Web corpora

CNPq and CAPES: Funding

DARPA: Funding

Google: Funding

Yahoo!: PhD Student Fellowship, M45 Cluster

Thank you

Online Materials:

http://rtw.ml.cmu.edu/wsdm10_online

(includes seed ontology, promoted items, learned patterns, Mechanical Turk templates)

Questions?