Extracting Meronyms for a Biology Knowledge Base Using Distant Supervision

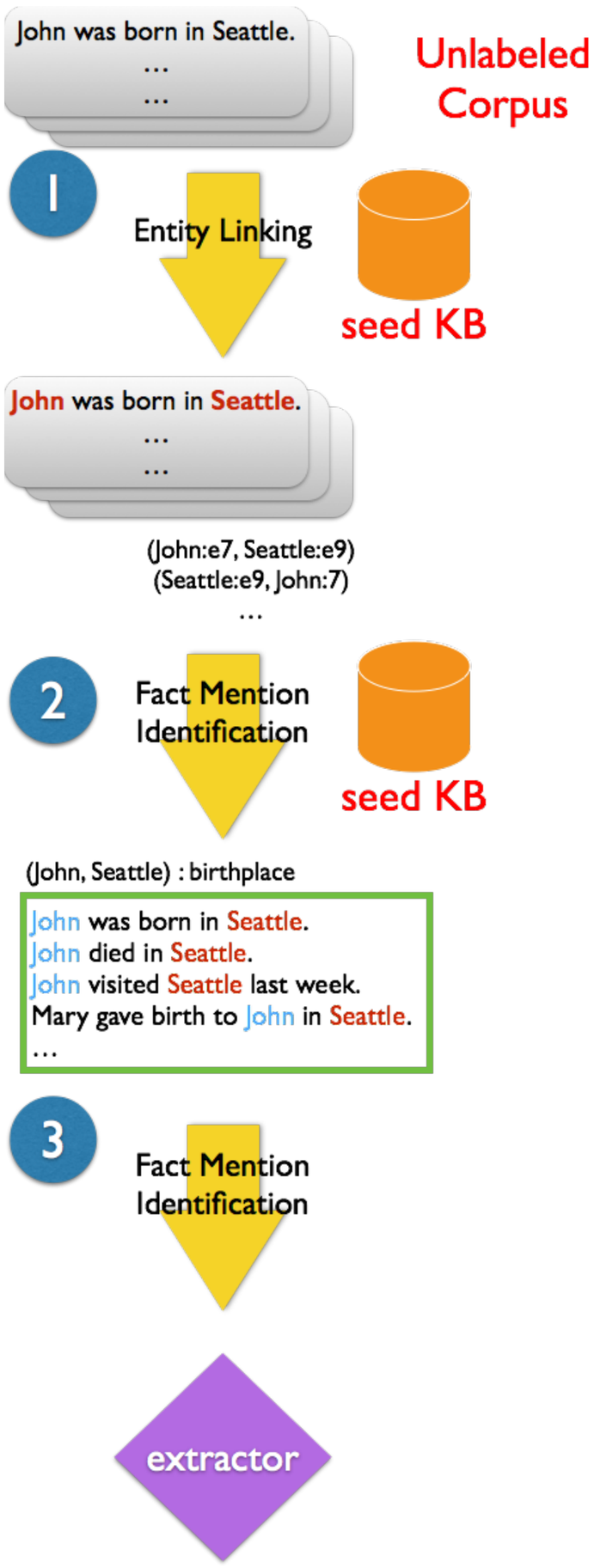
Xiao Ling† Peter Clark* Daniel S. Weld†

†University of Washington *Allen Institute for Artificial Intelligence

Overview

- Goal: Build a **KB** of *meronym* **relations for the domain of** *biology*
- Methodology: use **distant supervision** to learn a meronym relation extractor.
- Data: 1) a seed KB of meronym facts; 2) a unlabeled text corpus (i.e. a textbook "Campbell Biology".)
- Evaluation: a held-out set (172 has-part, 206 NA)

Distant Supervision



Expanding mentions with co-reference

- Standard: NER + string match
- Named Entity Linking
- This work: dictionary max-span matching

	Recall	Precision	F1
CV	0.664	0.820	0.733
CV+COREF	**0.674**	**0.821**	**0.740**
TEST	0.663	**0.857**	0.748
TEST+COREF	**0.744**	0.795	**0.769**

Example:

One of the sex-determining chromosomes is X chromosome ; and it has around 1,100 genes.

X-Chromosome → Gene

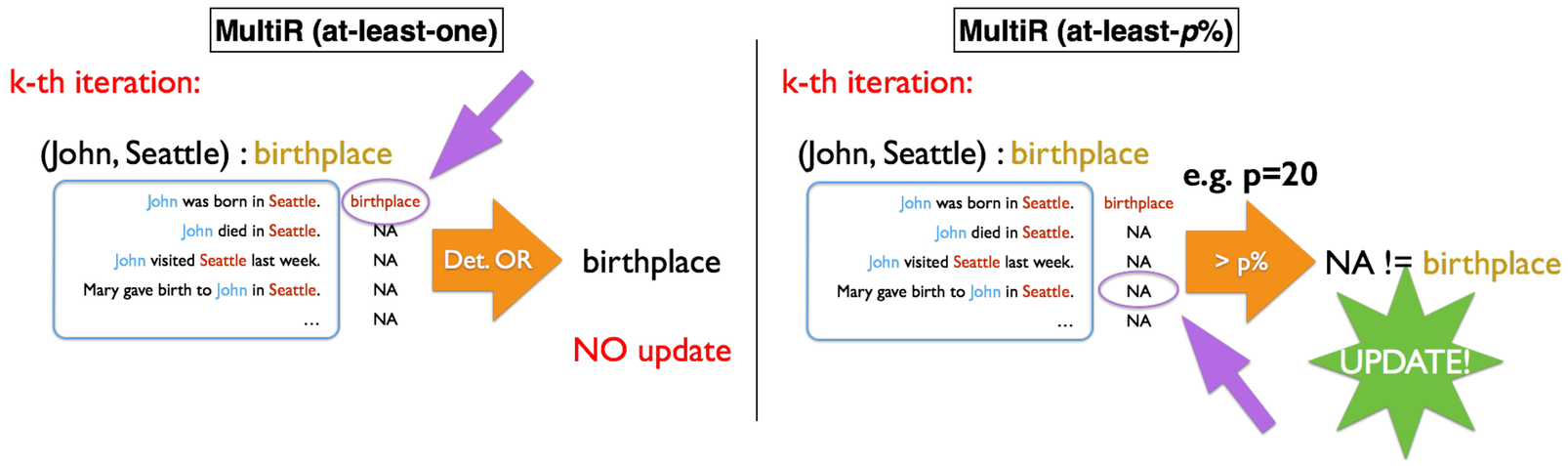
Generating negative data

- The seed KB provides mostly valid facts while invalid facts can be useful as negative data.
- To generate negative examples:
  - ▷ random sample (closed world)
  - ▷ functional relations

	#+	#-	Recall	Precision	F1
CV(BASE)	887	77	**0.856**	0.389	0.533
CV(REV)	887	963	0.706	0.748	0.725
CV(TRANS)	887	2566	0.674	**0.821**	**0.740**
TEST(BASE)	887	77	**0.884**	0.596	0.712
TEST(REV)	887	963	0.709	0.718	0.713
TEST(TRANS)	887	2566	0.744	**0.795**	**0.769**

This work:
- **reverse**: if $(e_1, \text{has-part}, e_2)$ then (e_2, NA, e_1)
- **transitive closure**

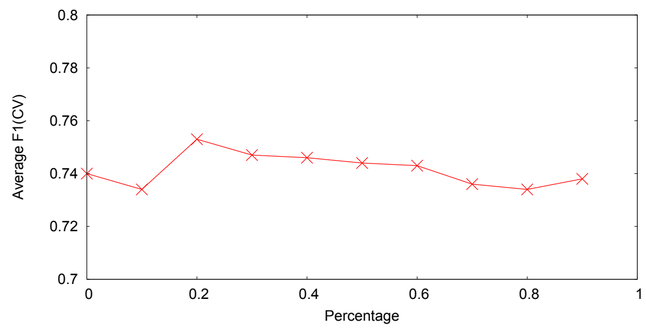
at-least-p% assumption



Leveraging supervision from out-of-domain

- additional training facts: meronyms from WordNet
- We match the WordNet meronyms to Wikipedia articles.

	p	#+	#-	Recall	Prec	F1
TEST	0.0	887	2566	0.384	**0.892**	0.537
TEST+WN	0.0	1578	2566	**0.401**	0.841	**0.543**
TEST	0.2	887	2566	0.523	**0.811**	0.636
TEST+WN	0.2	1578	2566	**0.558**	0.793	**0.655**



	Recall	Precision	F1
CV($p = 0$)	0.674	**0.821**	0.740
CV($p = 0.2$)	**0.730**	0.776	**0.753**
TEST($p = 0$)	0.744	**0.795**	0.769
TEST($p = 0.2$)	**0.791**	0.786	**0.788**

Automated Knowledge Base Construction (AKBC) 2013. xiaoling@cs.washington.edu, PeterC@vulcan.com, weld@cs.washington.edu