Bootstrapping Biomedical Ontologies for Scientific Text using NELL

Dana Movshovitz-Attias and William W. Cohen

Carnegie Mellon University
June 8, 2012

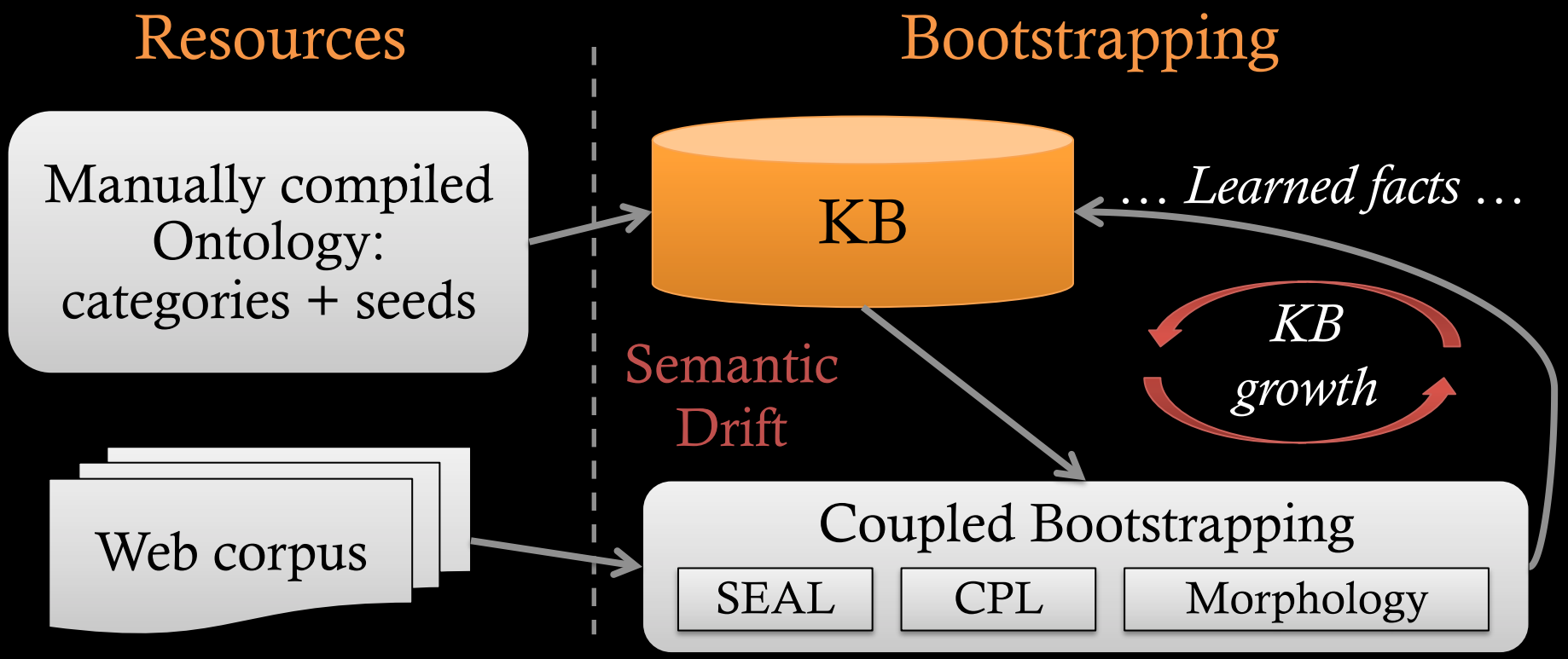
Goal

- Information extraction system for biomedical information
  - Learn a wide range of sub domains

- Approach
  - Adapt existing general purpose system (NELL) to biomedical domain

Never Ending Language Learner (NELL)

Semi-supervised learning system for extraction of information from the Web



Challenges

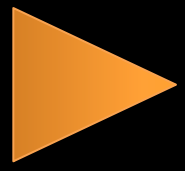
- Creating a biomedical ontology
  - Categories: interesting concepts
  - Seeds: examples for each category

  - Domain knowledge needed to build manually

- Ambiguity in biomedical terminology lead to *semantic drift* in KB

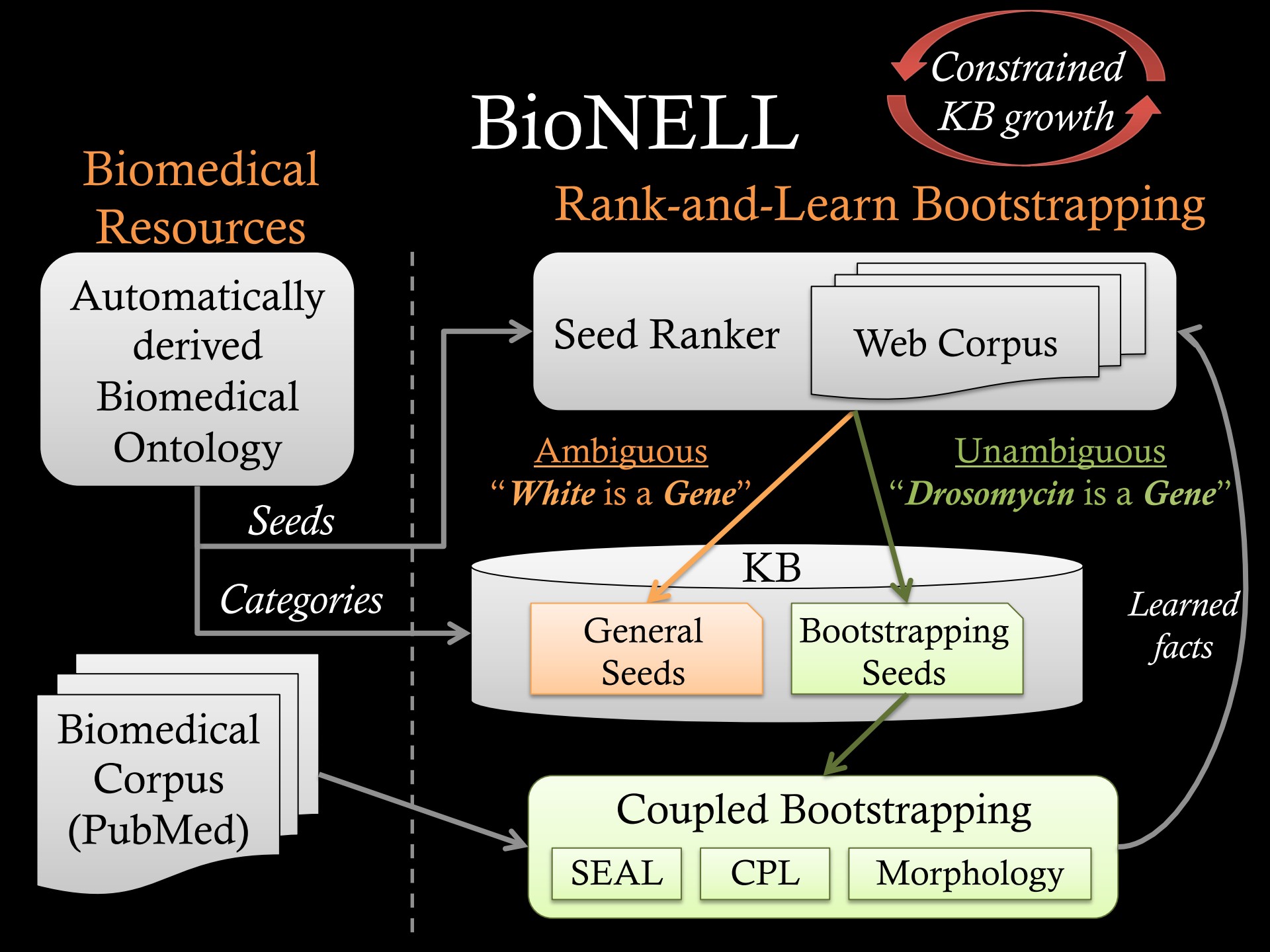
Ambiguity in Biomedical Terminology

- Sources of ambiguity:
  - Short form names and abbreviations
    - Non-meaningful morphological structure
    - Limited number of short abbreviations - overlap
  - Ambiguous names for genes, organisms, systems
    - "white" gene mutation
    - "peanut" is a plant and gene
    - Gene names often shared across species



Confusing training examples

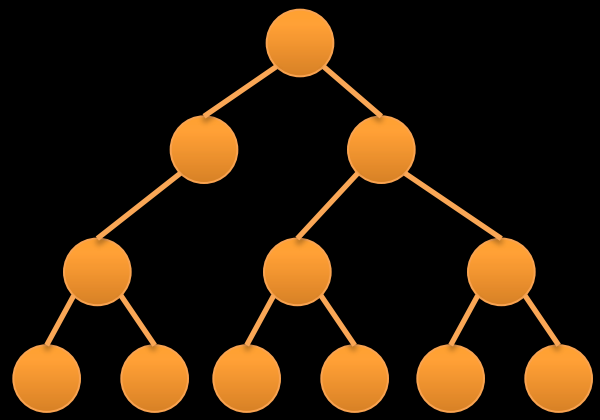
M. Krallinger, A. Valencia, and L. Hirschman. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol


BioNELL
Constrained KB growth
Biomedical Resources
Rank-and-Learn Bootstrapping
Automatically derived Biomedical Ontology
Seed Ranker
Web Corpus
Seeds
Categories
Ambiguous
"White is a Gene"
Unambiguous
"Drosomycin is a Gene"
KB
General Seeds
Bootstrapping Seeds
Learned facts
Biomedical Corpus (PubMed)
Coupled Bootstrapping
SEAL
CPL
Morphology

Ontology

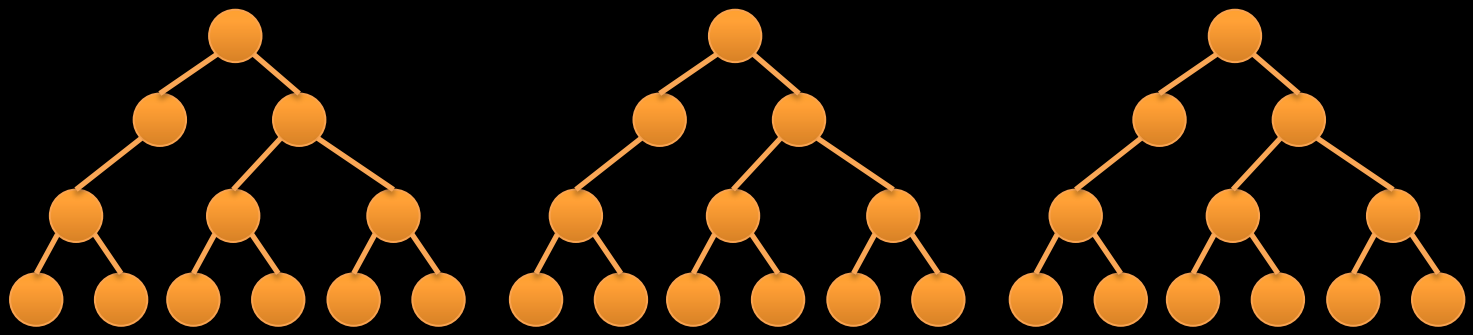
- Based on 6 common ontologies
- Cover a wide range of sub-domains
  - Gene Ontology (GO)
  - NCBI Taxonomy for model organisms
  - Chemical Entities of Biological Interest (ChEBI)
  - Sequence Ontology
  - Cell Type Ontology
  - Human Disease Ontology
- Source ontologies provide term hierarchy

Base Ontology

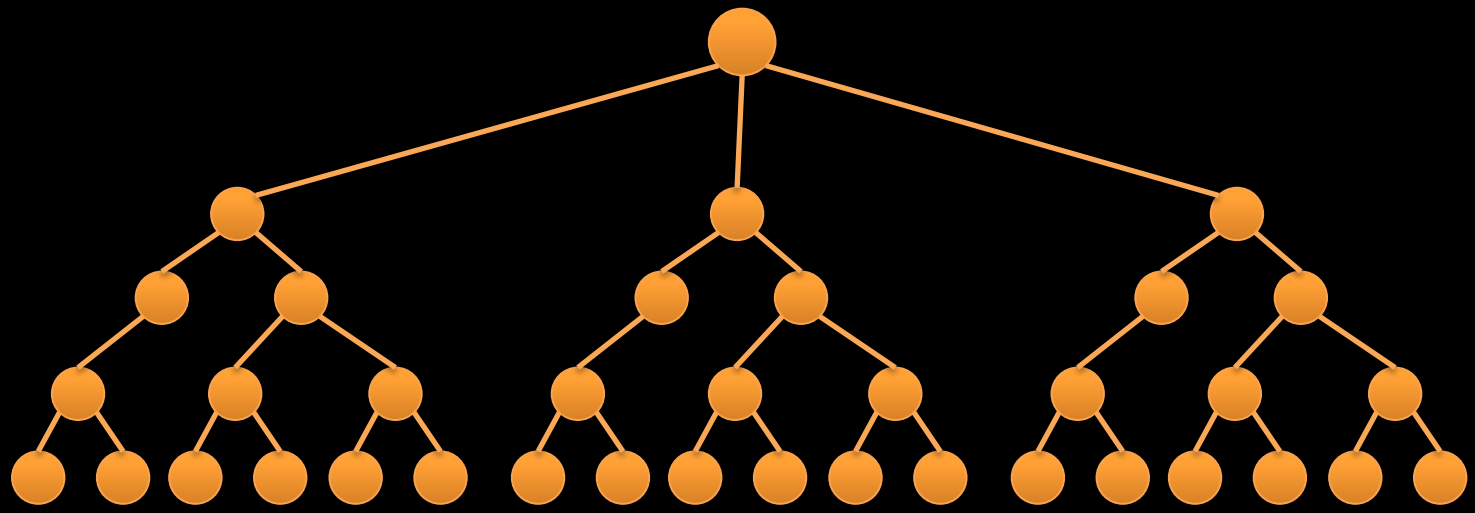


Ontology

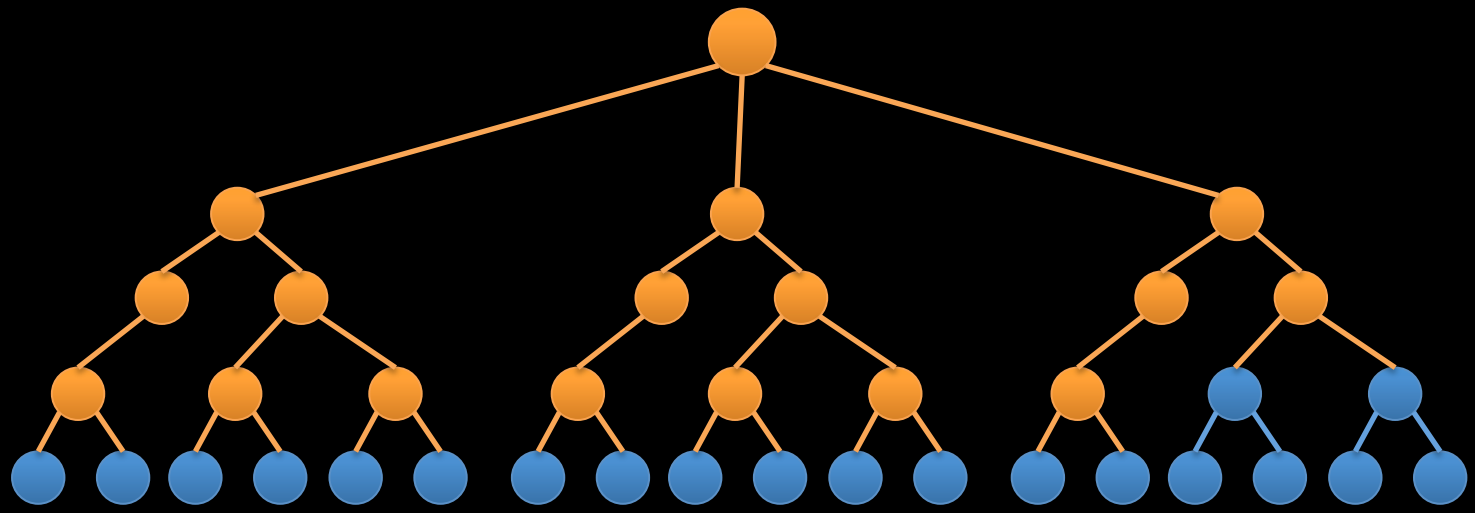
Base Ontology *Base Ontology* *Base Ontology*



Ontology



Ontology



- High level terms -- Categories/concepts
- Specific terms – Seeds/examples

- 109 categories

Ontology Stats

- Full tree: over 1 million terms
  - 856 K terms
  - 154 K synonyms

- In this study:
  - 109 categories (20 high-level terms from each ontology)
  - This leaves over 1 M seeds!

Seed Set Refinement

- Based on collocation of seed and a target category
- Using Pointwise Mutual Information

Seed = "white"
Category = "Gene"
D = document corpus (Web)
D(cat) = documents that mention *Category*

$$\text{PMI}(\textit{Seed}, \textit{Category}) \propto \frac{|\ \text{Occurrence}(\textit{Seed}, \textit{D(cat)})\ |}{|\ \text{Occurrence}(\textit{Seed}, \textit{D})\ |}$$

- PMI-Rank("white", "Gene") ≈ 0

Ranking Gene Names

- Ranking *D. Melanogaster* Genes
  - Data taken from the BioCreative Challenge

High PMI-Rank Genes	
SoxN	achaete
Pax-6	Drosomycin
BX-C	Ultrabithorax
D-Fos	sine oculis
Abd-A	dCtBP
PKAc	huckebein

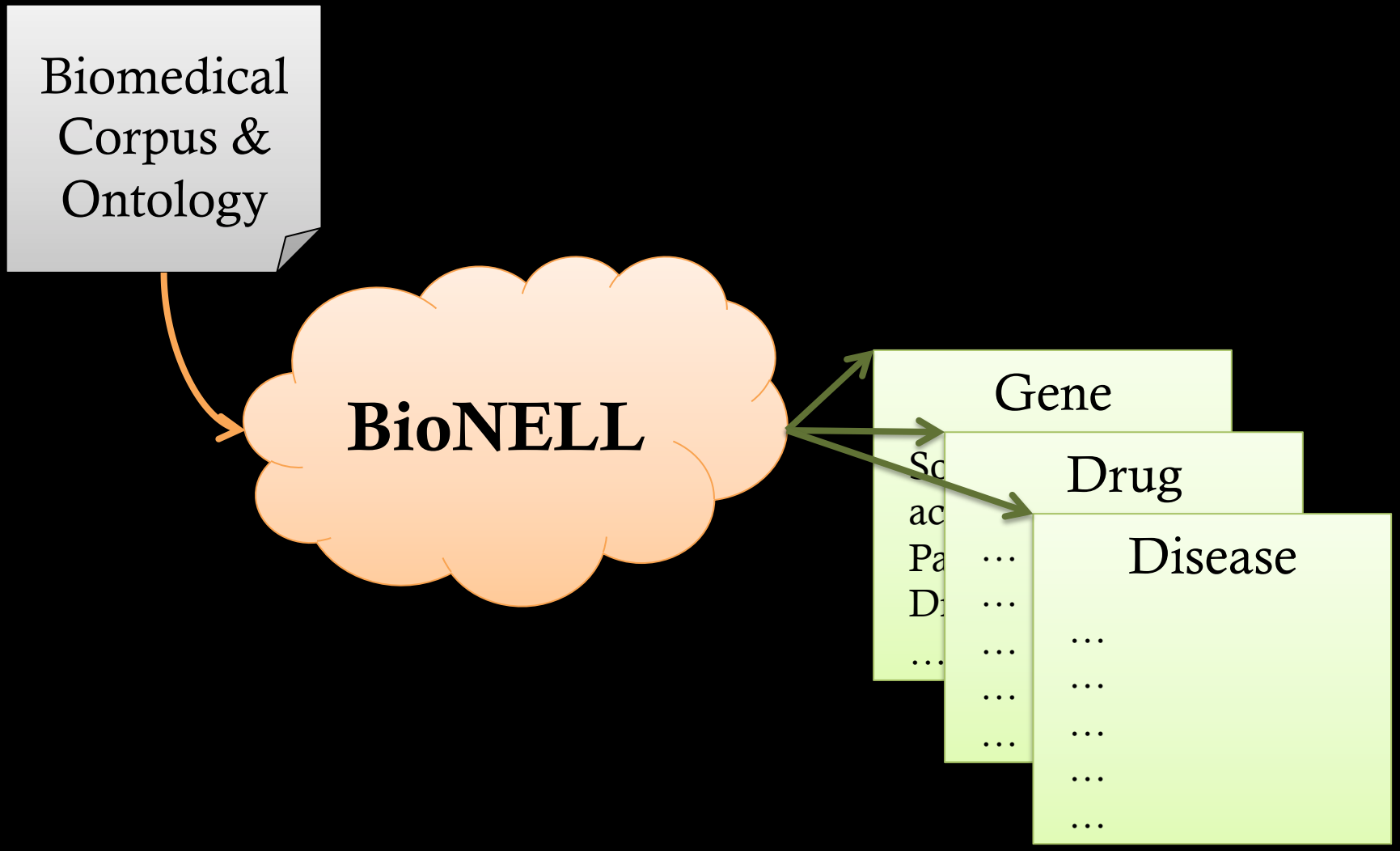
Random Sample	
section 33	crybaby
hv	Bob
ael	LRS
dip	chm
arm	3520

Evaluation

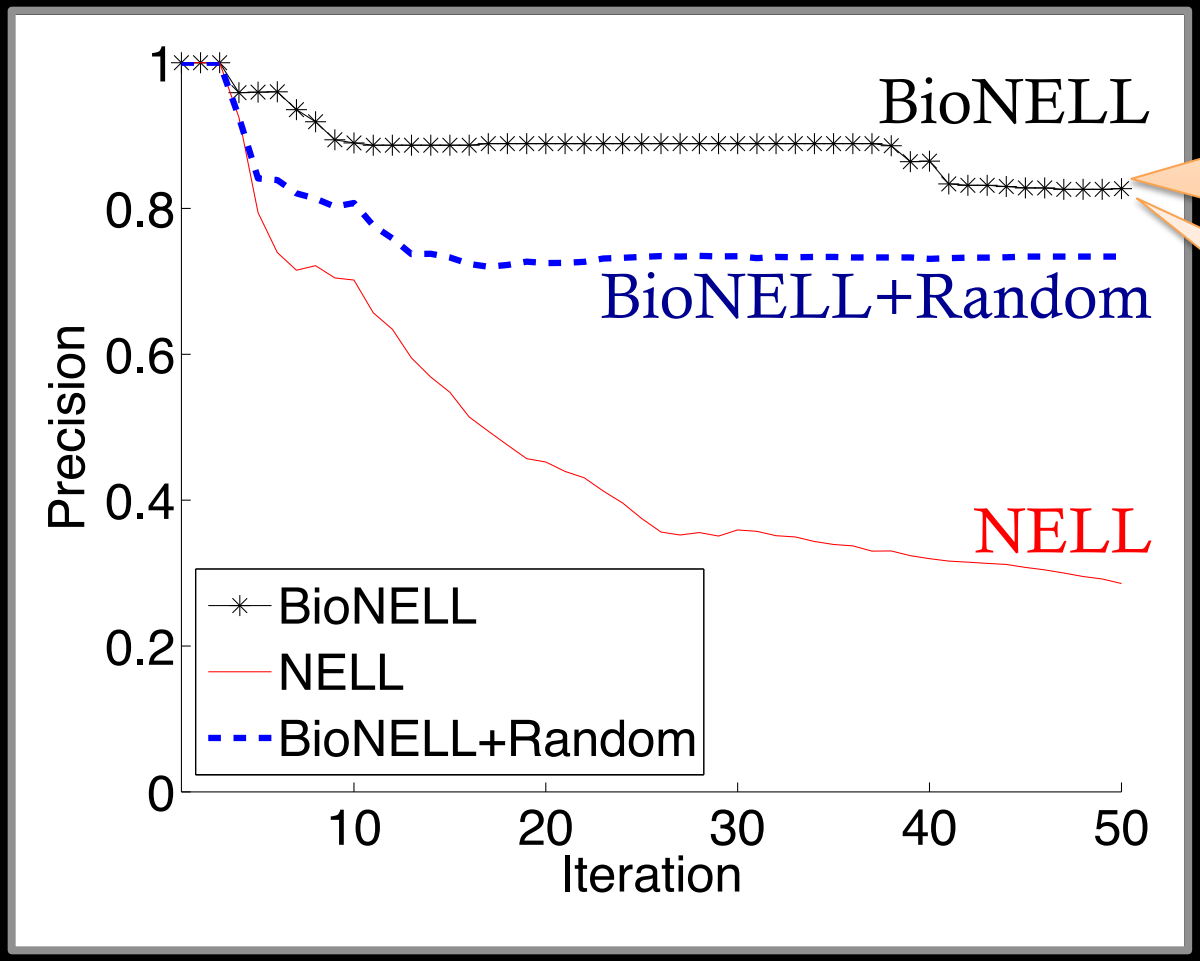
Learning System	Bootstrapping Algorithm	Initial Seeds	Corpus
BioNELL	Rank-and-Learn	PMI	PubMed
NELL	NELL's algo	Random	PubMed
BioNELL+Random	Rank-and-Learn	Random	PubMed

- All tested systems:
  - Run for 50 iterations
  - Use biomedical ontology & corpus
  - 50 initial seeds

Learning Biomedical Lexicons



D. Melanogaster Genes Lexicon



BioNELL has high precision

Precision is high throughout 50 iterations

Recall is low for all systems

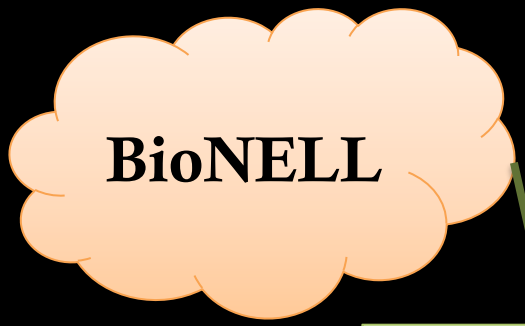
More Biomedical Lexicons

- More Categories:
  - Chemical Component (CC), Disease, Drug

System	Precision			Correct		
	CC	Drug	Disease	CC	Drug	Disease
BioNELL	.66	.52	.43	63	508	276
NELL	.15	.40	.37	74	522	288

- BioNELL has higher precision on all categories
- Recall is comparable

Named Entity Recognition



Genes

SoxN
achaete
Pax-6
Drosomycin
…

PubMed Abstracts

… the evolutionarily ancient role of Pax-6 was to regulate structural genes (e.g., rhodopsin) in …

Ambiguous terms in lexicon ?
"arm"

"arm" not in lexicon

… recessive and cell mutation armadillo (arm), detected by …

"arm" is in lexicon

… on the left arm of the third chromosome …

Named Entity Recognition

- Used learned lexicons for NER in text
- Simple method: string matching

Lexicon	Precision	Correct
BioNELL	.90	18
NELL	.02	5
BioNELL+Random	.03	3

Out of 1616

- BioNELL: Significantly higher precision

BioNELL: Main Advantages

- Automatically derived ontology
- Wide range of biomedical concepts
- Significantly reduces ambiguity in learned lexicons
  - Rank-and-Learn bootstrapping
  - PMI-based seed refinement

dma@cs.cmu.edu
www.cs.cmu.edu/~dmovshov