The Center For Language and Speech Processing
at the Johns Hopkins University

Boosting Biomedical Entity Extraction by Using Syntactic Patterns for Semantic Relation Discovery

Svitlana Volkova, PhD Student, CLSP JHU

Doina Caragea, William H. Hsu, John Drouhard, Landon Fowles

Department of Computing and Information Sciences, K-State

Research supported by: K-State National Agricultural Biosecurity Center (NABC) and the US Department of Defense

Agenda

I. Introduction

II. Related Work
   - Biomedical Entity Extraction
   - Ontology Learning

III. Methodology
   - Step 1: Manual Ontology Construction
   - Step 2: Automated Relationship Extraction
   - Step 3: Automated Ontology Construction
   - Step 4: Biomedical Entity Extraction

IV. Experimental Design and Results

V. Summary

Veterinary Medicine Data Online

Structured Data

- Official reports by different organizations:
  - state and federal laboratories, bioportals;
  - health care providers;
  - governmental agricultural or environmental agencies.

Unstructured Data

- Web-pages
- News

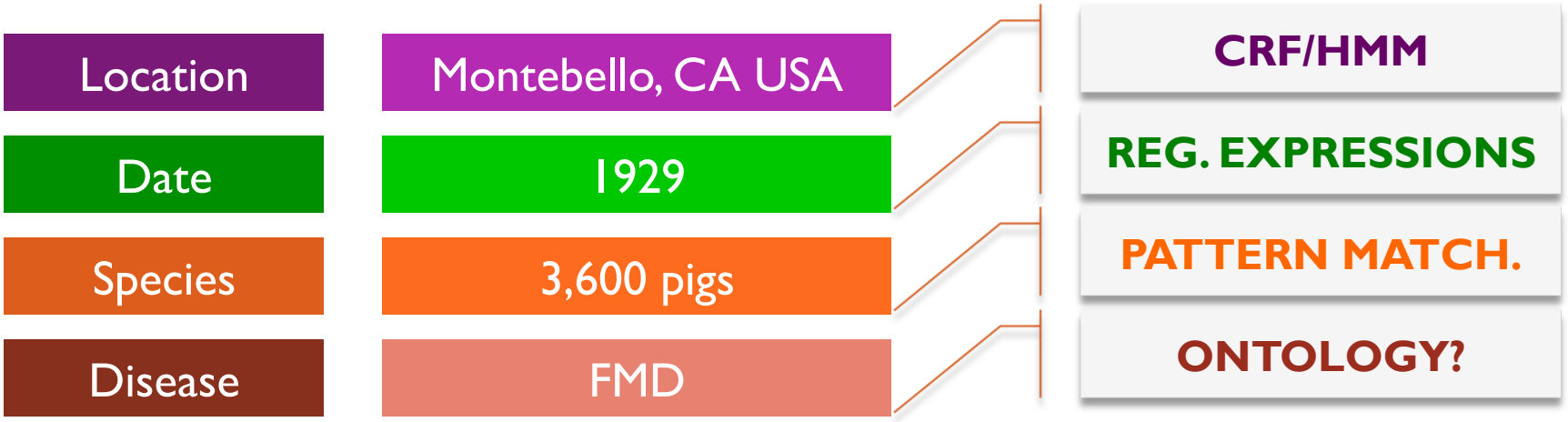
- E-mails (e.g., ProMed-Mail)
- Blogs

- Medical literature (e.g., books)

- Scientific papers (e.g., PubMed)

Entity Extraction + Event Recognition

"The **US** saw its latest **FMD** outbreak in **Montebello, CA** in **1929** where **3,600 pigs** were slaughtered".



- LACK OF ONTOLOGY IN THE DOMAIN IN VETERINARY MEDICINE
- LACK OF LABELED DATA FOR SEQUENCE LABELING

Related Work in Biomedical Entity Extraction

- Methods:
  - dictionary-based bio-entity name recognition in bio-literature
  - protein name recognition using gazetteer
  - gene-disease relation extraction
  - conditional random fields has been applied for identifying gene and protein mentions

- Limitations:
  - based on static dictionaries
  - limits the recall of the system by the size of the dictionary
  - requires annotated training corpora for learning

Emergency Surveillance Systems

- **BioCaster**
  - manually-constructed ontology of 50 animal diseases

- **Pattern-based Understanding & Learning System PULS**
  - a list of 2400 human and animal disease

- **HealthMap**
  - a list of 1100 human and animal disease

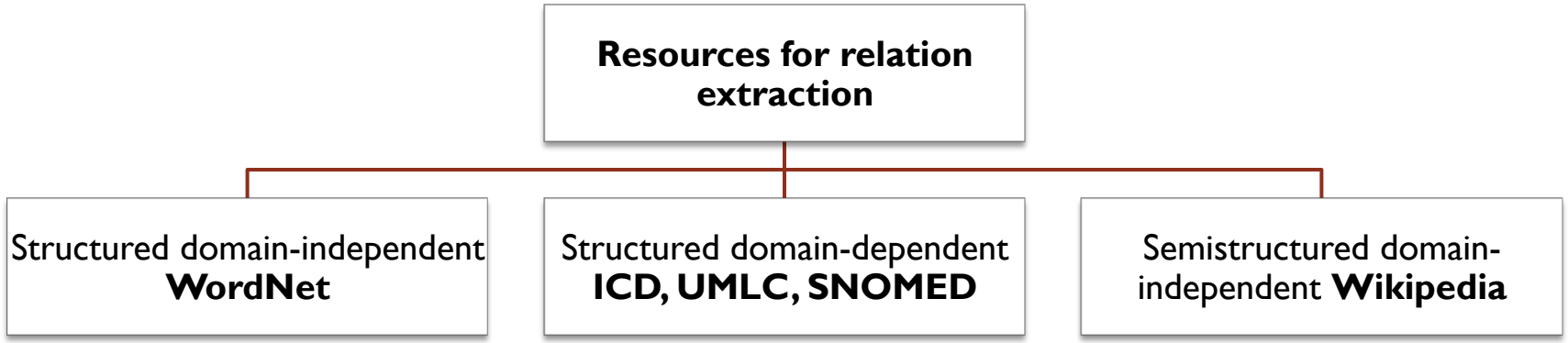
- **Limitations**
  - based on dictionary lookup
  - Do not extract disease synonyms, viruses and serotypes

Automated Ontology Learning for Boosting Biomedical Entity Extraction

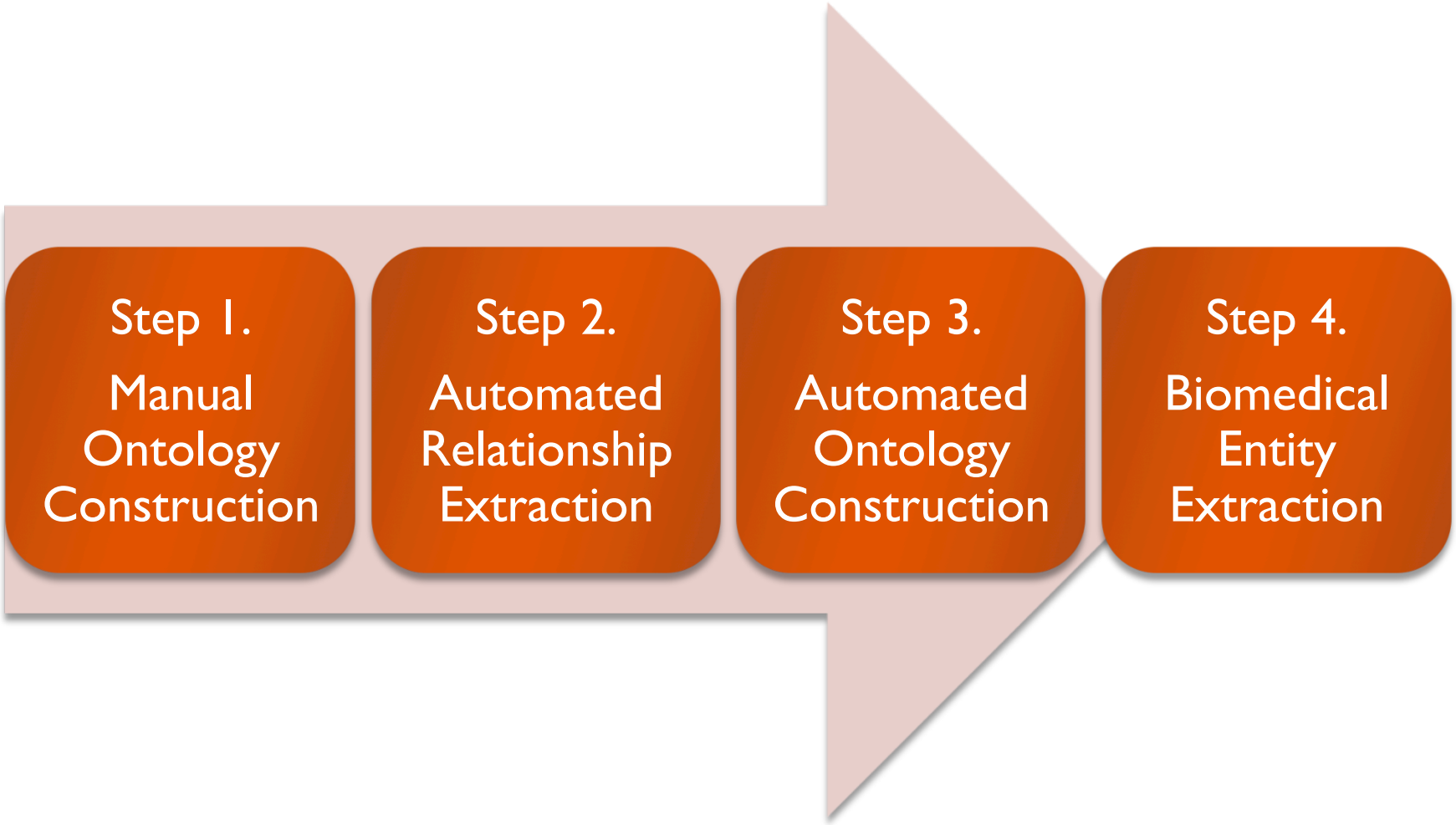
Learn
animal disease ontology
automatically
from web using syntactic
pattern matching
for semantic relation
discovery

Related Work: Relation Extraction

OntoLearn OntoMiner	• extract concepts with taxonomic (synonymic) "is-a" relations
Text-To-Onto Text2Onto	• extract non-taxonomic (hyponymic) relations between concepts
Concept Tuple-based Ontology Learning	• performs full-text parsing using statistical and rule-based syntactic analysis of documents



Methodology



Step1. Manual Ontology Construction

Disease names from Iowa State University Center for Food Security and Public Health (CFSPH)

Word Organization of Animal Health (OIE) Animal Disease Data

Ontology Sources

Department for Environmental Food and Rural Affairs, UK (DEFRA)

Wikipedia

$|O_{INIT}|$ = 429 terms

$|O_{Syn}|$ = 453 terms

$|O_{Abbr}|$ = 581 terms

$|O_{S+A}|$ = 605 terms

Step 2. Automated Relationship Extraction

- **Synonymic relationships –** "E1 is a kind of E2"

  E1 = "swine influenza" is a kind of E2 = "swine fever"

- **Hyponymic relationships –** "E1 and E1 are diseases"

  E1 = "anthrax", E2 = "yellow fever" are diseases

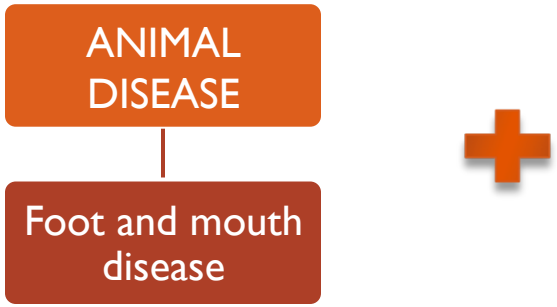
- **Causal relationships –** "E1 is caused by E2"

  E1 = "Ovine epididymitis" is caused by E2 = "Brucella ovis"

Synonymic	Hyponymic	Causal
• "is a", "and" • "also known as" • "is also called "	• "such as" • "for example" • "including"	• "is caused by" • "causes"

Step 3. Automated Ontology Construction Synonymic Relationship "is a kind of"

O_{INIT} = {"foot and mouth disease"}



"Foot-and-mouth disease, FMD or hoof-and-mouth disease (Aphtae epizooticae) is a highly contagious and sometimes fatal viral disease".

O_R = {"foot and mouth disease", "FMD", "hoof-and-mouth disease", "Aphtae epizooticae"}

Step 3. Automated Ontology Construction Causative Relationship "is caused by"

O'_{INIT} = {"foot and mouth disease","FMD", "hoof-and-mouth disease", "Aphtae epizooticae"}

"FMD is caused by foot-and-mouth disease virus (FMDV)"

O_R = {"foot and mouth disease","FMD","hoof-and-mouth disease", "Aphtae epizooticae","foot-and-mouth disease virus","FMDV"}

Step 4. Biomedical Entity Extraction

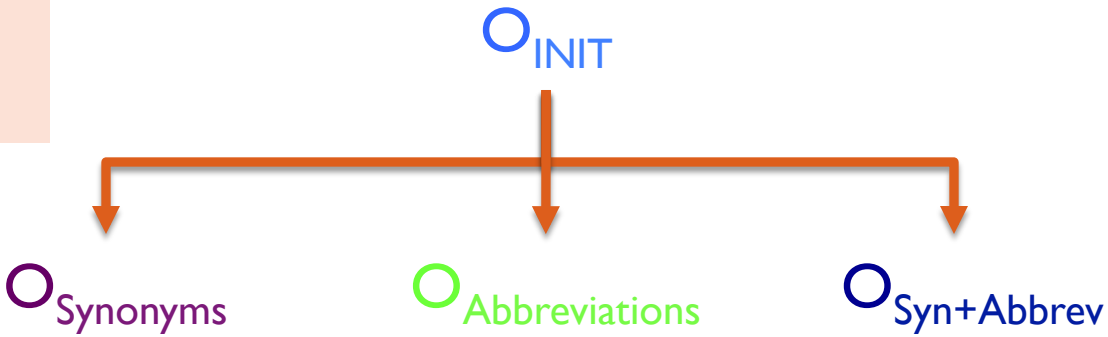
"Species infecting domestic livestock are B. melitensis (goats and sheep, see Brucella melitensis), B. suis (pigs, see Swine brucellosis), B. abortus (cattle and bison), B. ovis (sheep), and B. canis (dogs)"

- Terminology Extraction – "B. melitensis", "B.suis" …
- Segmentation – [43..54], [98..105]
- Association Extraction – e.g. "B. melitensis" is a synonym of "Brucella melitensis"
- Normalization – "Brucellosis" - "B. melitensis" - "B.suis"

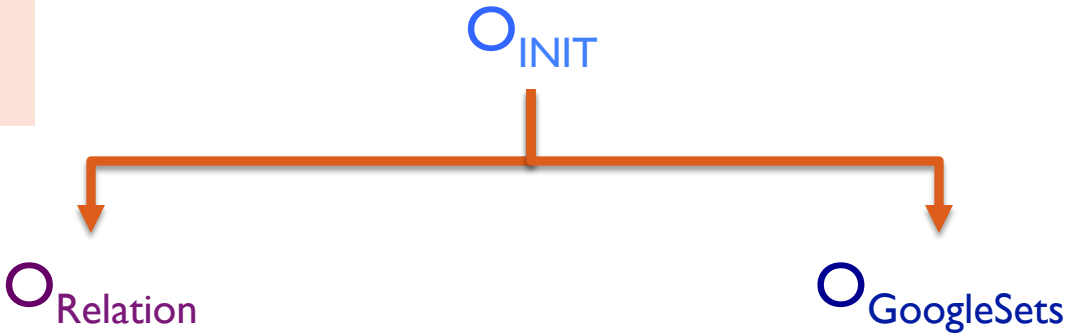
Experiment

- 100 unlabeled documents for ontology expansion - D_{Ont}
- 100 manually labeled document for entity extraction - D_{Ext}

Manually constructed

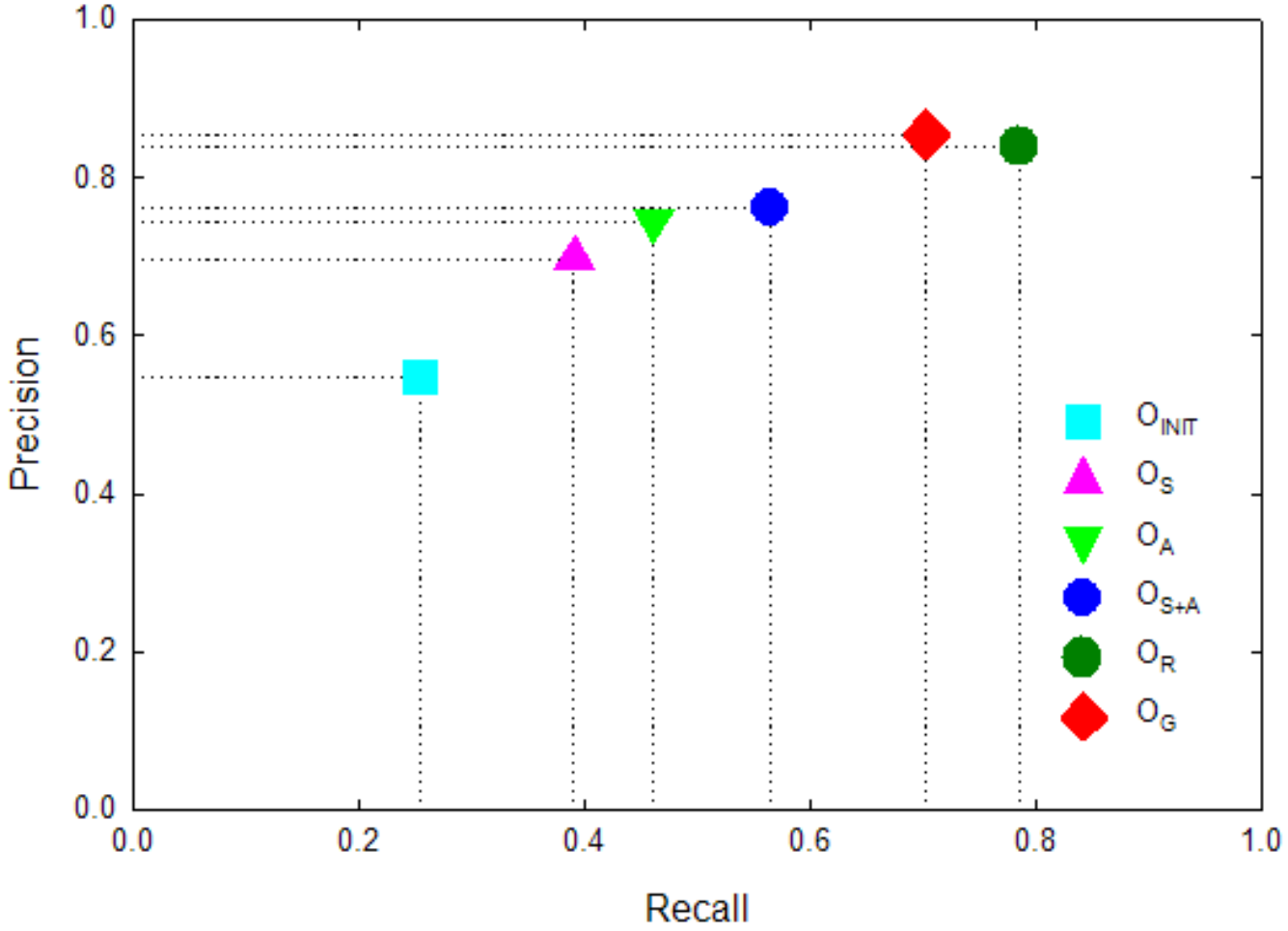


Automatically constructed



Google Sets expansion approach

Entity Extraction Results

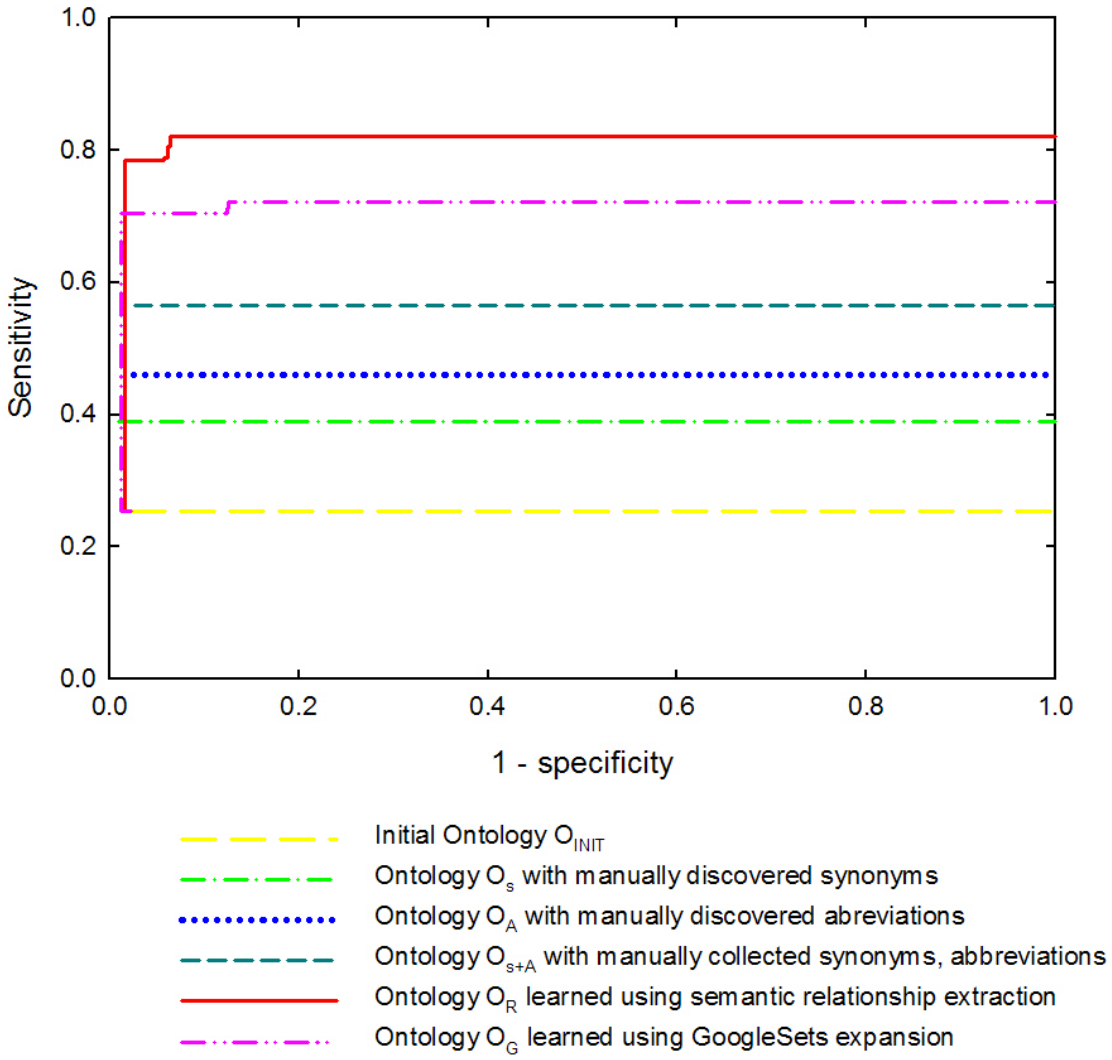


O_{INIT}
Precision – 0.54
Recall – 0.25

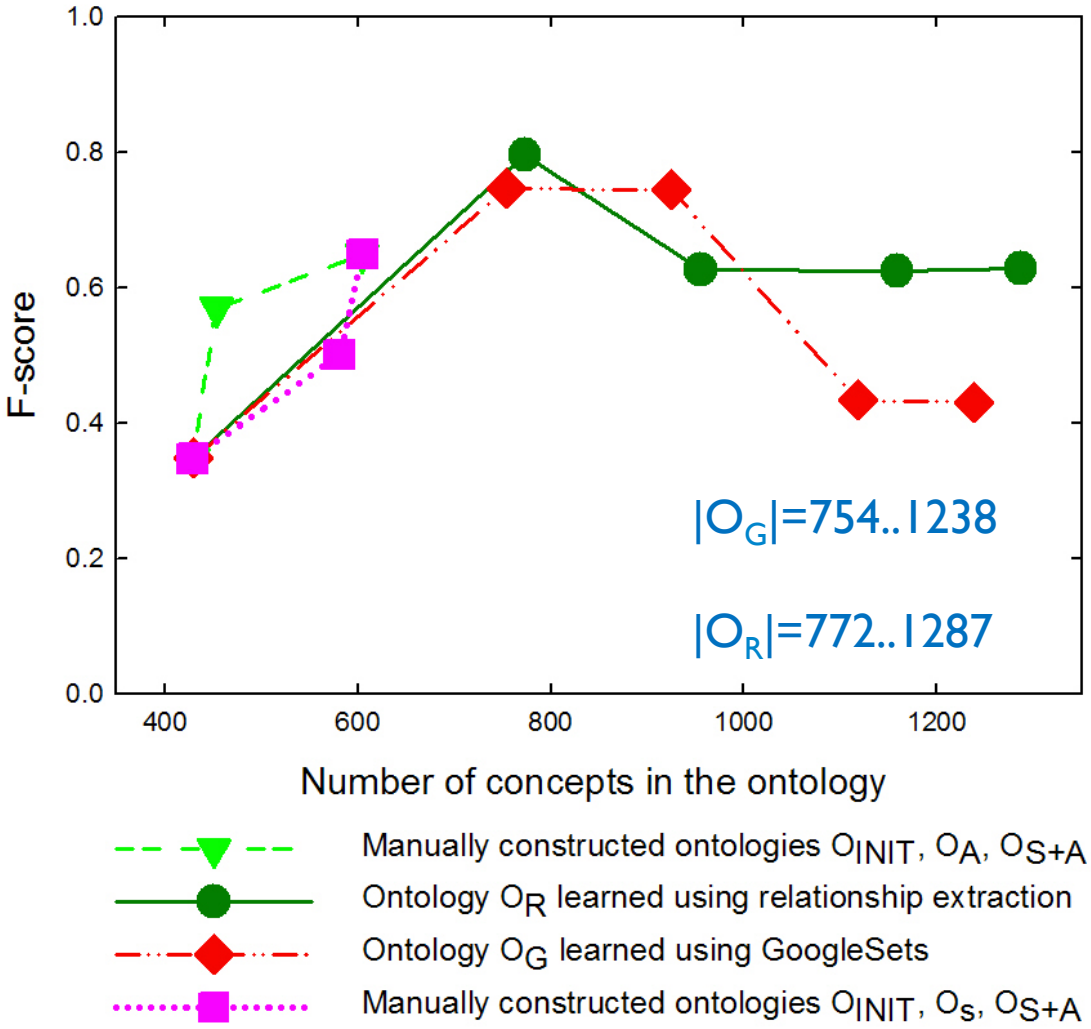
O_G
Precision – 0.85
Recall – 0.71

O_R
Precision – 0.85
Recall – 0.79

Entity Extraction Results: ROC Curves



Entity Extraction Results: Learning Curves



Summary & Future Work

- Our results:
  - O_R – Precision – 84.8, Recall – 78.9 and F-score – 81.7
  - O_G – Precision – 84.7, Recall – 71.3
- BioCaster
  - 200 news articles, F-score – 76.9
- DNA, RNA, cell type extraction
  - SVM and orthographic features, F-score – 66.5
- Biomedical Entity Extraction
  - multilingual ontology construction using Wikipedia
- Automated Ontology Construction
  - generalize for other named entities

Thank you!

Svitlana Volkova, svitlana@jhu.edu

http://people.cis.ksu.edu/~svitlana