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
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

<table>
<thead>
<tr>
<th>Structured Data</th>
<th>Unstructured Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Official reports by different organizations:</td>
<td>- Web-pages</td>
</tr>
<tr>
<td>- state and federal laboratories, bioportals;</td>
<td>- News</td>
</tr>
<tr>
<td>- health care providers;</td>
<td>- E-mails (e.g., ProMed-Mail)</td>
</tr>
<tr>
<td>- governmental agricultural or environmental agencies.</td>
<td>- Blogs</td>
</tr>
<tr>
<td></td>
<td>- Medical literature (e.g., books)</td>
</tr>
<tr>
<td></td>
<td>- Scientific papers (e.g., PubMed)</td>
</tr>
</tbody>
</table>
“The **US** saw its latest **FMD** outbreak in **Montebello, CA** in **1929** where **3,600 pigs** were slaughtered”.

- **Location**: Montebello, CA USA
- **Date**: 1929
- **Species**: 3,600 pigs
- **Disease**: FMD

- **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**
- extract concepts with taxonomic (synonymic) “is-a” relations

**OntoMiner**

**Text-To-Onto**
- extract non-taxonomic (hyponymic) relations between concepts

**Text2Onto**
- performs full-text parsing using statistical and rule-based syntactic analysis of documents

**Concept Tuple-based Ontology Learning**

**Resources for relation extraction**

- Structured domain-independent: WordNet
- Structured domain-dependent: ICD, UMLC, SNOMED

2010 IEEE / WIC / ACM International Conference on Web Intelligence, 31 Aug – 03 Sept 2010, Toronto Canada
Methodology

Step 1. Manual Ontology Construction
Step 2. Automated Relationship Extraction
Step 3. Automated Ontology Construction
Step 4. Biomedical Entity Extraction
Step 1. 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

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

- Wikipedia

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $|O_{INIT}| = 429$ 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”

<table>
<thead>
<tr>
<th>Synonymic</th>
<th>Hyponymic</th>
<th>Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “is a”, “and”</td>
<td>• “such as”</td>
<td>• “is caused by”</td>
</tr>
<tr>
<td>• “also known as”</td>
<td>• &quot;for example&quot;</td>
<td>• “causes”</td>
</tr>
<tr>
<td>• “is also called”</td>
<td>• “including”</td>
<td></td>
</tr>
</tbody>
</table>
Step 3. Automated Ontology Construction

Synonymic Relationship “is a kind of”

\[ O_{INIT} = \{\text{“foot and mouth disease”}\} \]

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”.

\[ O_R = \{\text{“foot and mouth disease”}, \text{“FMD”}, \text{“hoof-and-mouth disease”}, \text{“Aphtae epizooticae”}\} \]
Step 3. Automated Ontology Construction
Causative Relationship “is caused by”

\[ O'_{\text{INIT}} = \{\text{“foot and mouth disease”}, \text{“FMD”}, \text{“hoof-and-mouth disease”}, \text{“Aphtae epizooticae”}\} \]

\[ + \]

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

\[ O_R = \{\text{“foot and mouth disease”}, \text{“FMD”}, \text{“hoof-and-mouth disease”}, \text{“Aphtae epizooticae”}, \text{“foot-and-mouth disease virus”}, \text{“FMDV”}\} \]
Step 4. Biomedical Entity Extraction

“Species infecting domestic livestock are \textit{B. melitensis} (goats and sheep, see \textit{Brucella melitensis}), \textit{B. suis} (pigs, see Swine brucellosis), \textit{B. abortus} (cattle and bison), \textit{B. ovis} (sheep), and \textit{B. canis} (dogs)”

- Terminology Extraction – “\textit{B. melitensis}”, “\textit{B.suis}” …

- Segmentation – [43..54], [98..105]

- Association Extraction – e.g. “\textit{B. melitensis}” is a synonym of “\textit{Brucella melitensis}”

- Normalization – “\textit{Brucellosis}” - “\textit{B. melitensis}” - “\textit{B.suis}”
Experiment

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

[Diagram showing ROC curves for different ontologies, including:
- Initial Ontology $O_{\text{INIT}}$
- Ontology $O_6$ with manually discovered synonyms
- Ontology $O_A$ with manually discovered abbreviations
- Ontology $O_{s+A}$ with manually collected synonyms, abbreviations
- Ontology $O_R$, learned using semantic relationship extraction
- Ontology $O_G$, learned using GoogleSets expansion]
Entity Extraction Results: Learning Curves

![Graph showing learning curves for different ontology sizes.](image)

- $|O_G| = 754..1238$
- $|O_R| = 772..1287$

Legend:
- Manually constructed ontologies $O_{INIT}, O_A, O_{S+A}$
- Ontology $O_R$ learned using relationship extraction
- Ontology $O_G$ learned using GoogleSets
- Manually constructed ontologies $O_{INIT}, O_S, O_{S+A}$
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