Unsupervised Ontology Induction From Text

Hoifung Poon
University of Washington

(Joint work with Pedro Domingos)
Extracting Knowledge From Text

\[ \text{INDUCE}(e1) \land \text{IL-4}(e2) \land \text{CD11B}(e3) \land \text{INDUCE}(e1,e2) \land \text{INDUCED}(e1,e3) \]

\[ \text{REGULATE} \land \text{regulate, control, govern, modulate} \]

\[ \text{INDUCE} \land \text{inhibit, block, suppress, prevent, abolish, abrogate, down-regulate} \]

\[ \text{ACTIVATE} \land \text{activate} \]

\[ \ldots \ldots \]

\[ \ldots \ldots \]
REGULATE

regulate, control, govern, modulate

induce, enhance, trigger, augment, up-regulate

inhibit, block, suppress, prevent, abolish, abrogate, down-regulate

activate

ACTIVATE

INHIBIT

INDUCE
<table>
<thead>
<tr>
<th>Action</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGULATE</td>
<td>regulate, control, govern, modulate</td>
</tr>
<tr>
<td>INDUCE</td>
<td>induce, enhance, trigger, augment, up-regulate</td>
</tr>
<tr>
<td>INHIBIT</td>
<td>inhibit, block, suppress, prevent, abolish, abrogate, down-regulate</td>
</tr>
<tr>
<td>ACTIVATE</td>
<td>activate</td>
</tr>
</tbody>
</table>
Extracting Knowledge From Text
Extracting Knowledge From Text

- **Wanted**: Automatic, end-to-end solution
- Manual engineering: Costly and limited
- Supervised learning
  - Bottleneck: Labeled examples
  - Infeasible for large-scale, open-domain knowledge extraction
Unsupervised Learning for Knowledge Extraction

- TextRunner [Banko et al. 2007]
  - State-of-the-art open information extraction
  - Only extracts triples
  - Extractions are largely unstructured and noisy

- USP [Poon & Domingos 2009]
  - Form complete, detailed meaning representation
  - More robust to noise
  - Still limited to extractions with substantial evidence
  - Lacks ontological structures
Why Ontology?

- Compact representation and efficient reasoning [Staab & Studer 2004]
- Better generalization

Interestingly, the DEX-mediated IkappaBalpha induction was completely inhibited by IL-2, but not IL-4, in Th1 cells, while the reverse profile was seen in Th2 cells.

Q: What does IL-2 regulate?

A: The DEX-mediated IkappaBalpha induction
Ontology Learning


- **Induction:** Construct an ontology

- **Population:** Map textual expressions to concepts and relations in the ontology

- Limitations in existing approaches
  - Require heuristic patterns or existing KBs
  - Pursue each task in isolation

Knowledge representation

NLP
This Talk: OntoUSP

- Jointly conducts: Ontology induction, population, and knowledge extraction
- Learns ISA hierarchy over logical expressions
- Extends USP with hierarchical clustering
- Hierarchical smoothing encoded in a few high-order formulas in Markov Logic [Richardson & Domingos, 2006]
- Sole input is dependency trees

Five times as many correct answers as TextRunner

Improves on the recall of USP by 47%
Outline

- **Background: USP**
- Unsupervised ontology induction
- Conclusion
IL-4 protein induces CD11b

Structured prediction: Partition + Assignment
Challenge: Same Meaning, Many Variations

IL-4 induces CD11b
Protein IL-4 enhances the expression of CD11b
CD11b expression is induced by IL-4 protein
The cytokin interleukin-4 induces CD11b expression
IL-4’s up-regulation of CD11b, ...

......
Unsupervised Semantic Parsing

**USP** = Recursively cluster *arbitrary* expressions composed with / by similar expressions

- *IL-4 induces CD11b*
- *Protein IL-4 enhances the expression of CD11b*
- *CD11b expression is enhanced by IL-4 protein*
- *The cytokin interleukin-4 induces CD11b expression*
- *IL-4’s up-regulation of CD11b,* ...
Unsupervised Semantic Parsing

- **USP = Recursively** cluster arbitrary expressions composed with / by similar expressions

  *IL-4 induces CD11b*
  
  *Protein IL-4 enhances the expression of CD11b*
  
  *CD11b expression is enhanced by IL-4 protein*
  
  *The cytokin interleukin-4 induces CD11b expression*
  
  *IL-4’s up-regulation of CD11b, ...*

Cluster same forms at the atom level
Unsupervised Semantic Parsing

- USP = *Recursively* cluster arbitrary expressions composed with / by similar expressions

*IL-4 induces* CD11b

*Protein IL-4 enhances* the expression of CD11b

*CD11b expression is enhanced by* IL-4 protein

*The cytokin interleukin-4 induces* CD11b expression

*IL-4’s up-regulation of* CD11b, ...

Cluster forms in composition with same forms
Unsupervised Semantic Parsing

- **USP** = **Recursively** cluster **arbitrary** expressions composed with / by similar expressions

\[
\text{IL-4 induces CD11b}
\]

\[
\text{Protein IL-4 enhances the expression of CD11b}
\]

\[
\text{CD11b expression is enhanced by IL-4 protein}
\]

\[
\text{The cytokin interleukin-4 induces CD11b expression}
\]

\[
\text{IL-4’s up-regulation of CD11b, ...}
\]

Cluster forms in composition with same forms
Unsupervised Semantic Parsing

- USP = **Recursively** cluster **arbitrary** expressions composed with / by similar expressions

  *IL-4 induces CD11b*

  *Protein IL-4 enhances* the expression of CD11b

  *CD11b expression is enhanced by IL-4 protein*

  *The cytokin interleukin-4 induces CD11b expression*

  *IL-4’s up-regulation of CD11b, ...*

Cluster forms in composition with same forms
Unsupervised Semantic Parsing

- **USP = Recursively** cluster *arbitrary* expressions composed with / by similar expressions

- *IL-4 induces CD11b*
  - *Protein IL-4 enhances the expression of CD11b*
  - *CD11b expression is enhanced by IL-4 protein*
  - *The cytokin interleukin-4 induces CD11b expression*
  - *IL-4’s up-regulation of CD11b, ...*

Cluster forms in composition with same forms
Probabilistic Model for USP

- Joint probability distribution over input dependency trees and their semantic parses

- **Use Markov logic**

- A Markov Logic Network (MLN) is a set of pairs \((F_i, w_i)\) where
  - \(F_i\) is a formula in higher-order logic
  - \(w_i\) is a real number

\[
P(x) = \frac{1}{Z} \exp \left( \sum_i w_i \cdot N_i(x) \right)
\]
Unsupervised Semantic Parsing

- Exponential prior on number of parameters
- Cluster mixtures:
  \[ \text{InClust}(e,+c) \wedge \text{HasValue}(e,+v) \]

**Object/Event Cluster:** INDUCE

- **Induces:** 0.1
- **Enhances:** 0.4

**Property Cluster:** INDUCER

- nsubj: 0.5
- Agent: 0.4
- IL-4: 0.2
- IL-8: 0.1
- None: 0.1
- One: 0.8
Inference: Hill-Climb Probability

Initialize

Search Operator

Lambda reduction
Learning: Hill-Climb Likelihood

Initialize

<table>
<thead>
<tr>
<th>induces</th>
<th>enhances</th>
<th>IL-4</th>
<th>protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Search Operator

MERGE

<table>
<thead>
<tr>
<th>induces</th>
<th>enhances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

COMPOSE

<table>
<thead>
<tr>
<th>IL-4</th>
<th>protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

IL-4 protein 1
Outline

- Background: USP
- **Unsupervised ontology induction**
- Conclusion
OntoUSP

- USP + Hierarchical clustering + Shrinkage
- Modify the cluster mixture formula
  \[ \text{InClust}(e,c) \land \text{ISA}(c,d) \land \text{HasValue}(e,v) \]
New Operator: Abstraction

Captures substantial similarities
Experiments

- **Evaluate on an end task**: Question answering
  Applied OntoUSP to extract knowledge from text and answer questions

- **Evaluation**: Number of answers and accuracy

- **GENIA dataset**: 1999 Pubmed abstracts

- **Questions**
  - Use simple questions in this paper, e.g.:
    - *What does anti-STAT1 inhibit?*
    - *What regulates MIP-1 alpha?*
  - Sample 2000 questions according to frequency
Total vs. Correct Answers

Implements recall over USP by 47%.

Five times as many correct answers as TextRunner.

Highest accuracy of 91%.
Induced Ontology (Partial)

REGULATE

regulate, control, govern, modulate

ISA

induce, enhance, trigger, augment, up-regulate

ISA

inhibit, block, suppress, prevent, abolish, abrogate, down-regulate

ISA

activate

INHIBIT

INDUCE

ACTIVATE
Question-Answer: Example

Q: What does IL-2 control?
A: The DEX-mediated IkappaBalpha induction

Sentence:

Interestingly, the DEX-mediated IkappaBalpha induction was completely inhibited by IL-2, but not IL-4, in Th1 cells, while the reverse profile was seen in Th2 cells.
Conclusion

- **OntoUSP**: Unsupervised ontology induction
- **USP + hierarchical clustering / smoothing**
- Jointly conducts ontology induction, population, and knowledge extraction
- **See you at poster 46**