Unsupervised Semantic Parsing

Hoifung Poon
Microsoft Research

(Joint work with Pedro Domingos)
Overview

- Motivation
- Semantic parsing
- USP: Unsupervised Semantic Parsing
- Research directions
“Drowning in Information, Starved for Knowledge”
Example: Biomedical Research

- PubMed contains 20 millions abstracts
- Adds 2000 - 4000 every day
Many More …

- Finance
- Legal
- Patent
- News
- Etc.
Semantic Parsing

\[ \text{INDUCE}(e_1) \land \text{IL-4}(e_2) \land \text{CD11B}(e_3) \land \text{INDUCER}(e_1, e_2) \land \text{INDUCED}(e_1, e_3) \]

Text $\rightarrow$ Knowledge
Applications

- Natural language interfaces to database
- Knowledge extraction from
  - Wikipedia: 2 million articles
  - PubMed: 20 million biomedical abstracts
  - Web: Unlimited amount of information
- Machine reading: Learning by reading
- Question answering
- Help solve AI
Example: Literature-Based Discovery

Propose new hypotheses by assembling knowledge across subfields \[\text{[Swanson & Smalheiser, 1997]}\]

Semantic parsing can revolutionize literature-based discovery
This Talk

- **USP:** First unsupervised approach for semantic parsing
- Read text, extract knowledge, answer questions, all without any labeled examples
- Substantially outperformed state of the art
  - Extracted five times as many correct answers
  - Raised accuracy from below 60% to 90%
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 control?
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 control?
A: The DEX-mediated IkappaBalpha induction

- **REGULATE**
  - regulate, control, govern, modulate
  - ISA
  - ISA
  - ISA
- **INDUCE**
  - induce, enhance, trigger, augment, up-regulate
- **INHIBIT**
  - inhibit, block, suppress, prevent, abolish, abrogate, down-regulate
- **ACTIVATE**
  - activate
Overview

- Motivation
- **Semantic parsing**
- USP: Unsupervised Semantic Parsing
- Research directions
Semantic Parsing: Goal

Microsoft buys Skype

\[ \text{BUY}(n_1) \land \text{MICROSOFT}(n_2) \land \text{SKYPE}(n_3) \land \text{BUYER}(n_1, n_2) \land \text{BOUGHT}(n_1, n_3) \]
Input: Dependency Parse

buys

nsubj  dobj

Microsoft  Skype
Generate Quasi-Logical Form

Convert each node into a unary atom

buys

nsubj
Microsoft
dobj
Skype
Generate Quasi-Logical Form

buys (n₁)

nsubj
Microsoft (n₂)
dobj
Skype (n₃)
Generate Quasi-Logical Form

Convert each edge into a binary atom
Generate Quasi-Logical Form

\[
\begin{align*}
\text{buys} (n_1) \\
\text{nsubj}(n_1, n_2) & \quad \text{dobj}(n_1, n_3) \\
\text{Microsoft} (n_2) & \quad \text{Skype} (n_3)
\end{align*}
\]

Convert each edge into a binary atom
Decompose into Subformulas

buys \( (n_1) \)

nsubj \((n_1, n_2)\)  
Microsoft \((n_2)\)

dobj \((n_1, n_3)\)  
Skype \((n_3)\)
Subformula $\Rightarrow$ Lambda Form

$\text{buys}(n_1)$

\[ \text{nsubj}(n_1,n_2) \quad \text{dobj}(n_1,n_3) \]

Replace constants not in unary atom with lambda variables
Subformula \implies \text{Lambda Form}

\text{buys}(n_1)

\lambda x_2.\text{nsubj}(n_1, x_2) \quad \lambda x_3.\text{dobj}(n_1, x_3)

Replace constants not in unary atom with lambda variables
Lambda Form

- Remember how to combine with arguments
- Denote a function

\[ \lambda x_2. \text{nsubj}(n_1,x_2) = f(x_2) \]

**Input:** Microsoft\( (n_2) \)

**Output:** \( \text{nsubj}(n_1,n_2) \land \text{Microsoft}(n_2) \)

- Known as lambda reduction
Core And Arguments

\[ \lambda x_2. \text{nsubj}(n_1, x_2) \quad \lambda x_3. \text{dobj}(n_1, x_3) \]

Core form: No lambda variable
Argument form: One lambda variable
Core And Arguments

Core form: No lambda variable
Argument form: One lambda variable

\[ \text{Argument form: } \lambda x_2. \text{nsubj} (n_1, x_2) \]

\[ \text{Argument form: } \lambda x_3. \text{dobj} (n_1, x_3) \]
Semantic Grammar

buys(n₁)

λx₂.nsubj(n₁,x₂)

λx₂.BUYER(n₁,x₂)

λx₃.dobj(n₁,x₃)

λx₃.BOUGHT(n₁,x₃)

Microsoft(n₂)

MICROSOFT(n₂)

Skype(n₃)

SKYPE(n₃)
Translate into Canonical Form

\[ \text{BUY} (n_1) \]

\[ \lambda x_2 . \text{BUYER} (n_1, x_2) \quad \lambda x_3 . \text{BOUGHT} (n_1, x_3) \]

\[ \text{MICROSOFT} (n_2) \quad \text{SKYPE} (n_3) \]
Translate into Canonical Form

\[
\text{BUY} (n_1) \\
\text{BUYER} (n_1, n_2) \quad \text{BOUGHT} (n_1, n_3) \\
\text{MICROSOFT} (n_2) \quad \text{SKYPE} (n_3)
\]
Challenge: Structured Prediction

buys(n₁)

nsubj(n₁,n₂)  dobj(n₁,n₃)

Microsoft(n₂)  Skype(n₃)
Challenge: Structured Prediction

+nsubj(n₁,n₂) +dobj(n₁,n₃)

Microsoft(n₂) Skype(n₃)

buys(n₁)
Challenge: Structured Prediction

```
[buys(n₁)]

nsubj(n₁,n₂)   dobj(n₁,n₃)

Microsoft(n₂)   Skype(n₃)
```
Challenge: Many Variations

E.g., Microsoft buys Skype
Microsoft acquires the VoIP company Skype
Skype is acquired by Microsoft Corporation
The Redmond software giant buys Skype
Microsoft’s purchase of Skype, ...

......
Challenge: Many Variations

\[ \text{buys}(n_1) \quad \rightarrow \quad \text{BUY}(n_1) \]
\[ \text{acquires}(n_1) \quad \rightarrow \quad \text{BUY}(n_1) \]
\[ \text{purchases}(n_1) \quad \rightarrow \quad \text{BUY}(n_1) \]
\[ \vdots \]
\[ \lambda x_2. \text{nsubj}(n_1, x_2) \quad \rightarrow \quad \lambda x_2. \text{BUYER}(n_1, x_2) \]
\[ \lambda x_2. \text{agent}(n_1, x_2) \quad \rightarrow \quad \lambda x_2. \text{BUYER}(n_1, x_2) \]
\[ \vdots \]
Supervised Learning

- **Examples:**
  - Zelle & Mooney [1993]
  - Wong & Mooney [2007]
  - Lu et al. [2008]
  - Ge & Mooney [2009]

- Applicable to restricted domains only

- For general text
  - Not clear what predicates and objects to use
  - Hard to produce sufficient labeled examples
Overview

- Motivation
- Semantic parsing
- **USP: Unsupervised Semantic Parsing**
- Research directions
Unsupervised Semantic Parsing

- **USP**: First unsupervised approach


- **OntoUSP = USP + Ontology Induction**


- Encoded in a few Markov logic formula
USP: Key Idea # 1

Target predicates and objects are clusters of linguistic variations for the same meaning

**BUY**

= \{buys, acquires, ’s purchase of, ...\}

= Cluster of various expressions for acquisition

**MICROSOFT**

= \{Microsoft, the Redmond software giant, ...\}

= Cluster of various mentions of Microsoft
USP: Key Idea # 2

- **USP** = *Recursive relational clustering*
- Recursively cluster arbitrary expressions composed with or by similar expressions

*Microsoft buys Skype*

*Microsoft acquires the VoIP company Skype*

*Skype is acquired by Microsoft Corporation*

*The Redmond software giant buys Skype*

*Microsoft’s purchase of Skype, ...*
USP: Key Idea # 2

- **USP = Recursive relational clustering**
- Recursively cluster arbitrary expressions composed with or by similar expressions

  *Microsoft buys Skype*

  *Microsoft acquires the VoIP company Skype*

  *Skype is acquired by Microsoft Corporation*

  *The Redmond software giant buys Skype*

  *Microsoft’s purchase of Skype, ...*

Cluster same forms at the atom level
USP: Key Idea # 2

- **USP** = Recursive relational clustering
- Recursively cluster arbitrary expressions composed with or by similar expressions

*Microsoft buys Skype*

*Microsoft acquires the VoIP company Skype*

*Skype is acquired by Microsoft Corporation*

*The Redmond software giant buys Skype*

*Microsoft’s purchase of Skype, ...*

Cluster forms in composition with same forms
**USP: Key Idea # 2**

- **USP = Recursive relational clustering**
- Recursively cluster arbitrary expressions composed with or by similar expressions

*Microsoft buys Skype*

*Microsoft acquires the VoIP company Skype*

*Skype is acquired by Microsoft Corporation*

*The Redmond software giant buys Skype*

*Microsoft’s purchase of Skype, ...*

Cluster forms in composition with same forms
USP: Key Idea # 2

- **USP** = Recursive relational clustering
- Recursively cluster arbitrary expressions composed with or by similar expressions

**Microsoft buys Skype**

**Microsoft acquires the VoIP company Skype**

**Skype is acquired by Microsoft Corporation**

**The Redmond software giant buys Skype**

**Microsoft’s purchase of Skype, ...**

Cluster forms in composition with same forms
USP: Key Idea # 3

- **Start directly from syntactic analyses**
- Focus on translating them to semantics
- Leverage rapid progress in syntactic parsing
- Much easier than learning both
Probabilistic Model

- Exponential prior on number of parameters
- Cluster mixtures:

Cluster: **BUY**

- **buys** 0.1
- **acquires** 0.4

Argument Type: **BUYER**

- **nsubj** 0.5
- **agent** 0.4
- **MICROSOFT** 0.2
- **GOOGLE** 0.1
- **Zero** 0.1
- **One** 0.8
### Cluster: BUY

<table>
<thead>
<tr>
<th>Form</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>buys($n_1$)</td>
<td>0.1</td>
</tr>
<tr>
<td>acquires($n_1$)</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Distribution over core forms
Argument Type: BUYER

\[ \lambda x_2.\text{nsubj}(n_1, x_2) \quad 0.5 \]
\[ \lambda x_2.\text{agent}(n_1, x_2) \quad 0.4 \]

Distributions over argument forms, clusters, and number

MICROSOFT 0.2
GOOGLE 0.1
Zero 0.1
One 0.8
ISA → Can Inherit Parameters

REGULATE

... REGULATED ...

INDUCE

...

INHIBIT

...
ISA → Can Inherit Parameters
Inference and Learning

- **Challenge:** State space too large
- **But:** Meaning units and clusters often small
  ⇒ **Use combinatorial search**
Inference: Most Probable Parse

induces

protein

IL-4

CD11B
Inference: Most Probable Parse

Initialize by decomposing into atomic formulas
Inference: Most Probable Parse

Induces protein IL-4

Search Operator: Lambda reduction
Inference: Most Probable Parse

- Protein
- IL-4
- CD11B
- Induces

Search Operator: Lambda reduction
induces protein IL-4

C_1' nsubj protein IL-4

C_2' nn IL-4

C_1' induces CD11B

C_2' CD11B

A_1' nsubj protein IL-4

C_2' nn IL-4

A_1' induces CD11B

C_1' nsubj protein IL-4

C_3' CD11B

A_2' dobj CD11B

A_3' C_3' CD11B

C_1' induces CD11B

A_2' dobj CD11B

A_3' C_3' CD11B

execute the one with highest gain in probability
induces protein CD11B

Inference: Most Probable Parse

Stop when no improvement is possible
Learning: Maximize Posterior

Initialize

- induces 56
- enhances 37
- amino 62
- acid 45

......

Search Operators

- MERGE
- COMPOSE
- ABSTRACT
COMPOSE

amino 62 \rightarrow acid 45

amino 39 \rightarrow acid 22 \rightarrow amino acid 23
ABSTRACT

Captures substantial similarities
Evaluation

- No gold annotation of semantic parses
- End-to-end evaluation:
  Extract knowledge and answer questions
Dataset

- **GENIA**: 1999 PubMed abstracts
- Generate questions by sampling
- Used factoid questions, e.g.:
  - *What does anti-STAT1 inhibit?*
  - *What regulates MIP-1 alpha?*
Systems

- **KW-SYN**: Baseline by keyword matching
- **DIRT** [Lin & Pantel, 2001]
  Resolves synonymous binary relations
- **TextRunner** [Banko et al., 2008]
  State-of-the-art unsupervised extraction
- **RESOLVER** [Yates & Etzioni, 2009]
  Relational clustering on TextRunner output
Total and Correct Answers

KW-SYN  TextRunner  RESOLVER  DIRT  USP
USP increased recall five-fold and raised precision to 91%
Why Did USP Do Well?

- Resolved many nontrivial variations
- Active vs. passive voices
- Argument forms that mean the same, e.g.,
  
  - expression of $X = X$ expression
  - $X$ stimulates $Y = Y$ is stimulated with $X$
- Synonymous expressions
- Etc.
Clusters and Compositions

- Clusters in core forms
  - Investigate, examine, evaluate, analyze, study, assay
  - Diminish, reduce, decrease, attenuate
  - Dramatically, substantially, significantly
  - Synthesis, production, secretion, release

- Compositions
  - Amino acid, T cell, immune response, transcription factor, initiation site, binding site …
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 control?
A: The DEX-mediated IkappaBalpha induction

REGULATE
- regulate, control, govern, modulate

INDUCE
- induce, enhance, trigger, augment, up-regulate

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

ACTIVATE
- activate
Related Work

- Bayesian model for USP
  [Titov & Klementiev 2011, 2012]

- Distant supervision [Mintz et al. 2009]

- Grounded language learning
  - Semantic parsing of questions to database
  - Known schema and example question-answers
  - E.g., Clark et al. [2010], Liang et al. [2011]
Limitations of USP

- Greedy approach in learning and inference
- Not yet scalable to the Web
- Need to handle temporal, quantifiers, etc.
- Need to incorporate available knowledge
Overview

- Motivation
- Semantic parsing
- USP: Unsupervised Semantic Parsing
- Research directions
Learning for Efficient Inference

- **Bottleneck**: Inference complexity
- **Solution**: Compactly represent computation by learning a deep network
- Generalize USP for learning
- Preliminary work:

Leverage Knowledge Bases

- **USP + Grounded Learning**
- Bootstrap semantic parser using existing KBs
- Extend KB content by parsing new texts
- Extend KB schema by inducing new relations and attributes
Machine Science

- Automate scientific discovery


- First: Apply USP to pathway-based genome-wide association studies (GWAS)

- How far can we get?
Summary

- **USP**: First approach for unsupervised semantic parsing
- Learn target logical forms by recursively clustering variations of same meaning
- Extract knowledge and answer questions
- Substantially outperforms state of the art
- Exciting research directions