Unsupervised Semantic Parsing

Hoifung Poon
University of Washington

(Joint work with Pedro Domingos)
Outline

- **Motivation**
- Unsupervised semantic parsing
- Learning and inference
- Experimental results
- Conclusion
Semantic Parsing

- Natural language text $\Rightarrow$ Formal and detailed meaning representation (MR)
- Also called **logical form**
- Standard MR language: First-order logic
- E.g.,

  *Microsoft buys Powerset.*
Semantic Parsing

- Natural language text $\Rightarrow$ Formal and detailed meaning representation (MR)
- Also called logical form
- Standard MR language: First-order logic
- E.g.,

$\text{Microsoft buys Powerset.}$

$\text{BUYS(MICROSOFT, POWERSET)}$
Shallow Semantic Processing

- Semantic role labeling
  - Given a relation, identify arguments
  - E.g., agent, theme, instrument

- Information extraction
  - Identify fillers for a fixed relational template
  - E.g., seminar (speaker, location, time)

- **In contrast, semantic parsing is**
  - **Formal:** Supports reasoning and decision making
  - **Detailed:** Obtains far more information
Applications

- Natural language interfaces
- Knowledge extraction from
  - Wikipedia: 2 million articles
  - PubMed: 18 million biomedical abstracts
  - Web: Unlimited amount of information
- Machine reading: Learning by reading
- Question answering
- Help solve AI
Traditional Approaches

- Manually construct a grammar
- **Challenge:** Same meaning can be expressed in many different ways

  *Microsoft buys Powerset*
  *Microsoft acquires semantic search engine Powerset*
  *Powerset is acquired by Microsoft Corporation*
  *The Redmond software giant buys Powerset*
  *Microsoft’s purchase of Powerset, ...*

- Manual encoding of variations?
Supervised Learning

- User provides:
  - Target predicates and objects
  - Example sentences with meaning annotation
- System learns grammar and produces parser
- Examples:
  - Zelle & Mooney [1993]
  - Wong & Mooney [2007]
  - Lu et al. [2008]
  - Ge & Mooney [2009]
Limitations of Supervised Approaches

- Applicable to restricted domains only
- For general text
  - Not clear what predicates and objects to use
  - Hard to produce consistent meaning annotation
  - Crucial to develop unsupervised methods
- Also, often learn both syntax and semantics
  - Fail to leverage advanced syntactic parsers
  - Make semantic parsing harder
Unsupervised Approaches

- For shallow semantic tasks, e.g.:
  - **Open IE**: TextRunner [Banko et al. 2007]
  - **Paraphrases**: DIRT [Lin & Pantel 2001]
  - **Semantic networks**: SNE [Kok & Domingos 2008]
- Show promise of unsupervised methods
- **But … none for semantic parsing**
This Talk: USP

● First unsupervised approach for semantic parsing

Based on Markov Logic [Richardson & Domingos, 2006]
- Sole input is dependency trees
- Can be used in general domains
- Applied it to extract knowledge from biomedical abstracts and answer questions
- Substantially outperforms TextRunner, DIRT

Three times as many correct answers as second best
Outline

- Motivation
- **Unsupervised semantic parsing**
- Learning and inference
- Experimental results
- Conclusion
USP: Key Idea # 1

- Target predicates and objects can be learned
- Viewed as clusters of syntactic or lexical variations of the same meaning

\[ \text{BUYS} (-, -) \]
\[ = \{ \text{buys, acquires, ’s purchase of, } \ldots \} \]
\[ = \text{Cluster of various expressions for acquisition} \]

\[ \text{MICROSOFT} \]
\[ = \{ \text{Microsoft, the Redmond software giant, } \ldots \} \]
\[ = \text{Cluster of various mentions of Microsoft} \]
USP: Key Idea # 2

- **Relational clustering** = Cluster relations with same objects

- **USP** = *Recursively* cluster *arbitrary* expressions with similar subexpressions

  Microsoft buys Powerset
  Microsoft acquires semantic search engine Powerset
  Powerset is acquired by Microsoft Corporation
  The Redmond software giant buys Powerset
  Microsoft’s purchase of Powerset, ...
USP: Key Idea # 2

- **Relational clustering** = Cluster relations with same objects
- **USP** = **Recursively** cluster expressions with similar subexpressions

Microsoft _buys_ Powerset
Microsoft _acquires_ semantic search engine _Powerset_
Powerset _is acquired by_ Microsoft Corporation
The Redmond software giant _buys_ Powerset
Microsoft’s purchase of _Powerset_, ...

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

- **Relational clustering** = Cluster relations with same objects
- **USP** = **Recursively** cluster expressions with similar subexpressions

*Microsoft buys Powerset*
*Microsoft acquires* semantic search engine *Powerset*
*Powerset is acquired by Microsoft Corporation*
*The Redmond software giant buys Powerset*
*Microsoft’s purchase of Powerset, ...*

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

- **Relational clustering** = Cluster relations with same objects
- **USP** = **Recursively** cluster expressions with similar subexpressions

Microsoft buys Powerset
Microsoft acquires semantic search engine Powerset
Powerset is acquired by Microsoft Corporation
The Redmond software giant buys Powerset
Microsoft’s purchase of Powerset, ...

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

- **Relational clustering** = Cluster relations with same objects
- **USP** = *Recursively* cluster expressions with similar subexpressions

*Microsoft buys Powerset*

*Microsoft acquires semantic search engine Powerset*

*Powerset is acquired by Microsoft Corporation*

*The Redmond software giant buys Powerset*

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

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
USP: System Overview

- **Input:** Dependency trees for sentences
- Converts dependency trees into quasi-logical forms (QLFs)
- QLF subformulas have natural lambda forms
- Starts with lambda-form clusters at atom level
- Recursively builds up clusters of larger forms

- **Output:**
  - Probability distribution over lambda-form clusters and their composition
  - MAP semantic parses of sentences
Probabilistic Model for USP

- Joint probability distribution over a set of QLFs 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 first-order logic
  - $w_i$ is a real number

$$P(x) = \frac{1}{Z} \exp \left( \sum_i w_i \cdot N_i(x) \right)$$
Generating Quasi-Logical Forms

Convert each node into an unary atom
Generating Quasi-Logical Forms

buys \( n_1 \)

nsubj

Microsoft \( n_2 \)

doobj

Powerset \( n_3 \)

\( n_1, n_2, n_3 \) are Skolem constants
Generating Quasi-Logical Forms

Convert each edge into a binary atom

```
buy(n₁)

nsubj
Microsoft(n₂)
dobj
Powerset(n₃)
```
Generating Quasi-Logical Forms

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

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

Microsoft(\(n_2\)) Powerset(\(n_3\))

Convert each edge into a binary atom
A Semantic Parse

Partition QLF into subformulas
A Semantic Parse

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

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

Microsoft\((n_2)\) \quad \text{Powerset}(n_3) \]

Subformula \Rightarrow \text{Lambda form:}
 Replace Skolem constant not in unary atom with a unique lambda variable
A Semantic Parse

Subformula $\Rightarrow$ Lambda form:
Replace Skolem constant not in unary atom with a unique lambda variable
A Semantic Parse

Follow Davidsonian Semantics

Core form: No lambda variable
Argument form: One lambda variable
A Semantic Parse

buys(n₁)

λx₂.nsubj(n₁,x₂)  λx₃.dobj(n₁,x₃)

∈ C_BUYS

Microsoft(n₂)

∈ C_MICROSOFT

Powerset(n₃)

∈ C_POWERSET

Assign subformula to lambda-form cluster
Lambda-Form Cluster

One formula in MLN

Learn weights for each pair of cluster and core form

Distribution over core forms
Lambda-Form Cluster

May contain variable number of argument types
Argument Type: \( A_{\text{BUYER}} \)

\[ \lambda x_2. \text{prohibit}(x_1, x_2) \] 0.5

\[ \lambda x_2. \text{agent}(x_1, x_2) \] 0.8

None 0.1

One 0.8

Three MLN formulas

Distributions over argument forms, clusters, and number
USP MLN

- Four simple formulas
- Exponential prior on number of parameters
Abstract Lambda Form

Final logical form is obtained via lambda reduction

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

\[ C_{\text{BUYS}}(n_1) \land \lambda x_2 \cdot A_{\text{BUYER}}(n_1, x_2) \land \lambda x_3 \cdot A_{\text{BOUGHT}}(n_1, x_3) \]
Outline

- Motivation
- Unsupervised semantic parsing
- Learning and inference
- Experimental results
- Conclusion
Learning

- **Observed:** $Q$ (QLFs)
- **Hidden:** $S$ (semantic parses)
- Maximizes log-likelihood of observing the QLFs

$$L_\Theta(Q) = \log \sum_S P_\Theta(Q, S)$$
Use Greedy Search

- Search for $\Theta, S$ to maximize $P_\Theta(Q, S)$
- Same objective as hard EM
- Directly optimize it rather than lower bound
- For fixed $S$, derive optimal $\Theta$ in closed form
- Guaranteed to find a local optimum
Search Operators

- **MERGE($C_1, C_2$):** Merge clusters $C_1, C_2$
  
  E.g.: \{buys\}, \{acquires\} $\Rightarrow$ \{buys, acquires\}

- **COMPOSE($C_1, C_2$):** Create a new cluster resulting from composing lambda forms in $C_1, C_2$
  
  E.g.: \{Microsoft\}, \{Corporation\} $\Rightarrow$ \{Microsoft Corporation\}
USP-Learn

- **Initialization**: Partition = Atoms
- **Greedy step**: Evaluate search operations and execute the one with highest gain in log-likelihood
- **Efficient implementation**: Inverted index, etc.
MAP Semantic Parse

- **Goal**: Given QLF $Q$ and learned $\Theta$, find semantic parse $S$ to maximize $P_{\Theta}(Q, S)$
- Again, use greedy search
Outline

● Motivation
● Unsupervised semantic parsing
● Learning and inference
● **Experimental results**
● Conclusion
Task

- No predefined gold logical forms
- **Evaluate on an end task:** Question answering
- **Applied USP to extract knowledge from text and answer questions**
- **Evaluation:** Number of answers and accuracy
Dataset

- **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
Systems

- Closest match in aim and capability: TextRunner [Banko et al. 2007]
- Also compared with:
  - Baseline by keyword matching and syntax
  - RESOLVER [Yates and Etzioni 2009]
  - DIRT [Lin and Pantel 2001]
Total Number of Answers

![Bar chart showing the total number of answers for KW-SYN, TextRunner, RESOLVER, DIRT, and USP. USP has the highest count, followed by RESOLVER, TextRunner, DIRT, and KW-SYN.]
Number of Correct Answers
Three times as many correct answers as second best
Highest accuracy: 88%
Qualitative Analysis

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

● Clusters in core forms
  { investigate, examine, evaluate, analyze, study, assay }
  { diminish, reduce, decrease, attenuate }
  { synthesis, production, secretion, release }
  { dramatically, substantially, significantly }

● Compositions
  amino acid, t cell, immune response, transcription factor,
  initiation site, binding site …
Question-Answer: Example

Q: What does IL-13 enhance?
A: The 12-lipoxygenase activity of murine macrophages

Sentence:

The data presented here indicate that (1) the 12-lipoxygenase activity of murine macrophages is upregulated in vitro and in vivo by IL-4 and/or IL-13, (2) this upregulation requires expression of the transcription factor STAT6, and (3) the constitutive expression of the enzyme appears to be STAT6 independent.
Future Work

- Learn subsumption hierarchy over meanings
- Incorporate more NLP into USP
- Scale up learning and inference
- Apply to larger corpora (e.g., entire PubMed)
Conclusion

- **USP**: The first approach for unsupervised semantic parsing
- Based on Markov Logic
- Learn target logical forms by recursively clustering variations of same meaning
- Novel form of relational clustering
- Applicable to general domains
- Substantially outperforms shallow methods