Unsupervised Semantic Parsing

Hoifung Poon

Dept. Computer Science & Eng.

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.,

Microsoft buys Powerset.

$\downarrow$

`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]
  - Zettlemoyer & Collins [2005, 2007, 2009]
  - 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**
- Ba s, 2006]
- S
- C

Three times as many correct answers as second best

- Applied it to extract knowledge from biomedical abstracts and answer questions
- Substantially outperforms TextRunner, DIRT

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

BUYS(-,-)

= {*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

- **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 $\boxed{\textit{Microsoft Corporation}}$

$\boxed{\textit{The Redmond software giant}}$ *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

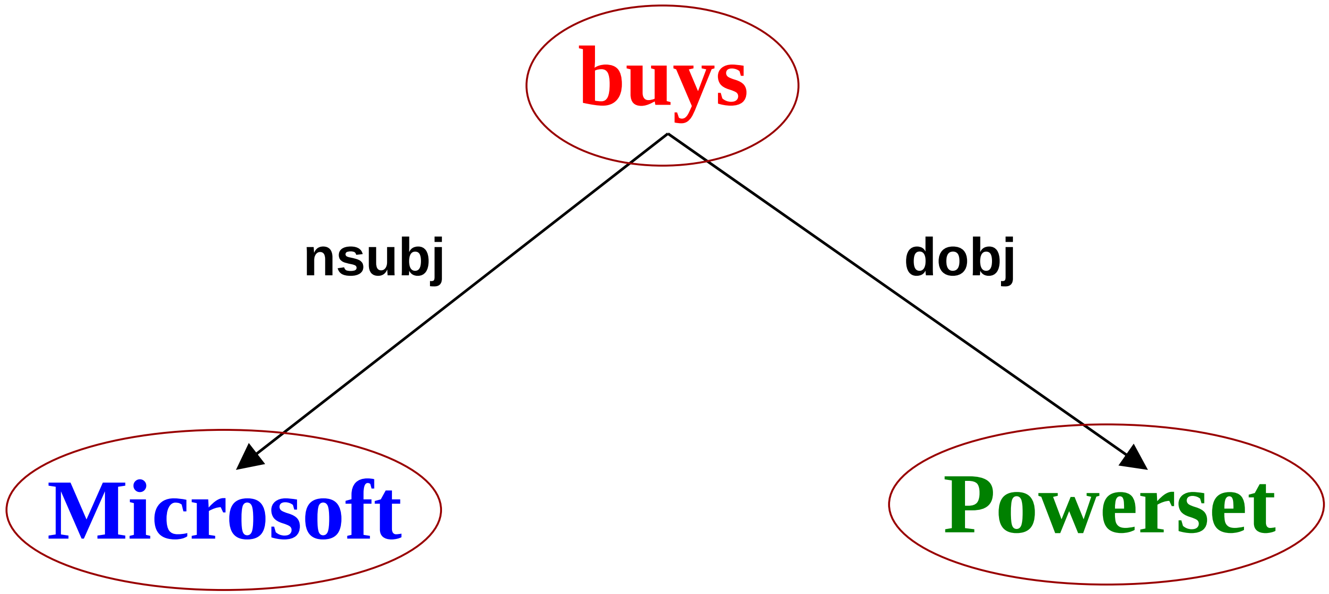
Probabilistic Model for USP

- 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
  - $\mathbf{F_i}$ is a formula in first-order l
  - $\mathbf{w_i}$ is a real number

Number of true groundings of $\mathbf{F_i}$

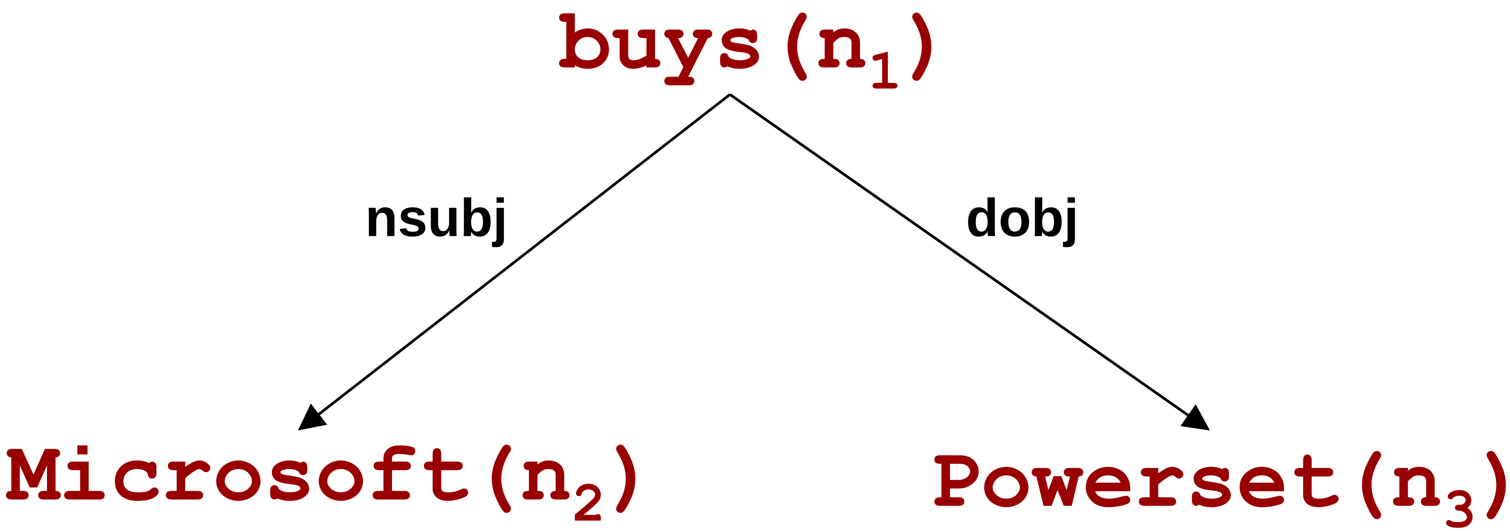
$$P(x) = \frac{1}{Z} \exp\left(\sum_i w_i \cdot N_i(x) \right)$$

Generating Quasi-Logical Forms



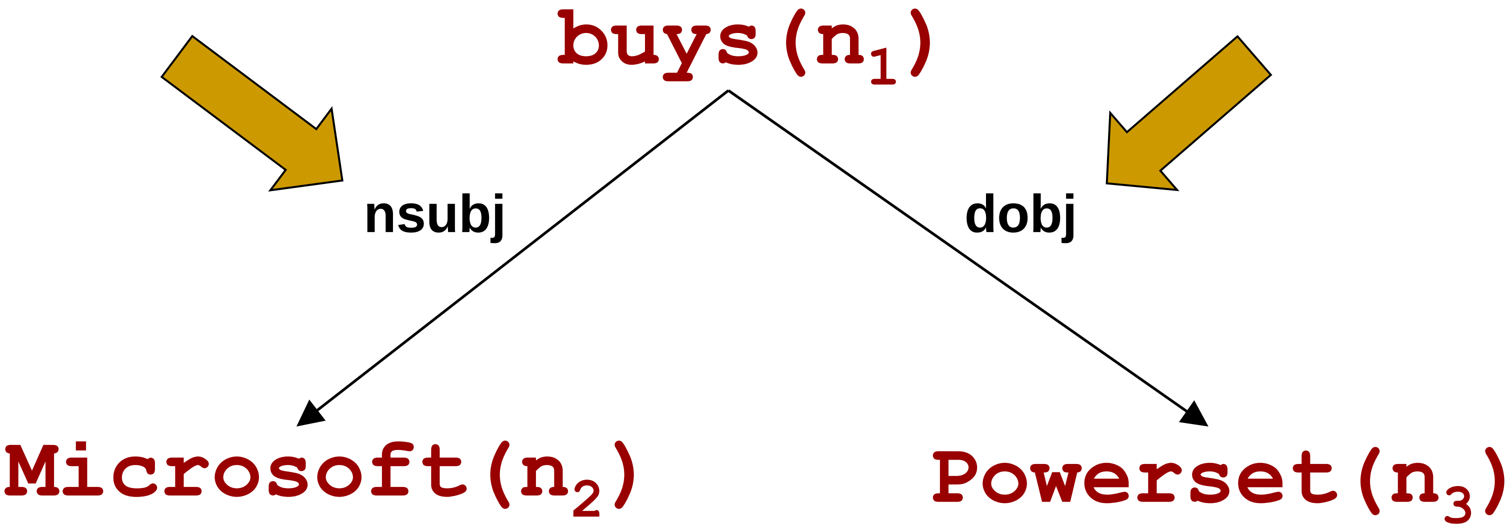
Convert each node into an unary atom

Generating Quasi-Logical Forms



n_1, n_2, n_3 are Skolem constants

Generating Quasi-Logical Forms



Convert each edge into a binary atom

Generating Quasi-Logical Forms

$\texttt{buys}(n_1)$

$\texttt{nsubj}(n_1, n_2)$ $\texttt{dobj}(n_1, n_3)$

$\texttt{Microsoft}(n_2)$ $\texttt{Powerset}(n_3)$

Convert each edge into a binary atom

A Semantic Parse

$\mathtt{buys}(n_1)$

$\mathtt{nsubj}(n_1, n_2)$ $\mathtt{dobj}(n_1, n_3)$

$\mathtt{Microsoft}(n_2)$ $\mathtt{Powerset}(n_3)$

Partition QLF into subformulas

A Semantic Parse

$\mathtt{buys(n_1)}$

$\mathtt{nsubj(n_1, n_2)}$ $\mathtt{dobj(n_1, n_3)}$

$\mathtt{Microsoft(n_2)}$ $\mathtt{Powerset(n_3)}$

Subformula ⇒ Lambda form:

Replace Skolem constant not in unary atom with a unique lambda variable

A Semantic Parse

$\mathbf{buys(n_1)}$

$\lambda \mathbf{x}_2.\mathbf{nsubj(n_1,x_2)}$ $\lambda \mathbf{x}_3.\mathbf{dobj(n_1,x_3)}$

$\mathrm{Microsoft(n_2)}$ $\mathrm{Powerset(n_3)}$

Subformula ⇒ Lambda form:

Replace Skolem constant not in unary atom with a unique lambda variable

A Semantic Parse

Core form

$\mathbf{buys}(n_1)$

Argument form

Argument form

$\lambda x_2.\mathbf{nsubj}(n_1, x_2)$ $\lambda x_3.\mathbf{dobj}(n_1, x_3)$

$\mathbf{Microsoft}(n_2)$ $\mathbf{Powerset}(n_3)$

Follow Davidsonian Semantics

Core form: No lambda variable

Argument form: One lambda variable

A Semantic Parse

$$\left[\begin{array}{c} \texttt{buys}(n_1) \\ \\ \lambda x_2.\texttt{nsubj}(n_1, x_2) \quad \lambda x_3.\texttt{dobj}(n_1, x_3) \end{array}\right] \in C_{\mathrm{BUYS}}$$

$$\boxed{\texttt{Microsoft}(n_2)} \in C_{\mathrm{MICROSOFT}}$$

$$\boxed{\texttt{Powerset}(n_3)} \in C_{\mathrm{POWERSET}}$$

Assign subformula to lambda-form cluster

Lambda-Form Cluster

C_{BUYS}

buys(n₁) 0.1

One formula in MLN

Learn weights for each pair of cluster and core form

:

Distribution over core forms

Lambda-Form Cluster

C_{BUYS}

`buys(n`$_1$`)`	0.1
`acquires(n`$_1$`)`	0.2
⋮	

A_{BUYER}

A_{BOUGHT}

A_{PRICE}

⋮

May contain variable number of argument types

Argument Type: A$_{BUYER}$

None 0.1

ne 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

$\mathtt{buys}(\mathtt{n}_1)$

Final logical form is obtained via lambda reduction

$C_{BUYS}(\mathtt{n}_1)$
$\wedge\ \lambda \mathtt{x}_2 . A_{BUYER}(\mathtt{n}_1, \mathtt{x}_2)$
$\wedge\ \lambda \mathtt{x}_3 . A_{BOUGHT}(\mathtt{n}_1, \mathtt{x}_3)$

Outline

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

Learning

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

$$L_{\Theta}(Q) = \log \sum_{S} P_{\Theta}(Q, S)$$

Use Greedy Search

- Search for $\boldsymbol{\Theta}$, $\boldsymbol{S}$ to maximize $\boldsymbol{P_{\Theta}(Q, S)}$
- Same objective as hard EM
- Directly optimize it rather than lower bound
- For fixed $\boldsymbol{S}$, derive optimal $\boldsymbol{\Theta}$ 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 $\boldsymbol{Q}$ and learned $\boldsymbol{\Theta}$, find semantic parse $\boldsymbol{S}$ to maximize $\boldsymbol{P_{\Theta}(Q, S)}$
- 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

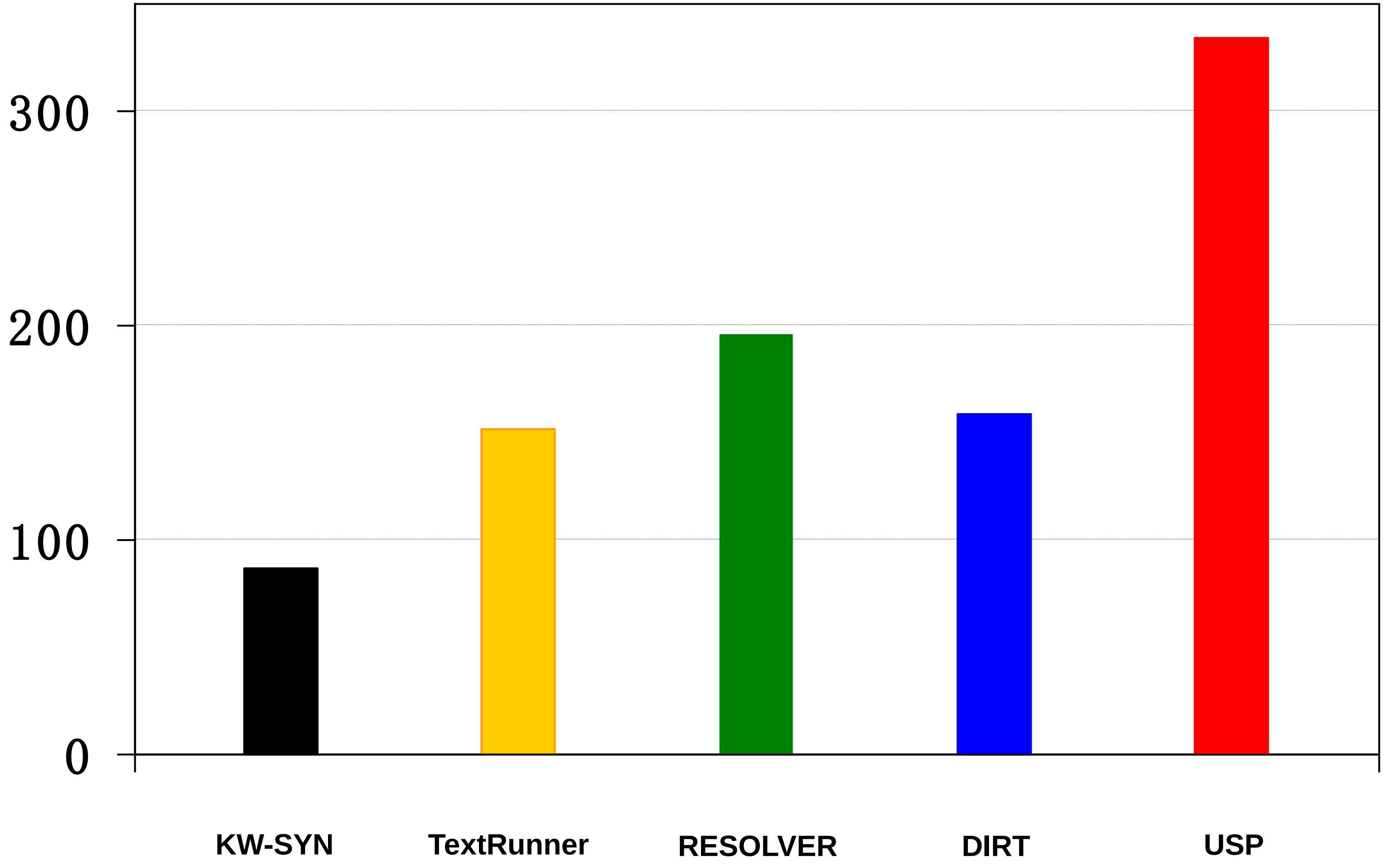
Dataset

- **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

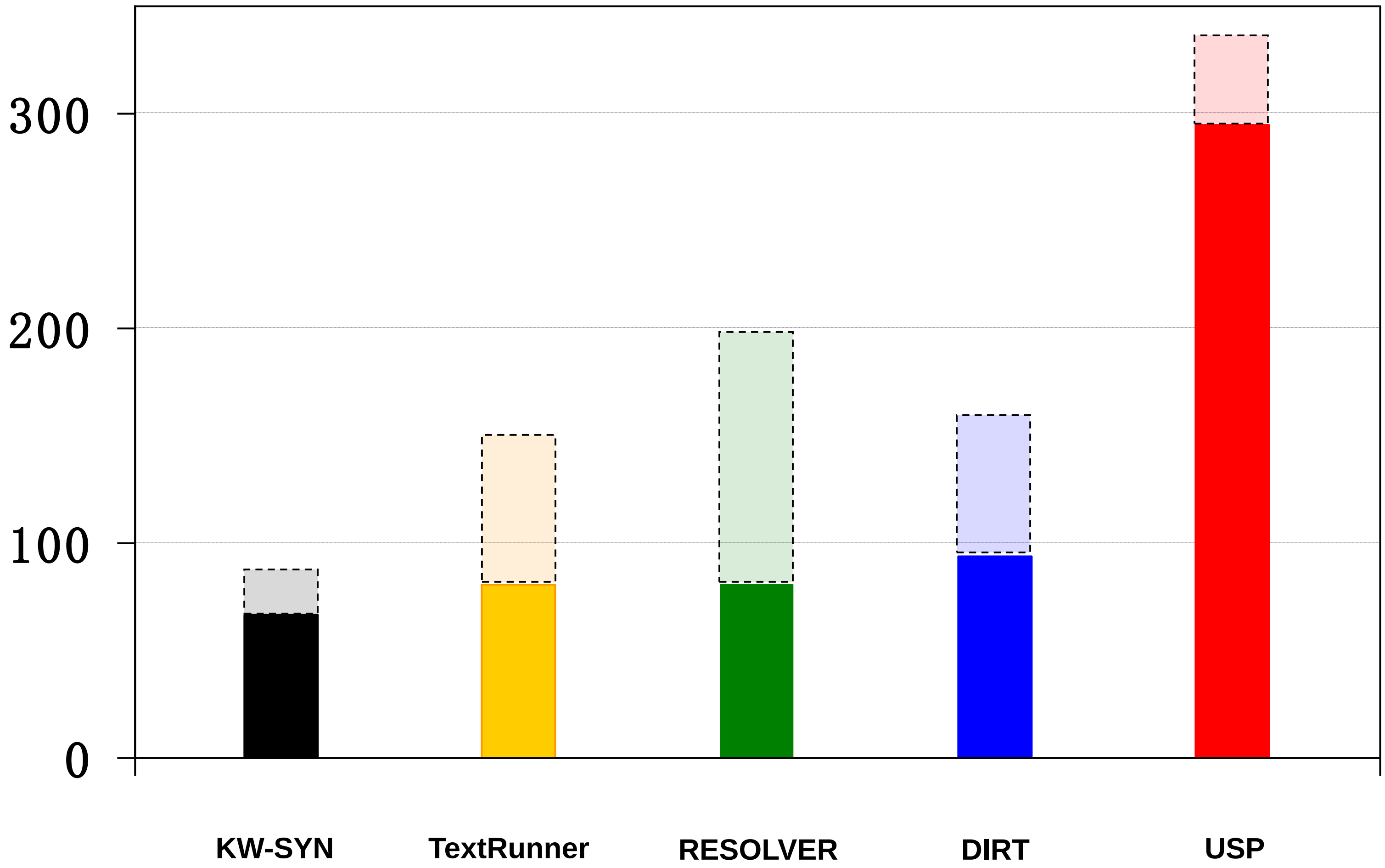
Systems

- **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]

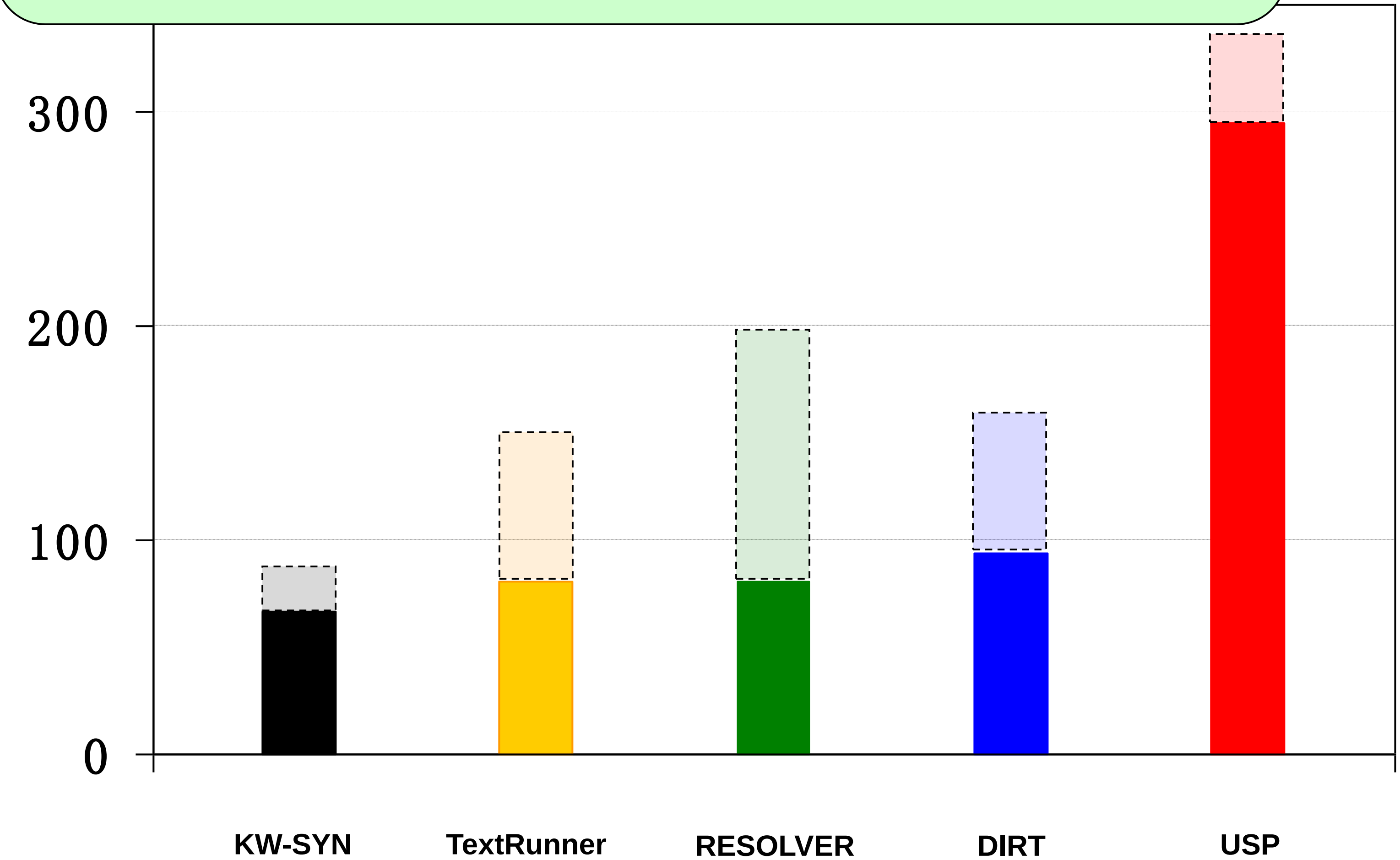
Total Number of Answers



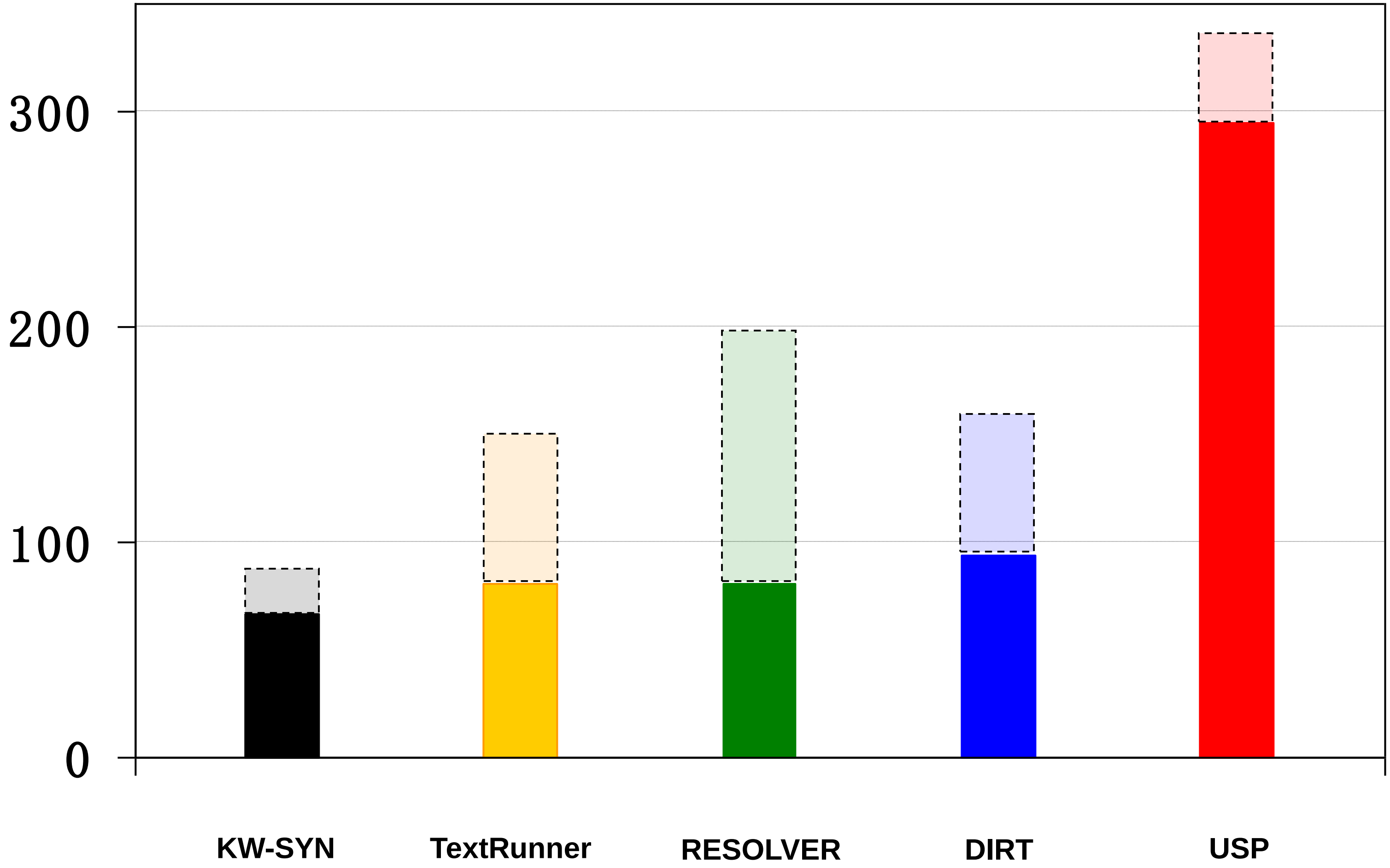
Number of Correct Answers



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