Learning to Execute Natural Language

Percy Liang

Computer Forum
April 16, 2013

Why natural language?

Information extraction

Observation: free-form text contains a wealth of information (e.g., news articles, blogs, scientific journals, etc.)

Goal: extract structured information from unstructured natural language

Intelligent user interfaces

Tasks: find information, perform data analytics, buy tickets, send email, etc.

Goal: allow people to accomplish more complex tasks efficiently using natural language

Book a non-stop flight from SF to Tokyo when the price drops by 20%.

Web search

Blocks world [Winograd 1971]

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: OK.
Person: I don't understand which pyramid you mean.
Person: (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: By “it”, I assume you mean the block which is taller than the one I am holding.
Computer: OK.
Person: What does the box contain?
Computer: The blue pyramid and the blue block.
Person: What is the pyramid supported by?
Computer: The box.
Person: How many blocks are not in the box?
Computer: Four of them.
Person: Is at least one of them narrower than the one which I told you to pick up?
Computer: Yes, the red cube.
How do we get deep and broad systems?

Models in NLP

Of countries that don't border an ocean, which has the most people?

Basic models:
- Topic models (e.g., Latent Dirichlet Allocation)
- n-gram language models
- Sequence models (e.g., HMM, conditional random fields)

More structured models (our focus):
- Syntactic models over parse trees
- Semantic models over logical forms

Deep question answering

Of countries that don't border an ocean, which has the most people?

Semantic parsing

\[
\text{argmax}(\lambda.x. \text{Country}(x) \land \neg \exists y. \text{Border}(x, y) \land \text{Ocean}(y), \lambda.x. \text{Population}(x))
\]

execute database query

Egypt

Point: to answer question, need to model the logical form

Training a semantic parser

Detailed supervision: manually annotate logical forms

What's Bulgaria's capital? Capital(Bulgaria)
When was Google started? DateFounded(Google)
What movies has Tom Cruise been in? \( \lambda.x. \text{Movie}(x) \land \text{ActedIn}(\text{Tom Cruise}, x) \)

Requires experts — slow and expensive, doesn't scale up!

Example: Penn Treebank (50K sentences annotated with parse trees) took 3 years

Training a semantic parser

Shallow supervision: question/answers pairs

What's Bulgaria's capital? Sofia
When was Google started? 1998
What movies has Tom Cruise been in? TopGun, VanillaSky, ...

- Get answers via crowdsourcing (no expertise required) or by scraping the web — fast and cheap (but noisy), scales up
- Logical forms modeled as latent variables

Summary so far:
- Modeling deep semantics of natural language is important
- Need to learn from natural/weak supervision to obtain broad coverage

Rest of talk:
- Spectral methods for learning latent-variable models
- Learning a broad coverage semantic parser
Spectral methods for learning latent-variable models

(joint work with Daniel Hsu, Sham Kakade, Arun Chaganty)

Latent-variable models

natural/weak supervision $\Rightarrow$ latent variables

Many applications:
- Semantic parsing
- Relation extraction
- Machine translation
- Speech recognition
- ...

Unsupervised learning

In general, latent-variable models lead to non-convex optimization problems (finding global optimum is NP hard)

Local optimization

Algorithms: EM, Gibbs sampling, variational methods

Problem: get stuck in local optima

Solution (heuristic): careful initialization, annealing, multiple restarts

Method of moments (global)

Algorithm (has rigorous theoretical guarantees):
- Compute aggregate statistics over data (trivial to parallelize)
- Perform simple linear algebra operations to obtain parameter estimates

Method of moments (global)

Use of data
- Global optimization efficient
- Local optimization no guarantees
- Method of moments inefficient

Computation
- Efficient
- Inefficient

In Big Data regime, method of moments is a win!

Missing: structural uncertainty, discriminative modeling
Structural uncertainty

\[ x: \text{I like algorithms.} \]

Our algorithm: unmixing [NIPS 2012]

Generative models (e.g., Naive Bayes):

\[ p(x, y, z) \]

Discriminative models (e.g., logistic regression, SVMs):

\[ p(y, z | x) \]

Our algorithm: for mixture of linear regressions [ICML 2013]

Semantic parsing

\[ \text{Of countries that don’t border an ocean, which has the most people?} \]

\[ \text{argmax}(\lambda z. \text{Country}(z) \land \neg \exists y. \text{Border}(z, y) \land \text{Ocean}(y), \lambda z. \text{Population}(z)) \]

\[ \text{execute database query} \]

Egypt

Semantic parsing (joint work with Jonathan Berant, Andrew Chou)

Training data

Expensive: logical forms

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005]
[Wong & Mooney, 2006; Kistadkowski et al., 2010]

What is the most populous city in California?
\[ \Rightarrow \text{argmax}(\lambda z. \text{city}(z) \land \text{loc}(z, \text{CA}), \lambda z. \text{pop.}(z)) \]

What is the most populous city in California?
\[ \Rightarrow \text{Los Angeles} \]

How many states border Oregon?
\[ \Rightarrow 3 \]

Can we learn with no annotated logical forms?

Cheap: answers

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005]
[Clarke et al., 2010]
[Liamb et al., 2011]

What is the most populous city in California?
\[ \Rightarrow \text{Los Angeles} \]

How many states border Oregon?
\[ \Rightarrow 3 \]

Experimental results

Task: US geography question/answering benchmark

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
<th>Logical</th>
<th>Location</th>
<th>Logical forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZOC</td>
<td>CCC [Zettlemoyer &amp; Collins, 2005]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ZOC/IT</td>
<td>replied CCC [Zettlemoyer &amp; Collins, 2005]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ZOC/IT/CO</td>
<td>CCC, worse [Kistadkowski et al., 2011]</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ZOC/IT/CS</td>
<td>our system</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ZOC/IT/CS/CL</td>
<td>our system</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Punchline: our system (without logical forms) matches previous work (with logical forms)
Towards broad coverage

Collecting question answering dataset from the Web:

- What shows has David Spade been in?
- What are the main rivers in Egypt?
- What year did Vince Young get drafted?
- In what year was President Kennedy shot?

Compared to previous datasets:

- Domain: from US geography to general facts
- Database size: from 500 to 400,000,000 (Freebase)
- Number of database predicates: from 40 to 30,000

Alignment

Challenge: figure out how words (e.g., born) map onto predicates (e.g., PlaceOfBirth)

Raw text: 1B web pages
Freebase: 400M assertions

Output: noisy mapping from words to predicates
Final step: train semantic parser using this mapping

Experimental results

Punchline: using alignment, can get same accuracy with 10 times fewer question/answer pairs

Summary

- Goal: deep natural language semantics from shallow supervision
- Consequence: need to learn latent-variable models
- Spectral methods: from intractable to easy by trading off computation and information — paradigm shift in learning
- Semantic parsing: state-of-the-art results learning only from question-answer pairs

Real-world impact

Increasing demand for deep language understanding

Thank you!