Learning to Classify from Natural Language Explanations

Shashank Srivastava

Joint work with Igor Labutov, Tom Mitchell
**Is this email important?**

- ‘Emails from my boss are usually important’
- ‘Such emails mention a deadline or a meeting’
- ‘The subject might say urgent …’
Towards Conversational ML?

- Traditional dependence on ‘big data’
  - Widely successful
  - Infeasible for long tail of learning problems

- Inherent statistical limitations
  - Coarsely, $n \approx \log(H)$
  - Intractable for representations like ontologies

- Extend ML to richer forms of input
  - Explanations, instructions, clarifications ...
Learning from Language

- Much of human learning is through language
  - Think books, lectures, student-teacher dialogue
Why now?

• If there is a new publication relevant to my current project, email it to me

• Whenever it snows at night, wake me up 30 minutes earlier

• If I receive a late night email from my advisor, ring alarm at full blast

Every user can be a programmer
Core issues

- Learning to Interpret NL
  - Parsing of NL statements to formal semantic representations

  \[ \text{‘Emails from my boss are usually important’} \rightarrow \text{equals( email.sender, getContactEmail("boss") )} \]

- Using Language to Operationalize Learning
  - E.g., Learning classification tasks from language

  \[ \text{([{0,1}]} \rightarrow \text{Binary classification} \]
How can language operationalize learning?

① By defining expressive features for learning tasks

Joint Concept Learning and Semantic Parsing from Natural Language Explanations
EMNLP 2017

② By specifying model constraints that can supervise training

Zero-shot Learning of Classifiers from Natural Language Quantification
ACL 2018
Part 1: Defining features using NL explanations
Defining features using NL

Is this email important?

‘Emails from my boss are usually important’
‘Such emails mention a deadline or a meeting’
‘The subject might say urgent ...’

NL explanations

Executable feature functions
NL Explanations as feature definitions

Semantic parsing maps NL to formal logical forms

Natural language statement (s)

‘three less than twenty times six’

‘What is the longest river that flows through Pittsburgh?’

‘Phishing emails often mention prices’

Logical form (l)

\[
\text{minus}( \text{prod}(20, 6), 3 )
\]

\[
\text{argmax}( \text{river}(x) \land \text{traverse}(x,y) \land \text{const}(y, \text{Pittsburgh}), \text{length})
\]

\[
\text{findSemanticCategory}( \text{MONEY, field:body})
\]

Evaluate in a context \( z = [l]_x \)

117

Ohio

Yes/No
How to interpret explanations?

- Pragmatics of language can guide parsing
  - A teacher’s intention would be use discriminative statements

**NL Explanation:** ‘Phishing emails often mention prices’

**Interpretation**

- l1: findWord(‘prices’, body)
- l2: findSemanticCategory(cat:MONEY, body)

**Discriminative?**

- ✗ l1: findWord(‘prices’, body)
- ✓ l2: findSemanticCategory(cat:MONEY, body)

**Jointly learn a classifier and a semantic parser!**

Don’t need annotated logical forms
No annotations of logical forms, supervision is only through concept labels \{0,1\} for examples
Coupled parsing and concept classification

<table>
<thead>
<tr>
<th>Input</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_j$</th>
<th>$S_m$</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$z_{11}$</td>
<td>$z_{12}$</td>
<td>$z_{1j}$</td>
<td>$z_{1m}$</td>
<td>$y_1$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>$x_i$</td>
<td>$z_{i1}$</td>
<td>$z_{i2}$</td>
<td>$z_{ij}$</td>
<td>$z_{im}$</td>
<td>$y_i$</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>---</td>
</tr>
<tr>
<td>$x_n$</td>
<td>$z_{n1}$</td>
<td>$z_{n2}$</td>
<td>$z_{nj}$</td>
<td>$z_{nm}$</td>
<td>$y_n$</td>
</tr>
</tbody>
</table>

$$\log P(y_i|x_i, s, \theta) = \log P(y_i|z_{i:}, \theta_{pred}) + \log P(z_{i:}|x_i, s, \theta_{parse})$$

**Classifier**

How likely are the observed concept labels, taking evaluations of NL statements as given?

**Parser**

How probable is a NL statement to apply for a given email (marginalized over all interpretations)?
Model training

- **Variational EM:**
  - **E-step:** Calculate estimates of $z_{ij}$ (evaluations of statements in different contexts)
    
    $q_j(z_j) \propto \exp \left( \mathbb{E}_{j' \neq j} [\log p_{\theta_c}(y|z, x)] + \log p_{\theta_p}(z_j|x, s_j) \right)$

    Prefer values that are discriminative
    Prefer interpretations supported by linguistic evidence

  - **M-step:** Updates concept classifier and semantic parsing models taking $z_{ij}$‘s as given.

Prefer interpretations of sentences that are both discriminative as well as supported by linguistic evidence
Concept to Learn: Phishing Emails

Phishing emails often contain mentions of prices

Executable feature functions $[l]$

findSemanticCategory(body cat:MONEY)

Instance feature Vector $[z]$

Classifier $\theta_c$

Feature Evaluator $\times$

Learning Algorithm $Y_{true}$ $Y_{pred}$

Update parameters

Natural language Statements $[s]$

Phishing emails often contain mentions of prices

Parser $\theta_p$

Update parameters
Data: Email classification

- Emails representing common email categories through AMT
  - Reminders, meeting invitations, requests from boss, internet humor, going out with friends, policy announcements, etc.
  - 1100 emails, 7 types

E.g. You are writing an email to yourself as a reminder to do something

<table>
<thead>
<tr>
<th><strong>Subject:</strong></th>
<th>Note to self - Move the Bodies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>From:</strong></td>
<td><a href="mailto:john@initech-corp.com">john@initech-corp.com</a></td>
</tr>
<tr>
<td><strong>To:</strong></td>
<td><a href="mailto:john@initech-corp.com">john@initech-corp.com</a></td>
</tr>
<tr>
<td><strong>Body:</strong></td>
<td>Blasted police. I need to pick up lye and move the bodies tonight. Forecast is rain and the swamp's filling up. Need to remember galoshes, too.</td>
</tr>
<tr>
<td><strong>Attachment:</strong></td>
<td>none</td>
</tr>
</tbody>
</table>
Data: NL Explanations

- Dataset of statements explaining each concept
- Turkers describe emails from each category
- 30 statements for each category

Sample explanations:
Most reminders mention a date and a time in the message of the email
The sender of the email is the same as the recipient
These emails usually close with a name or title
These emails sometimes have jpg attachments
The email likely has words like "policy" or "announcement" in the subject
Emails from a public domain are not office requests
Results: Email classification

- Average F1 across concepts
- Significantly better than best baseline for 6 of 7 categories
Learning from fewer examples

- LNL consistently outperforms BoW, especially with fewer examples
Predicted logical forms are often highly correlated

\[
\text{getPhraseMention(email, stringVal('meeting'))}
\]
\[
\text{getPhraseMention(body, stringVal('meeting'))}
\]
Summary

- NL explanations can define executable feature functions that improve concept learning performance

- Pragmatic context can guide learning of semantic parsers even with very weak supervision (class-labels only)

- Each domain requires specifying a DSL (one-time effort)
  - Reusable across long tail of categories
Part 2: Incorporating model constraints from NL
NL advice as defining model constraints

Show my important emails.

What are important emails?

If the subject says ‘urgent’, it is almost certainly important.

Most emails from John are important.

Emails that I reply to are usually important.

Unimportant emails are often sent to a list.

- Potentially enable learning without labeled examples?
- Leverage quantifier expressions in language
Sequential Approach

Emails that I reply to are **usually important**

Mapping language to quantitative constraints

Incorporating constraints in model training

**Semantic Parser**

\[ x \rightarrow (\text{email.replied} == \text{true}) \]
\[ y \rightarrow \text{important: true} \]
\[ \mathbb{E}_{y|x}[\phi(x, y)] = b_{usually} \]

**Posterior Regularization**

[\( \theta \)]

**Classifier**

\[ f : x \rightarrow y \]

Unlabeled data
Emails that I reply to are usually important

Mapping language to quantitative constraints

Semantic Parser

\[ x \rightarrow (\text{email.replied} == \text{true}) \]
\[ y \rightarrow \text{important:true} \]
\[ \mathbb{E}_{y|x}[\phi(x, y)] = b_{usually} \]

Posterior Regularization

Classifier

\[ f : x \rightarrow y \]

Unlabeled data
Training classifiers from declarative NL

- Explanations encode multiple properties that can aid statistical learning

- ‘Emails that I reply to are usually important’

1. Features important for a learning problem
   - x : repliedTo:true

2. Class labels
   - y : Important

3. Type of Relationship b/w features and labels
   - P(y|x)

4. Strength of Relationship
   - Specified by quantifier?
Semantic parsing

- **Constraint types:**
  1. *About a third of the emails that I get are important*: \( P(y) \)
  2. *Emails that I reply to are usually important*: \( P(y \mid x) \)
  3. *I almost always reply to important emails*: \( P(x \mid y) \)

- **Novelty largely in identifying the type of the assertion**
  - Primarily depends on syntactic features
    - Features based on dependency paths
    - Presence/absence of negation words
    - Identifying active/passive voice
    - Order of occurrence of triggers for x and y

  ‘*Emails that I reply to are usually important*’

  \[ P(\text{important} \mid \text{replied: true}) \approx p_{\text{usually}} \]
Semantic parsing

- Leverage semantics of linguistic quantifiers
  - Associate point probability estimates for frequency adverbs and determiners

<table>
<thead>
<tr>
<th>Frequency quantifier</th>
<th>Probability value</th>
</tr>
</thead>
<tbody>
<tr>
<td>always, certainly, definitely, all</td>
<td>0.95</td>
</tr>
<tr>
<td>usually, normally, generally, likely</td>
<td>0.70</td>
</tr>
<tr>
<td>most, majority</td>
<td>0.60</td>
</tr>
<tr>
<td>often, half</td>
<td>0.50</td>
</tr>
<tr>
<td>many</td>
<td>0.40</td>
</tr>
<tr>
<td>sometimes, frequently, some</td>
<td>0.30</td>
</tr>
<tr>
<td>few, occasionally</td>
<td>0.20</td>
</tr>
<tr>
<td>rarely, seldom</td>
<td>0.10</td>
</tr>
<tr>
<td>never</td>
<td>0.05</td>
</tr>
</tbody>
</table>

- Purely subjective beliefs, not calibrated on any data
Sequential Approach

Incorporating constraints in model training
Posterior Regularization

- Use the posterior regularization (PR) principle to imbue human-provided advice in learned models
  - Unobserved class labels as latent variables

- PR optimizes a latent variable model subject to a set of constraints on the posterior distribution $p_{\theta}(y \mid x)$

$y_1 = ?$
$y_2 = ?$
$y_3 = ?$
Probability Assertions as PR Constraints

- PR can handle linear constraints over distributions of latent variables

\[ Q := \{ q_x(y) : \mathbb{E}_q[\phi(x, y)] \leq b \} \]

Linear bounds on expected values of features under q

- Can convert each constraint type to this form:

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(y</td>
<td>x)</td>
<td>Emails that I reply to are usually important</td>
</tr>
<tr>
<td>P(x</td>
<td>y)</td>
<td>I almost always reply to important emails</td>
</tr>
<tr>
<td>P(y)</td>
<td>About a third of all emails I get are important</td>
<td>Same as P(y</td>
</tr>
</tbody>
</table>
Posterior Regularization

- Each constraint from the semantic parser can be expressed in the form compatible with PR
  - Conjunction of all such constraints specifies $Q$

- Train with modified EM to maximize PR objective:

$$J_Q(\theta) = \mathcal{L}(\theta) - \min_{q \in Q} KL(q \mid p_\theta(Y \mid X))$$

- Improve data likelihood
- Emulate human advice
Synthetic shape classification

Turkers observe samples of shapes from synthetically generated datasets, and describe them through statements.

- Selected shapes are almost always a square
- Other shapes rarely have a blue border
- If a shape has a red fill, it is most likely not a selected shape...

- 50 datasets
- 30 workers
- 4.3 statements per task on average
Each dot represents a dataset (and corresponding classification task) generated from a known distribution.
# Average Classification Accuracy (Shapes data)

## Table

<table>
<thead>
<tr>
<th>Approach</th>
<th>Avg Accuracy</th>
<th>Access to labels</th>
<th>Access to statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNQ</td>
<td>0.751</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Bayes Optimal</td>
<td>0.831</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.737</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Random</td>
<td>0.524</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
### Average Classification Accuracy (Shapes data)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Avg Accuracy</th>
<th>Access to labels</th>
<th>Access to statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNQ</td>
<td>0.751</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Bayes Optimal</td>
<td>0.831</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.737</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Random</td>
<td>0.524</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>LNQ (no quantification)</td>
<td>0.545</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>LNQ (coarse quantification)</td>
<td>0.679</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
## Average Classification Accuracy (Shapes data)

<table>
<thead>
<tr>
<th>Approach</th>
<th>Avg Accuracy</th>
<th>Access to labels</th>
<th>Access to statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNQ</td>
<td>0.751</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Bayes Optimal</td>
<td>0.831</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.737</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Random</td>
<td>0.524</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>LNQ (no quantification)</td>
<td>0.545</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>LNQ (coarse quantification)</td>
<td>0.679</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Human teacher</td>
<td>0.802</td>
<td>yes</td>
<td>yes (writes descriptions)</td>
</tr>
<tr>
<td>Human learner</td>
<td>0.734</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
Real classification tasks

- A specimen that has a striped crown is likely to be a selected bird
- Birds in the other category rarely ever have dagger-shaped beaks

Example statements:
Results: Bird Species Identification

![Diagram showing bird species identification results for LNQ, LR, and FLGE+ methods. The x-axis represents bird species including Sayornis, Frigatebird, Gadwell, White-br..., Mocking..., Nighthawk, Ovenbird, Harris Sp..., Cape Glo..., and Cerulean. The y-axis represents Avg F1, ranging from 0.5 to 0.9.]
Results: Emails Categorization

Performance by training from both quantification and labels

- About a third of statements used quantifiers
Empirical distributions of probability values

**Rarely**

$$\mu = 0.06$$  
$$\sigma = 0.05$$

**Sometimes**

$$\mu = 0.29$$  
$$\sigma = 0.18$$

**Often**

$$\mu = 0.46$$  
$$\sigma = 0.15$$

**Majority**

$$\mu = 0.73$$  
$$\sigma = 0.16$$

**Most**

$$\mu = 0.81$$  
$$\sigma = 0.15$$

**(most)Likely**

$$\mu = 0.86$$  
$$\sigma = 0.09$$
Summary

➢ Declarative NL can supervise learning in limited data settings

➢ Differential associative strengths of linguistic quantifiers can be effective towards zero-shot concept learning

➢ Possible to learn through a blend of strategies
Other directions

- Learning with mixed initiative dialog
  - Allow the learner to ask questions?

- Learning from multiple teachers
  - How to learn from contradictory advice?

- Pairing explanations with demonstrations, curricular learning,...
Questions?