Structured Output Learning with Indirect Supervision

Sriraam Appusamy Subramanian

School of Computing, University of Utah

November 18, 2014
Outline

• Motivation
• Joint-Learning Using Indirect Supervision
• Experiments
• Discussions
Supervised Learning

[Labeled Data]

Chang, M., Goldwasser, D., Roth, D., and Srikumar, V. Discriminative learning over constrained latent representations. In NAACL, 2010.]

[Training Process]

[Yu, C. and Joachims, T. Learning structural svms with latent variables. In ICML, 2009]

[Model]

[Yu, C. and Joachims, T. Learning structural svms with latent variables. In ICML, 2009]
Supervised Learning

• Large quantities of Annotated data can be tedious to obtain for structured output learning

• Other examples:
  • Tagging parts of a car (Object Recognition)
  • Identifying fields from Advertisement/Citation (Information Extraction)
Lafferty et al., and Taskar et al. proposed several discriminative algorithms for learning structured output predictors before 2005.

Trained Model

Does this output depend on something else besides the features in the input?
Idea

- Can knowing that it is/isn’t a citation help with learning to extract parts of a citation?

- Given an image, can knowing if there is/isn’t a car, help me learn a better model to predict parts of a car?
Indirect Supervision

- Companion Binary Problem: A related problem whose solution does not predict the structured output

- Indirect Supervision: Using labeled datasets for the companion binary problem in structured output learning

- *Can the structured output prediction task benefit from indirect supervision to the companion problem?*
Joint-Learning using Indirect Supervision

Generalizes SVM learning which is given by:

$$\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w)$$  \hspace{1cm} (1)

Allows specification of constraints on output space

$$\forall (x, -1)x \in B^-, \forall h \in H(x), w^T \Phi(x, h) \leq 0$$ \hspace{1cm} (2)

$$\forall (x, +1)x \in B^+, \exists h \in H(x), w^T \Phi(x, h) \geq 0$$ \hspace{1cm} (3)
JLIS learning involves minimization of

$$\frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B} L_B(x_i, y_i, w)$$  \hspace{1cm} (4)

where $L_B$ and $L_S$ are given by

$$L_B(x_i, y_i, w) = \ell(1 - y_i \max_{h \in H(x)} (w^T \Phi_B(x_i, h)))$$  \hspace{1cm} (5)

$$L_S(x_i, h_i, w) = \ell(\max_h (\Delta(h, h_i) - w^T (\Phi(x_i, h_i) - \Phi(x_i, h))))$$  \hspace{1cm} (6)
What does it mean to use JLIS?
What does it mean to use JLIS?
Objective Function $Q(w)$ is given by

$$
\frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B^+} L_B(x_i, y_i, w) + C_2 \sum_{i \in B^-} \ell(1 - \max_h (w^T \Phi_B(x_i, h)))
$$

(7)

Solved by:

- Cutting Plane Algorithm
- Dual Co-ordinate Descent Algorithm
Experiments

• Phonetic Alignment Transliteration

• Part-Of-Speech Tagging

• Information Extraction
Phonetic Alignment Transliteration

- Structured Output Problem: Given a pair of English-Hebrew words, find the best Phonetic alignment between character sequences.

- Companion Binary Task: Both the Named Entities refer to same entity?

- Direct Supervision: English-Hebrew Dataset(100)

- Indirect Supervision: Web Crawling (1000+, 10,000−)
## Phonetic Alignment Transliteration Results

**Figure: F Score**

| $|S|$ | Size of $B$ | 0 (SSVM) | 2$k$ | 4$k$ | 8$k$ |
|-----|-------------|---------|------|------|------|
| 10  |             | 72.9    | 78.8 | 79.8 | 80.0 (26.2%) |
| 20  |             | 82.1    | 84.6 | 84.7 | 85.4 (18.4%) |
| 40  |             | 85.7    | 86.9 | 87.2 | 87.4 (12.0%) |
| 80  |             | 88.6    | 89.4 | 89.0 | 89.4 (7.1%) |
Part of Speech Tagging

- **Structured Output Problem**: Find the correct sequence of POS tags for a given sentence.

- **Companion Binary Task**: Is there a valid tag sequence for a given sentence?

- **Direct Supervision**: Wall Street Journal corpus (1000 sentences)

- **Indirect Supervision**: $(1000^+, 1000^-)$
Part of Speech Tagging Results
## Information Extraction Results

<table>
<thead>
<tr>
<th>Advertisements</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>S</td>
</tr>
<tr>
<td>500</td>
<td>57.18</td>
</tr>
<tr>
<td>1000</td>
<td>64.93</td>
</tr>
<tr>
<td>2000</td>
<td>69.57</td>
</tr>
<tr>
<td>4000</td>
<td>74.24</td>
</tr>
<tr>
<td>19k</td>
<td>78.07</td>
</tr>
</tbody>
</table>
Importance of Negative Examples

![Graph showing the importance of negative examples](image-url)
Can it be reversed?

- Improve Binary classification performance using structured prediction task as a companion problem

- Joint approach performs significantly better.
Thank You