Structured Prediction with Indirect Supervision

Ming-Wei Chang
University of Illinois at Urbana-Champaign

Joint Work With James Clarke, Dan Goldwasser, Lev Ratinov, Vivek Srikumar, and Dan Roth

June 27th, 2011

Talk at the Joint ICML-ACL-ISCA symposium
Reducing supervision effort is crucial

**Semantic Parsing**

**INPUT**

What is the largest state that borders New York and Maryland?

**OUTPUT**

\[
\text{largest( state( next\_to( state(NY) ) AND next\_to(state(MD))) )}
\]
Reducing supervision effort is crucial

**Semantic Parsing**

**INPUT** What is the largest state that borders New York and Maryland?

**OUTPUT** \( \text{largest( state( next\_to( state(NY) ) AND next\_to(state(MD))))} \)

**A structured task: multiple interdependent decisions**

- city(NY) or state(NY)?
- state(next\_to(·)) \( \neq \) next\_to(state(·))
Reducing supervision effort is crucial

**Semantic Parsing**

**INPUT**

What is the largest state that borders New York and Maryland?

**OUTPUT**

```
largest( state( next_to( state(NY) ) AND next_to(state(MD))) )
```

**A structured task: multiple interdependent decisions**

- city(NY) or state(NY)?
- state(next_to(·)) \(\neq\) next_to(state(·))

**Supervision cost**

- Labeling data is **very expensive**!
- The annotators need to know how to write meaning representation
Main Idea: Indirect Supervision

Example

- Input: Human Query, Output: Meaning Representation
Main Idea: Indirect Supervision

Example

- Input Human Query, Output Meaning Representation
Main Idea: Indirect Supervision

Example

- Input: Human Query, Output: Meaning Representation
- (Indirect) Simple Output: Is the answer correct?

Use indirect supervision signals instead of supervising at the level of complex structures. Indirect supervision signals are easier to obtain.
Main Idea: Indirect Supervision

Example

- **Input** Human Query, **Output** Meaning Representation
- **(Indirect) Simple Output**: Is the answer correct?
Main Idea: Indirect Supervision

Example

- **Input** Human Query, **Output** Meaning Representation
- **(Indirect) Simple Output**: Is the answer correct?

Use indirect supervision signals

- Instead of supervising at the level of complex structures, use indirect supervision signals
- Indirect supervision signals are easier to obtain
Part I: Learning with Latent Structure

Part II: Learning with Indirect Supervision
Part I: Learning with Latent Structure

Part II: Learning with Indirect Supervision
Part I: Learning with Latent Structure

Part II: Learning with Indirect Supervision
Part I: Learning with Latent Structure

Input → Complex Structural Variables → Simple Output

Part II: Learning with Indirect Supervision

Input → Complex Structural Variables → (Indirect) Simple Output
Outline

Part I: Learning with Latent Structure

Input

Complex Structural Variables

Simple Output

Target

Part II: Learning with Indirect Supervision

Input

Complex Structural Variables

(Indirect) Simple Output

Target
Example task: Paraphrase Identification

Yes/NO

Alan will face murder charges, said Bob with murder charges will be charged.

Q: Are sentence 1 and sentence 2 paraphrases of each other?
Example task: Paraphrase Identification

Q: Are sentence 1 and sentence 2 paraphrases of each other?

- Yes, but why?
- They carry the same information!

Justifying the decision requires an intermediate representation.
Example task: Paraphrase Identification

Yes/NO

Alan will face murder charges.
Bob said he will be charged with murder.

Q: Are sentence 1 and sentence 2 paraphrases of each other?
   - Yes, but why?
   - They carry the same information!

Justifying the decision requires an intermediate representation.

Just an example; the real intermediate representation is more complicated.
Example task: Paraphrase Identification

Q: Are sentence 1 and sentence 2 paraphrases of each other?
- Yes, but why?
- They carry the same information!

Justifying the decision requires an intermediate representation.

Just an example; the real intermediate representation is more complicated.

Problem of interests
- Binary output problem: \( z \in \{-1, 1\} \)
- Intermediate representation: \( h \)
  - Some structure that justifies the positive label
  - The intermediate representation is latent (not present in the data)
The intuition behind the joint approach

Yes/NO

Alan will face murder charges, Bob said

Alan will be charged with murder

Bob said
The intuition behind the joint approach

intermediate representation ⇔ \{1, -1\}

- Only positive examples have good intermediate representations
- **No** negative example has a good intermediate representation
The intuition behind the joint approach

intermediate representation $\Leftrightarrow \{1, -1\}$

- Only positive examples have good intermediate representations
- **No** negative example has a good intermediate representation

$x$: a sentence pair
$h$: an alignment between two sentences
$\mathcal{H}(x)$: all possible alignments for $x$
The intuition behind the joint approach

intermediate representation $\Leftrightarrow \{1, -1\}$

- Only positive examples have good intermediate representations
- **No** negative example has a good intermediate representation

$x$: a sentence pair, **weight vector**: $w$

$h$: an alignment between two sentences

$\mathcal{H}(x)$: all possible alignments for $x$
The intuition behind the joint approach

**intermediate representation** $\Leftrightarrow \{1, -1\}$

- Only positive examples have good intermediate representations
- **No** negative example has a good intermediate representation

$x$: a sentence pair, **weight vector**: $w$

$h$: an alignment between two sentences

$\mathcal{H}(x)$: all possible alignments for $x$

- **Pair** $x_1$ is positive
  - There must exist a good explanation that justifies the positive label
  - $\exists h, w^T \Phi(x_1, h) \geq 0$

- **Pair** $x_2$ is negative
  - No explanation is good enough to justify the positive label
  - $\forall h, w^T \Phi(x_2, h) \leq 0$
Geometric interpretation: the case of two examples

- **Pair $x_1$ is positive**
  - There must exist a good explanation that justifies the positive label
  - $\exists h, w^T \Phi(x_1, h) \geq 0$, or $\max_h w^T \Phi(x_1, h) \geq 0$

- **Pair $x_2$ is negative**
  - No explanation is good enough to justify the positive label
  - $\forall h, w^T \Phi(x_2, h) \leq 0$, or $\max_h w^T \Phi(x_2, h) \leq 0$
Geometric interpretation: the case of two examples

- **Pair $x_1$ is positive**
  - There must exist a good explanation that justifies the positive label
  - $\exists h, w^T \Phi(x_1, h) \geq 0$, or $\max_h w^T \Phi(x_1, h) \geq 0$

- **Pair $x_2$ is negative**
  - No explanation is good enough to justify the positive label
  - $\forall h, w^T \Phi(x_2, h) \leq 0$, or $\max_h w^T \Phi(x_2, h) \leq 0$
Geometric interpretation: the case of two examples

- Pair $\mathbf{x}_1$ is positive
  - There must exist a good explanation that justifies the positive label
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$

- Pair $\mathbf{x}_2$ is negative
  - No explanation is good enough to justify the positive label
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$
Geometric interpretation: the case of two examples

- **Pair \( x_1 \) is positive**
  - There must exist a good explanation that justifies the positive label
  \[ \exists h, w^T \Phi(x_1, h) \geq 0, \text{ or } \max_h w^T \Phi(x_1, h) \geq 0 \]
- **Pair \( x_2 \) is negative**
  - No explanation is good enough to justify the positive label
  \[ \forall h, w^T \Phi(x_2, h) \leq 0, \text{ or } \max_h w^T \Phi(x_2, h) \leq 0 \]

\[ \{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \} \]
Geometric interpretation: the case of two examples

- **Pair $x_1$ is positive**
  - There must exist a good explanation that justifies the positive label
  - $\exists h, w^T \Phi(x_1, h) \geq 0$, or $\max_h w^T \Phi(x_1, h) \geq 0$

- **Pair $x_2$ is negative**
  - No explanation is good enough to justify the positive label
  - $\forall h, w^T \Phi(x_2, h) \leq 0$, or $\max_h w^T \Phi(x_2, h) \leq 0$

\[
\{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \} \\
\{ \Phi(x_2, h) \mid h \in \mathcal{H}(x_2) \}
\]
Geometric interpretation: the case of two examples

- **Pair \( x_1 \) is positive**
  - There must exist a good explanation that justifies the positive label
  - \( \exists h, w^T \Phi(x_1, h) \geq 0 \), or \( \max_h w^T \Phi(x_1, h) \geq 0 \)

- **Pair \( x_2 \) is negative**
  - No explanation is good enough to justify the positive label
  - \( \forall h, w^T \Phi(x_2, h) \leq 0 \), or \( \max_h w^T \Phi(x_2, h) \leq 0 \)
Geometric interpretation: the case of two examples

- **Pair \( x_1 \) is positive**
  - There must exist a good explanation that justifies the positive label
  - \( \exists h, w^T \Phi(x_1, h) \geq 0 \) or \( \max_h w^T \Phi(x_1, h) \geq 0 \)

- **Pair \( x_2 \) is negative**
  - No explanation is good enough to justify the positive label
  - \( \forall h, w^T \Phi(x_2, h) \leq 0 \) or \( \max_h w^T \Phi(x_2, h) \leq 0 \)
Geometric interpretation: the case of two examples

- **Pair \( x_1 \) is positive**
  - There must exist a good explanation that justifies the positive label
  - \( \exists \mathbf{h}, \mathbf{w}^T \Phi(x_1, \mathbf{h}) \geq 0 \), or \( \max_h \mathbf{w}^T \Phi(x_1, \mathbf{h}) \geq 0 \)

- **Pair \( x_2 \) is negative**
  - No explanation is good enough to justify the positive label
  - \( \forall \mathbf{h}, \mathbf{w}^T \Phi(x_2, \mathbf{h}) \leq 0 \), or \( \max_h \mathbf{w}^T \Phi(x_2, \mathbf{h}) \leq 0 \)
Geometric interpretation: the case of two examples

- **Pair \( x_1 \) is positive**
  - There must exist a good explanation that justifies the positive label
  - \( \exists \textbf{h}, \textbf{w}^T \Phi(x_1, \textbf{h}) \geq 0 \), or \( \max_{\textbf{h}} \textbf{w}^T \Phi(x_1, \textbf{h}) \geq 0 \)
- **Pair \( x_2 \) is negative**
  - No explanation is good enough to justify the positive label
  - \( \forall \textbf{h}, \textbf{w}^T \Phi(x_2, \textbf{h}) \leq 0 \), or \( \max_{\textbf{h}} \textbf{w}^T \Phi(x_2, \textbf{h}) \leq 0 \)
Geometric interpretation: the case of two examples

- Pair $x_1$ is positive
  - There must exist a good explanation that justifies the positive label
    - $\exists h, w^T \Phi(x_1, h) \geq 0$, or $\max_h w^T \Phi(x_1, h) \geq 0$

- Pair $x_2$ is negative
  - No explanation is good enough to justify the positive label
    - $\forall h, w^T \Phi(x_2, h) \leq 0$, or $\max_h w^T \Phi(x_2, h) \leq 0$

- The prediction function:
  $$\max_h w^T \Phi(x, h)$$

\[
\{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \} \\
\{ \Phi(x_2, h) \mid h \in \mathcal{H}(x_2) \}
\]
Find Structures

- In the learning algorithm, we need to solve
  \[
  \max_h \mathbf{w}^T \Phi(\mathbf{x}, h)
  \]
- A problem of assigning values to multiple interacting discrete variables

Constraint Based Declarative Framework

- We formulate this problem as an Integer Linear Programming problem (Roth and Yih 2004)
  1. Allow one to define the knowledge necessary for the problem declaratively
  2. Avoid designing a special purpose inference algorithm for each problem.
- Final System: Learning Constrained Latent Representation (LCLR)
Find Structures

- In the learning algorithm, we need to solve: \( \max_h w^T \Phi(x, h) \)
- A problem of assigning values to multiple interacting discrete variables

Constraint Based Declarative Framework

- We formulate this problem as an Integer Linear Programming problem (Roth and Yih 2004)
  1. Allow one to define the knowledge necessary for the problem declaratively
  2. Avoid designing a special purpose inference algorithm for each problem.
- Final System: Learning Constrained Latent Representation (LCLR)
Optimizing the objective function

\[
\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} L_B(x_i, y_i, w) = \\
\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \ell(-z_i \max_{h \in H} w^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x))
\]

- **Not a regular LR/SVM**: Inference procedures inside (pink boxed)
- **No shortcut** Calling a LR/SVM solver multiple times will not work
- Similar to MI-SVM and Latent-SVM
Optimizing the objective function

\[
\begin{align*}
\min_w & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l L_B(x_i, y_i, w) = \\
\min_w & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \ell(-z_i \max_{h \in \mathcal{H}} w^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x)) \\
\end{align*}
\]

- **Not a regular LR/SVM**: Inference procedures inside (pink boxed)
- **No shortcut** Calling a LR/SVM solver multiple times will not work
- Similar to MI-SVM and Latent-SVM

**Our solution**

- A new optimization algorithm: Focus on square-hinge loss
  - EM-like procedure + Cutting plane methods + Dual coordinate descent
  - \[
  \min_w \frac{1}{2} \|w\|^2 + C \sum_{z_i = -1} L_B(x_i, y_i, w) + C \sum_{z_i = +1} L_B(x_i, y_i, w)
  \]
- Code available:
  
  [http://cogcomp.cs.illinois.edu/page/software](http://cogcomp.cs.illinois.edu/page/software)
Experimental setting

Tasks

- Transliteration: Is named entity B a transliteration of A?
- Textual Entailment: Can sentence A entail sentence B?
- Paraphrase Identification

Goal of experiments

- Determine if a joint approach be better than a two-stage approach?
- Joint approach also learns latent structures automatically

Two-stage approach versus LCLR

- Exactly the same features and definition of latent structures
  - Our two-stage approach uses a domain-dependent heuristic to find an intermediate representation
  - LCLR finds the intermediate representation automatically
- Initialization of LCLR: two-stage
### Transliteration System

<table>
<thead>
<tr>
<th></th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Goldwasser and Roth 2008)</td>
<td>⋆</td>
<td>N/A</td>
<td>89.4</td>
<td></td>
</tr>
<tr>
<td>Our two-stage</td>
<td></td>
<td>⋆</td>
<td>80.0</td>
<td>85.7</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td>92.3</td>
<td>95.4</td>
</tr>
</tbody>
</table>

### Entailment System

<table>
<thead>
<tr>
<th></th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median of TAC 2009 systems</td>
<td></td>
<td>⋆</td>
<td>61.5</td>
</tr>
<tr>
<td>Our two-stage</td>
<td>⋆</td>
<td></td>
<td>65.0</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td>66.8</td>
</tr>
</tbody>
</table>

Experimental results
## Paraphrase Identification

<table>
<thead>
<tr>
<th>Paraphrase System</th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Experiments using (Dolan, Quirk, and Brockett 2004)</em></td>
<td></td>
<td></td>
<td>72.00</td>
</tr>
<tr>
<td>(Qiu, Kan, and Chua 2006)</td>
<td></td>
<td>⋆</td>
<td>73.86</td>
</tr>
<tr>
<td>(Das and Smith 2009)</td>
<td>⋆</td>
<td></td>
<td>76.23</td>
</tr>
<tr>
<td>(Wan, Dras, Dale, and Paris 2006)</td>
<td>⋆</td>
<td></td>
<td>76.41</td>
</tr>
<tr>
<td>Our two-stage</td>
<td></td>
<td>⋆</td>
<td>76.23</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td><strong>76.41</strong></td>
</tr>
<tr>
<td>Paraphrase System</td>
<td>Joint</td>
<td>ILP</td>
<td>Acc</td>
</tr>
<tr>
<td>----------------------------------------------------------------</td>
<td>-------</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>Paraphrase Identification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Experiments using (Dolan, Quirk, and Brockett 2004)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Qiu, Kan, and Chua 2006)</td>
<td></td>
<td></td>
<td>72.00</td>
</tr>
<tr>
<td>(Das and Smith 2009)</td>
<td>⋆</td>
<td></td>
<td>73.86</td>
</tr>
<tr>
<td>(Wan, Dras, Dale, and Paris 2006)</td>
<td></td>
<td></td>
<td>75.60</td>
</tr>
<tr>
<td>Our two-stage</td>
<td></td>
<td>⋆</td>
<td>76.23</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td><strong>76.41</strong></td>
</tr>
</tbody>
</table>
## Paraphrase Identification

<table>
<thead>
<tr>
<th>Paraphrase System</th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Experiments using (Dolan, Quirk, and Brockett 2004)</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Qiu, Kan, and Chua 2006)</td>
<td></td>
<td></td>
<td>72.00</td>
</tr>
<tr>
<td>(Das and Smith 2009)</td>
<td>⋆</td>
<td></td>
<td>73.86</td>
</tr>
<tr>
<td>(Wan, Dras, Dale, and Paris 2006)</td>
<td></td>
<td>⋆</td>
<td>75.60</td>
</tr>
<tr>
<td>Our two-stage</td>
<td></td>
<td>⋆</td>
<td>76.23</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td><strong>76.41</strong></td>
</tr>
<tr>
<td><em>Experiments using Noisy data set</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our two-stage</td>
<td></td>
<td>⋆</td>
<td>72.00</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td><strong>72.75</strong></td>
</tr>
</tbody>
</table>
Part I: Learning with Latent Structure

Input \rightarrow Complex Structural Variables \rightarrow Simple Output

Target

Part II: Learning with Indirect Supervision

Input \rightarrow Complex Structural Variables \rightarrow (Indirect) Simple Output

Target
Part I: Learning with Latent Structure

Part II: Learning with Indirect Supervision
Our Goal

- Given that supervising structures is time consuming and often requires expertise, our goal is to reduce the supervision effort for structured output learning.
- Reducing the supervision effort: A major challenge in many domains
Our Goal

- Given that supervising structures is time consuming and often requires expertise, our goal is to reduce the supervision effort for structured output learning.
- Reducing the supervision effort: A major challenge in many domains

Research Question

Is it possible to use (and gain from) additional cheap sources of supervision?
Example structured output problems

**Object Part Recognition**
Given a car image, where are the body, windows and wheels?
Example structured output problems

Object Part Recognition

Given a car image, where are the body, windows and wheels?
Example structured output problems

**Object Part Recognition**

Given a car image, where are the body, windows and wheels?

![Car Image with Part Recognition](image)

**Citation Recognition**

Example structured output problems

Object Part Recognition
Given a car image, where are the body, windows and wheels?

Citation Recognition
Task
Given a car image, where are the body, windows and wheels?
Supervising structured output problems

Task
Given a car image, where are the body, windows and wheels?
Supervising structured output problems

**Task**

Given a car image, where are the body, windows and wheels?

- Supervised Approach
Supervising structured output problems

Task
Given a car image, where are the body, windows and wheels?

- Supervised Approach is Expensive!
Task
Given a car image, where are the body, windows and wheels?

- Supervised Approach is Expensive!
Supervising structured output problems

Task
Given a car image, where are the body, windows and wheels?

- Supervised Approach is Expensive!
Supervising structured output problems

Task
Given a car image, where are the body, windows and wheels?

- Supervised Approach is Expensive!
Supervising structured output problems

Task
Given a car image, where are the body, windows and wheels?

- Supervised Approach is Expensive!
Supervising structured output problems

**Task**
Given a car image, where are the body, windows and wheels?

- Supervised Approach is Expensive!

**Indirect Supervision**
Use binary labeled data as indirect supervisions
Supervising structured output problems: Citation

Supervised Learning algorithms

OUTPUT: \( h \)

Labeled Citation

INPUT: \( x \)

Author

Author

Title

Author

Title

Lars

Ole

Andersen

Program

...
Semi-Supervised Learning algorithms

INPUT: $x$

OUTPUT: $h$

Labeled Citation:

Authors: Lars, Ole, Andersen
Title: Program

Unlabeled Citation: Positive Examples:

Authors: Ming, Wei, Chang
Title: Structured
Supervising structured output problems: Citation

Indirect Supervision algorithm

OUTPUT: \( h \)

INPUT: \( x \)

Labeled Citation

Author
Author
Author
Author
Title
Title

INPUT: \( x \)

Unlabeled Citation: Positive Examples

Ming
Wei
Chang

Program

Not a Citation: Negative Examples

Structured

Shuffling tokens of a citation entry
Key Intuition

Structured Output Task

Observation

Many structured output prediction problems have a companion binary decision problem: predicting whether an input possesses a good structure or not.

Why is this important

Binary labeled data is very easy to obtain.
Many structured output prediction problems have a companion binary decision problem: predicting whether an input possesses a good structure or not.

Why is this important?

Binary labeled data is very easy to obtain.
Key Intuition

Observation
Many structured output prediction problems have a companion binary decision problem: predicting whether an input possesses a good structure or not.
Key Intuition

Observation
Many structured output prediction problems have a companion binary decision problem: predicting whether an input possesses a good structure or not.

Why is this important
Binary labeled data is very easy to obtain
Observation

Many structured output prediction problems have a companion binary decision problem: predicting whether an input possesses a good structure or not.

Why is this important

Binary labeled data is very easy to obtain.
The role of binary labeled data

Structured Output Learning

Recognize Car parts

Companion Binary Output Problem

Is there a car in this image?

Companion Task: Does this example possess a good structure?

\[ x_1 \text{ is positive.} \]
\[ \exists h, w^T \Phi(x_1, h) \geq 0 \]

\[ x_2 \text{ is negative.} \]
\[ \forall h, w^T \Phi(x_2, h) \leq 0 \]
The role of binary labeled data

**Structured Output Learning**
- Recognize Car parts

**Companion Binary Output Problem**
- Is there a car in this image?

There must exist a good structure that justifies the positive label:

$$\exists h, w^T \Phi(x_1, h) \geq 0$$

$$x_2$$ is negative.

No structure is good enough,

$$\forall h, w^T \Phi(x_2, h) \leq 0$$
The role of binary labeled data

**Structured Output Learning**
- Recognize Car parts

**Companion Binary Output Problem**
- Is there a car in this image?

**Companion Task:** Does this example possess a good structure?
Structured Output Learning
- Recognize Car parts

Companion Binary Output Problem
- Is there a car in this image?

Companion Task: Does this example possess a good structure?

- $x_1$ is positive.
  - There must exist a good structure that justifies the positive label
  - $\exists h, w^T \Phi(x_1, h) \geq 0$
The role of binary labeled data

**Structured Output Learning**
- Recognize Car parts

**Companion Binary Output Problem**
- Is there a car in this image?

![Car Image](image1)

**Companion Task:** Does this example possess a good structure?

- \( x_1 \) is positive.
  - There must exist a good structure that justifies the positive label
  - \( \exists h, w^T \Phi(x_1, h) \geq 0 \)

- \( x_2 \) is negative.
  - No structure is good enough, \( \forall h, w^T \Phi(x_2, h) \leq 0 \)
Why is binary labeled data useful?

- **\(x_1\) is positive**: There exists a good structure
  - \(\exists h, \mathbf{w}^T \Phi(x_1, h) \geq 0\), or \(\max_h \mathbf{w}^T \Phi(x_1, h) \geq 0\)

- **\(x_2\) is negative**: No structure is good enough
  - \(\forall h, \mathbf{w}^T \Phi(x_2, h) \leq 0\), or \(\max_h \mathbf{w}^T \Phi(x_2, h) \leq 0\)
Why is binary labeled data useful?

\[ \{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \} \]

- **\( x_1 \) is positive**: There exists a good structure
  \[ \exists h, w^T \Phi(x_1, h) \geq 0, \text{ or } \max_h w^T \Phi(x_1, h) \geq 0 \]

- **\( x_2 \) is negative**: No structure is good enough
  \[ \forall h, w^T \Phi(x_2, h) \leq 0, \text{ or } \max_h w^T \Phi(x_2, h) \leq 0 \]
Why is binary labeled data useful?

Supervised Model: $\mathbf{w}$

$$\{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \}$$

- **$x_1$ is positive**: There exists a good structure
  - $\exists h, \mathbf{w}^T \Phi(x_1, h) \geq 0$, or $\max_h \mathbf{w}^T \Phi(x_1, h) \geq 0$

- **$x_2$ is negative**: No structure is good enough
  - $\forall h, \mathbf{w}^T \Phi(x_2, h) \leq 0$, or $\max_h \mathbf{w}^T \Phi(x_2, h) \leq 0$
Why is binary labeled data useful?

Supervised Model: \( w \)

\[
\{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \}
\]

Predict: \( \Phi(x_1, \hat{h}) \)

- **\( x_1 \) is positive**: There exists a good structure
  - \( \exists h, w^T \Phi(x_1, h) \geq 0 \), or \( \max_h w^T \Phi(x_1, h) \geq 0 \)
- **\( x_2 \) is negative**: No structure is good enough
  - \( \forall h, w^T \Phi(x_2, h) \leq 0 \), or \( \max_h w^T \Phi(x_2, h) \leq 0 \)
Why is binary labeled data useful?

Supervised Model: $w$

Gold: $\Phi(x_1, h^*_1)$

Predict: $\Phi(x_1, \hat{h})$

- $x_1$ is positive: There exists a good structure
  - $\exists h, w^T \Phi(x_1, h) \geq 0$, or $\max_h w^T \Phi(x_1, h) \geq 0$

- $x_2$ is negative: No structure is good enough
  - $\forall h, w^T \Phi(x_2, h) \leq 0$, or $\max_h w^T \Phi(x_2, h) \leq 0$
Why is binary labeled data useful?

Supervised Model: $w$

Gold: $\Phi(x_1, h^*_1)$

Predict: $\Phi(x_1, \hat{h})$

$x_1$ is positive: There exists a good structure

$\exists h, w^T \Phi(x_1, h) \geq 0$, or $\max_h w^T \Phi(x_1, h) \geq 0$

$x_2$ is negative: No structure is good enough

$\forall h, w^T \Phi(x_2, h) \leq 0$, or $\max_h w^T \Phi(x_2, h) \leq 0$
Why is binary labeled data useful?

Supervised Model: \( w \)

\[
\{ \Phi(x_1, h) \mid h \in H(x_1) \}
\]

Gold: \( \Phi(x_1, h^*_1) \)

Predict: \( \Phi(x_1, \hat{h}) \)

\[
\{ \Phi(x_2, h) \mid h \in H(x_2) \}
\]

- **\( x_1 \) is positive**: There exists a good structure
  \[
  \exists h, \ w^T \Phi(x_1, h) \geq 0, \text{ or } \max_h w^T \Phi(x_1, h) \geq 0
  \]

- **\( x_2 \) is negative**: No structure is good enough
  \[
  \forall h, \ w^T \Phi(x_2, h) \leq 0, \text{ or } \max_h w^T \Phi(x_2, h) \leq 0
  \]
Why is binary labeled data useful?

Supervised Model:

\[ w \text{: +Indirect Supervision} \]

\[
\Phi(x_1, h^*) \quad \text{Gold:} \\
\Phi(x_1, \hat{h}) \quad \text{Predict:}
\]

\[
\{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \}
\]

\[
\{ \Phi(x_2, h) \mid h \in \mathcal{H}(x_2) \}
\]

- \( x_1 \) is positive: There exists a good structure
  \[ \exists h, w^T \Phi(x_1, h) \geq 0, \text{ or } \max_h w^T \Phi(x_1, h) \geq 0 \]

- \( x_2 \) is negative: No structure is good enough
  \[ \forall h, w^T \Phi(x_2, h) \leq 0, \text{ or } \max_h w^T \Phi(x_2, h) \leq 0 \]
Why is binary labeled data useful?

\[ \{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \} \]

\[ \{ \Phi(x_2, h) \mid h \in \mathcal{H}(x_2) \} \]

\[ \text{Gold:} \Phi(x_1, h_1^*) \]

\[ \text{Predict:} \Phi(x_1, \hat{h}) \]

\[ w: +\text{Indirect Supervision} \]

\[ x_1 \text{ is positive: There exists a good structure} \]

\[ \exists h, w^T \Phi(x_1, h) \geq 0, \text{ or } \max_h w^T \Phi(x_1, h) \geq 0 \]

\[ x_2 \text{ is negative: No structure is good enough} \]

\[ \forall h, w^T \Phi(x_2, h) \leq 0, \text{ or } \max_h w^T \Phi(x_2, h) \leq 0 \]
Joint Learning with Indirect Supervision [ICML’10]

\[
\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B} L_B(x_i, z_i, w),
\]

- **Regularization**: measures the model complexity
- **Direct Supervision**: structured labeled data \( S = \{(x, h)\} \)
- **Indirect Supervision**: binary labeled data \( B = \{(x, z)\} \)
Joint Learning with Indirect Supervision [ICML’10]

\[
\min_w \frac{1}{2} \|w\|^2 + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B} L_B(x_i, z_i, w),
\]

- Regularization: measures the model complexity
- Direct Supervision: structured labeled data \( S = \{(x, h)\} \)
- Indirect Supervision: binary labeled data \( B = \{(x, z)\} \)
Joint Learning with Indirect Supervision [ICML’10]

\[
\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B} L_B(x_i, z_i, w),
\]

- **Regularization**: measures the model complexity
- **Direct Supervision**: structured labeled data \( S = \{(x, h)\} \)
- **Indirect Supervision**: binary labeled data \( B = \{(x, z)\} \)

**Share weight vector \( w \)**

Use the same weight vector for both structured labeled data and binary labeled data.
Joint Learning with Indirect Supervision [ICML’10]

\[
\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B} L_B(x_i, z_i, w),
\]

- **Regularization**: measures the model complexity
- **Direct Supervision**: structured labeled data \( S = \{(x, h)\} \)
- **Indirect Supervision**: binary labeled data \( B = \{(x, z)\} \)

**Share weight vector \( w \)**

Use the same weight vector for both structured labeled data and binary labeled data.

**Support Structured SVM**
Experimental Results

PA : Phonetic Alignment
ADS : Advertisement field recognition

Tasks

Accuracy

PA : 70
POS : 80
Citation : 60
ADS :

Structural SVM
Joint Learning with Indirect Supervision
Experimental Results

Tasks

Accuracy

PA : Phonetic Alignment

ADS : Advertisement field recognition

Structural SVM | Joint Learning with Indirect Supervision
Impact of negative examples

- J-LIS: takes advantage of both positively and negatively labeled data
- J-LIS: takes advantage of both positively and negatively labeled data

![Graph showing accuracy of Structural SVM and J-LIS](image)
J-LIS: takes advantage of both positively and negatively labeled data

Impact of negative examples

Number of tokens in the negative examples

Accuracy

Structural SVM

JLIS
Recent publications about indirect supervisions

- User Response as Indirect Supervisions
  - Application: Mapping natural language into logical forms
    - (Clarke, Goldwasser, Chang, and Roth 2010; Liang, Jordan, and Klein 2011)

- Constraints as Indirect Supervisions
  - Applications: Word Alignment, Dependency Parsing, Information Extraction
    - (Chang, Ratinov, and Roth 2007; Mann and McCallum 2008; Ganchev, Grača, Gillenwater, and Taskar 2010; Carlson, Betteridge, Wang, Jr., and Mitchell 2010)
Recent publications about indirect supervisions

- **User Response as Indirect Supervisions**
  - Application: Mapping natural language into logical forms
    - Clarke, Goldwasser, Chang, and Roth (2010; Liang, Jordan, and Klein 2011)

- **Constraints as Indirect Supervisions**
  - Applications: Word Alignment, Dependency Parsing, Information Extraction
    - Chang, Ratinov, and Roth (2007; Mann and McCallum 2008; Ganchev, Grača, Gillenwater, and Taskar 2010; Carlson, Betteridge, Wang, Jr., and Mitchell 2010)
Recent publications about indirect supervisions

**User Response as Indirect Supervisions**

- Application: Mapping natural language into logical forms
- (Clarke, Goldwasser, Chang, and Roth 2010; Liang, Jordan, and Klein 2011)

**Constraints as Indirect Supervisions**

- Applications: Word Alignment, Dependency Parsing, Information Extraction
- (Chang, Ratinov, and Roth 2007; Mann and McCallum 2008; Ganchev, Graça, Gillenwater, and Taskar 2010; Carlson, Betteridge, Wang, Jr., and Mitchell 2010)
**Conclusion**

**Target: Binary Output Variables**
- We can find intermediate representations that help the binary decisions the most!
- Use Integer Linear Programming: Easy to apply to a new task

**Target: Complex Structural Variables**
- We can invent easy output problems to supervise the model
- We have a framework that can accept both direct and indirect supervision signals
- The use of negative examples is important

**General Indirect Supervision**
- It is possible to invent new indirect supervision signals
- It has been shown to be useful in many applications
Conclusion

**Target: Binary Output Variables**
- We can find intermediate representations that help the binary decisions the most!
- Use Integer Linear Programming: Easy to apply to a new task

**Target: Complex Structural Variables**
- We can invent easy output problems to supervise the model
- We have a framework that can accept both direct and indirect supervision signals
- The use of negative examples is important

**General Indirect Supervision**
- It is possible to invent new indirect supervision signals
- It has been shown to be useful in many applications

Thank you!
Example: Transliteration

It a l y

איטליה
Example: Transliteration

**Italy**

איטליה

**Structured Output Learning**

Given one English NE and its Hebrew transliteration, tell me what are the phonetic alignments?
Example: Transliteration

**Structured Output Learning**

Given one English NE and its Hebrew transliteration, tell me what are the phonetic alignments?
Example: Transliteration

Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what are the phonetic alignments?
Example: Transliteration

Structured Output Learning
Given one English NE and its Hebrew transliteration, tell me what are the phonetic alignments?

Companion Binary Output Problem
Are these two NEs a transliteration pair?
Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what are the phonetic alignments?

Is there any connection between these two problems?

Companion Binary Output Problem

Are these two NEs a transliteration pair?
Structured Output Learning
Given one English NE and its Hebrew transliteration, tell me what are the phonetic alignments?

Companion Binary Output Problem
Are these two NEs a transliteration pair?

Relationships
- Only a transliteration pair can have good phonetic alignment!
- Non-transliteration pairs cannot have good phonetic alignment!
Coupled semi-supervised learning for information extraction.
In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*.

Chang, M., L. Ratinov, and D. Roth (2007).
Guiding semi-supervision with constraint-driven learning.
In *ACL*.

Driving semantic parsing from the world’s response.
In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010)*.

Paraphrase identification as probabilistic quasi-synchronous recognition.
In *ACL*.

Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.
In COLING.


A linear programming formulation for global inference in natural language tasks.
In H. T. Ng and E. Riloff (Eds.), CoNLL.

In Proc. of the Australasian Language Technology Workshop (ALTW).