Structured Output Learning with Indirect Supervision

Ming-Wei Chang, Vivek Srikumar, Dan Goldwasser and Dan Roth

Computer Science Department, University of Illinois at Urbana-Champaign
**Example**

**Input:** A Sentence, **Output:** Its Part-Of-Speech Tags

### OUTPUT: h
- JJ
- NN
- NN
- VBZ
- ADJ

### INPUT: x
- Natural language processing
- is
- fun
Review: structured output prediction

Example

Input: A Sentence, Output: Its Part-Of-Speech Tags

OUTPUT: h JJ NN NN VBZ ADJ
INPUT: x Natural language processing is fun

Properties of Structured Output Prediction

- Many interdependent decisions. **Expensive to label**
- Exponential number of structures for a given input
- Many important tasks in NLP, Computer Vision and other domains are structured output prediction tasks
Notation

**OUTPUT:** h

**INPUT:** x

Natural language processing is fun

Training model with feature vector \( \Phi(x, h) \)

Key idea: learn a scoring function over \((x, h)\) pairs

Scoring function:

\[
 w^T \Phi(x, h)
\]

Inference based prediction

Given \( x \), find \( h \) that maximizes the score

\[
 \arg \max_{h \in H(x)} w^T \Phi(x, h)
\]

\( H(x) \): A set of all possible structures for an example
Natural language processing is fun.

Training

- model \( \mathbf{w} \), feature vector \( \Phi(x, h) \)
- Key idea: learn a **scoring** function over \((x, h)\) pairs
- Scoring function: \( \mathbf{w}^T \Phi(x, h) \)
Natural language processing is fun.

Training

- model $w$, feature vector $\Phi(x, h)$
- Key idea: learn a **scoring** function over $(x, h)$ pairs
- Scoring function: $w^T \Phi(x, h)$

Inference based prediction

- Given $x$, find $h$ that maximizes the score

$$\arg \max_{h \in \mathcal{H}(x)} w^T \Phi(x, h)$$

- $\mathcal{H}(x)$: A set of all possible structures for an example $x$. 
Motivation

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
Motivation

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?
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?
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!
Supervising structured output problems

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

- Supervised Approach is **Expensive**!
- Semi-Supervised Approach
Supervising structured output problems

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

- Supervised Approach is Expensive!
- Semi-Supervised Approach
Supervising structured output problems

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

- Supervised Approach is Expensive!
- Semi-Supervised Approach
Supervising structured output problems

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

- Supervised Approach is Expensive!
- Semi-Supervised Approach
Supervising structured output problems

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

- Supervised Approach is **Expensive**!
- Semi-Supervised Approach **ignores invalid data**!
Supervising structured output problems

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

- Supervised Approach is Expensive!
- Semi-Supervised Approach ignores invalid data!

Can we use invalid data to improve the model?
Supervising structured output problems

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

- Supervised Approach is **Expensive**!
- Semi-Supervised Approach ignores invalid data!

Can we use invalid data to improve the model?
Outline

1. Motivation
2. Structured Output Prediction and Its Companion Task
3. Joint Learning with Indirect Supervision
4. Optimization
5. Experiments
Outline

1 Motivation

2 Structured Output Prediction and Its Companion Task

3 Joint Learning with Indirect Supervision

4 Optimization

5 Experiments
Example: Object Part Recognition

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

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

Only a car image can contain car parts in the right position!
A non-car image cannot have the car parts in the right position.
Example: Object Part Recognition

Structured Output Learning

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

Structured Output Learning

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

Structured Output Learning

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

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

Companion Binary Output Problem
Is there a car in this image?
Example: Object Part Recognition

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

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

Is there any connection between these two problems?
Example: Object Part Recognition

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

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

- Only a car image can contain car parts in the right position!
- A non-car image cannot have the car parts in the right position
Example: Phonetic Alignment

Italy

איטליה
Example: Phonetic Alignment

It a l y

איטליה

Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?
Example: Phonetic Alignment

Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?
Example: Phonetic Alignment

Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?

Italy

Israel

Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?
Example: Phonetic Alignment

I t a l y

Israel

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

Companion Binary Output Problem
Are these two NEs a transliteration pair?
Example: Phonetic Alignment

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

Israel
Yes/No

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

Is there any connection between these two problems?
Example: Phonetic Alignment

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

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!
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.
Key Intuition

Structured Output Task

Companion Binary 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.
Observation

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

Structured Output Task

Companion Binary Task

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

Why is this important
Binary labeled data is very easy to obtain
Key Intuition

How to exploit it???

Structured Output Task → Companion Binary Task

Observation

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

Why is this important

Binary labeled data is very easy to obtain
Geometric Interpretation for SSVM

Decision Function

\[
\arg \max_{\mathbf{h} \in \mathcal{H}(x_i)} \mathbf{w}^T \Phi(x_i, \mathbf{h})
\]

Training: Intuition

Given an example \((x_i, h_i)\), find a \(\mathbf{w}\) such that the gold structure \(h_i\) has the highest score!
Geometric Interpretation for SSVM

Decision Function
\[
\arg \max_{h \in \mathcal{H}(x_i)} w^T \Phi(x_i, h)
\]

Training: Intuition
Given an example \((x_i, h_i)\), find a \(w\) such that the gold structure \(h_i\) has the highest score!

\[
\{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \}
\]
Geometric Interpretation for SSVM

**Decision Function**

\[
\text{arg max}_{h \in H(x_i)} w^T \Phi(x_i, h)
\]

**Training: Intuition**

Given an example \((x_i, h_i)\), find a \(w\) such that the gold structure \(h_i\) has the highest score!
**Decision Function**

\[
\arg \max_{h \in \mathcal{H}(x_i)} w^T \Phi(x_i, h)
\]

**Training: Intuition**

Given an example \((x_i, h_i)\), find a \(w\) such that the gold structure \(h_i\) has the highest score!
Geometric Interpretation for SSVM

**Decision Function**
\[
\arg \max_{h \in \mathcal{H}(x_i)} w^T \Phi(x_i, h)
\]

**Training: Intuition**
Given an example \((x_i, h_i)\), find a \(w\) such that the gold structure \(h_i\) has the highest score!

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

\[
\{ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \}
\]
Geometric Interpretation for SSVM

Decision Function

\[ \arg \max_{h \in \mathcal{H}(x_i)} w^T \Phi(x_i, h) \]

Training: Intuition

Given an example \((x_i, h_i)\), find a \(w\) such that the gold structure \(h_i\) has the highest score!

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

\[ \Phi(x_1, h_1^*) \]
Structural SVM

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

- **Regularization**: Measures the model complexity
- **Structural Loss**: 
  - \( S \) is the set of structured labeled examples: 
  - \( L_S(x_i, h_i, w) \): Measures “the distance” between the current best prediction and the gold structure \( h_i \)
  - \( L_S \) can use hinge or square hinge functions or others
  - A convex optimization problem
\[
\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w)
\]

- **Regularization**: Measures the model complexity

- **Structural Loss**:  
  - \(S\) is the set of *structured* labeled examples:  
  - \(L_S(x_i, h_i, w)\): Measures “the distance” between the current best prediction and the gold structure \(h_i\)
  - \(L_S\) can use hinge or square hinge functions or others
  - A convex optimization problem

Now, add supervision from the companion task!
The role of binary labeled data

Structured Output Problem

I t a l y

Companion Binary Output Problem

A r s e n i e

Israel

Yes/No

A z i l i n o v
The role of binary labeled data

Structured Output Problem

It a l y

Companion Binary Output Problem

Israel

Yes/No

Companion Task: Does this example possess a good structure?
The role of binary labeled data

Structured Output Problem

Companion Binary Output Problem

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 Problem

Italy

Companion Binary Output Problem

Israel

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?

- **x₁ is positive**: There exists a good structure
  \[ \exists h, w^T \Phi(x_1, h) \geq 0, \text{ or } \max_h w^T \Phi(x_1, h) \geq 0 \]

- **x₂ 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) \}

- **\(x_1\) 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?

\[ \Phi(x_1, h) \mid h \in \mathcal{H}(x_1) \]
Why is binary labeled data useful?

SSVM: $w$

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

**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?

**SSVM:** $w$

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

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

**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?

SSVM: $w$

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

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?

SSVM: \( w \)

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

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

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

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

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

- $\Phi(x_1, h)$, $h \in H(x_1)$
- $\Phi(x_2, h)$, $h \in H(x_2)$

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

**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 $\max_h w^T \Phi(x_2, h) \leq 0$

$w$: SSVM + Indirect Supervision
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) \}

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

\[ \text{Gold: } \Phi(x_1, h^*_1) \]

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

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

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

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

\[ \forall h, w^T \Phi(x_2, h) \leq 0, \text{ or } \max_h w^T \Phi(x_2, h) \leq 0 \]
Outline

1 Motivation

2 Structured Output Prediction and Its Companion Task

3 Joint Learning with Indirect Supervision

4 Optimization

5 Experiments
Binary and structured labeled data

**Direct Supervision: \( S \)**
- **Target Task**

**Indirect Supervision: \( B \)**
- **Companion Task**

**Goal:**
- For **Direct Supervision**: \( w^T \Phi(x, h) \geq \max_{h \in H} (x, h) \)
- For **Indirect Supervision**: \( y_i \max_{h \in H} (x, h) \geq 0 \)

Both \( L_S \) and \( L_B \) can use hinge, square-hinge, logistic, ...
Binary and structured labeled data

Direct Supervision: $S$

- **Target Task**
- An example: $(x_i, h_i)$

Indirect Supervision: $B$

- **Companion Task**
- An example: $(x_i, y_i)$
Binary and structured labeled data

### Direct Supervision: $S$

#### Target Task
- An example: $(x_i, h_i)$
- Goal:

$$w^T \Phi(x_i, h_i) \geq \max_{h \in \mathcal{H}(x_i)} w^T \Phi(x_i, h).$$

### Indirect Supervision: $B$

#### Companion Task
- An example: $(x_i, y_i)$
- Goal:

$$y_i \max_{h \in \mathcal{H}(x_i)} w^T \Phi(x_i, h) \geq 0.$$
Direct Supervision: \( S \)

- **Target Task**
- An example: \((x_i, h_i)\)
- Goal:

\[
\mathbf{w}^T \Phi(x_i, h_i) \geq \max_{h \in \mathcal{H}(x_i)} \mathbf{w}^T \Phi(x_i, h).
\]

- Structural Loss: \( L_S \)

Indirect Supervision: \( B \)

- **Companion Task**
- An example: \((x_i, y_i)\)
- Goal:

\[
y_i \max_{h \in \mathcal{H}(x_i)} \mathbf{w}^T \Phi(x_i, h) \geq 0
\]

- Binary Loss: \( L_B \)
Binary and structured labeled data

Direct Supervision: $S$

- **Target Task**
  - An example: $(x_i, h_i)$
  - Goal:
    \[ w^T \phi(x_i, h_i) \geq \max_{h \in \mathcal{H}(x_i)} w^T \phi(x_i, h). \]
  - Structural Loss: $L_S$

Indirect Supervision: $B$

- **Companion Task**
  - An example: $(x_i, y_i)$
  - Goal:
    \[ y_i \max_{h \in \mathcal{H}(x_i)} w^T \phi(x_i, h) \geq 0 \]
  - Binary Loss: $L_B$

Both $L_S$ and $L_B$ can use hinge, square-hinge, logistic, ...
Joint Learning with Indirect Supervision

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

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

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

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

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

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

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

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

1 Motivation

2 Structured Output Prediction and Its Companion Task

3 Joint Learning with Indirect Supervision

4 Optimization

5 Experiments
Convexity Properties

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

\[
L_S(x_i, h_i, w) = \ell \left( \max_h (\Delta(h, h_i) - w^T \Phi(x_i, h_i) + w^T \Phi(x_i, h)) \right) \quad (1)
\]

\[
L_B(x_i, y_i, w) = \ell \left( 1 - y_i \max_{h \in \mathcal{H}(x)} (w^T \Phi_B(x_i, h)) \right) \quad (2)
\]
Convexity Properties

Convex Parts

Regularization, Direct Supervision, Negative Data $B^-$

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

Neither convex nor concave

Positive Data $B^+$
JLIS: optimization procedure

**Algorithm**

1: Find the best structures for **positive examples**

2: Find the weight vector using the structure found in Step 1.
   - Still need to do **inference** for structured examples and negative examples

3: Repeat!
JLIS: optimization procedure

Algorithm

1: Find the best structures for **positive examples**
2: Find the weight vector using the structure found in Step 1.
   - Still need to do **inference** for structured examples and negative examples
3: Repeat!

This algorithm converges when $\ell$ is monotonically increasing and convex.
JLIS: optimization procedure

Algorithm

1: Find the best structures for **positive examples**

2: Find the weight vector using the structure found in Step 1.
   - Still need to do **inference** for structured examples and negative examples

3: Repeat!

This algorithm converges when $\ell$ is monotonically increasing and convex.

Properties of the algorithm: Asymmetric nature

- Converting a non-convex problem into a series of smaller convex problems
- Inference allows incorporating constraints on the output space. (Chang, Goldwasser, Roth, and Srikumar NAACL 2010)
Solving the convex sub-problem

\[
\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B^-} L_B(x_i, y_i, w) + C_2 \sum_{i \in B^+} L_B(x_i, y_i, w)
\]
Solving the convex sub-problem

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

with fixed structures
Solving the convex sub-problem

\[
\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B^-} L_B(x_i, y_i, w) + C_2 \sum_{i \in B^+} L_B(x_i, y_i, w) \text{ with fixed structures}
\]

**Cutting plane method**

- Find the “best structure” for examples in $S$ and $B^-$ with the current $w$
- Add chosen structure into the cache and solve it again!
Solving the convex sub-problem

\[
\min_w \frac{\|w\|^2}{2} + C_1 \sum_{i \in S} L_S(x_i, h_i, w) + C_2 \sum_{i \in B^-} L_B(x_i, y_i, w) + C_2 \sum_{i \in B^+} L_B(x_i, y_i, w) \quad \text{with fixed structures}
\]

**Cutting plane method**
- Find the “best structure” for examples in \( S \) and \( B^- \) with the current \( w \)
- Add chosen structure into the cache and solve it again!

**Dual coordinate descent method**
- Simple implementation with square (L2) hinge loss
Outline

1 Motivation

2 Structured Output Prediction and Its Companion Task

3 Joint Learning with Indirect Supervision

4 Optimization

5 Experiments
Experimental Setting

Tasks

- **Task 1**: Phonetic alignment
- **Task 2**: Part-of-speech Tagging
- **Task 3**: Information Extraction
  - Citation recognition
  - Advertisement field recognition

Companion Tasks

- **Phonetic alignment**: Transliteration pair or not
- **POS Tagging**: Has a legitimate POS tag sequence or not
- **IE**: Is a legitimate Citation/Advertisement or not
Experimental Results

Tasks

- PA: Phonetic Alignment
- POS
- Citation
- ADS: Advertisement field recognition

Accuracy

- Structural SVM
- Joint Learning with Indirect Supervision
Experimental Results

### Accuracy of Tasks

**Tasks**
- **PA**: Phonetic Alignment
- **ADS**: Advertisement field recognition

**Comparison Methods**
- **Blue**: Structural SVM
- **Red**: Joint Learning with Indirect Supervision

<table>
<thead>
<tr>
<th>Task</th>
<th>Accuracy (Structural SVM)</th>
<th>Accuracy (Joint Learning)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>POS</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>Citation</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>ADS</td>
<td>50</td>
<td>40</td>
</tr>
</tbody>
</table>
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.
J-LIS: takes advantage of both positively and negatively labeled data.
Comparison to other learning framework

**Generalization over several frameworks**

- \( B = \emptyset \Rightarrow \text{Structured SVM} \) (Tsochantaridis, Hofmann, Joachims, and Altun 2004)
- \( S = \emptyset \Rightarrow \text{Latent SVM/LR} \) (Felzenszwalb, Girshick, McAllester, and Ramanan 2009) (Chang, Goldwasser, Roth, and Srikumar NAACL 2010)
Comparison to other learning framework

### Generalization over several frameworks

- $B = \emptyset \Rightarrow$ Structured SVM (Tsochantaridis, Hofmann, Joachims, and Altun 2004)
- $S = \emptyset \Rightarrow$ Latent SVM/LR (Felzenszwalb, Girshick, McAllester, and Ramanan 2009) (Chang, Goldwasser, Roth, and Srikumar NAACL 2010)

### Semi-Supervised Learning methods

- (Zien, Brefeld, and Scheffer 2007): Transductive Structural SSVM, (Brefeld and Scheffer 2006): co-Structural SVM
- J-LIS uses “negative” examples
Comparison to other learning framework

**Generalization over several frameworks**

- $B = \emptyset \Rightarrow$ Structured SVM (Tsochantaridis, Hofmann, Joachims, and Altun 2004)
- $S = \emptyset \Rightarrow$ Latent SVM/LR (Felzenszwalb, Girshick, McAllester, and Ramanan 2009) (Chang, Goldwasser, Roth, and Srikumar NAACL 2010)

**Semi-Supervised Learning methods**

- (Zien, Brefeld, and Scheffer 2007): Transductive Structural SSVM,
  (Brefeld and Scheffer 2006): co-Structural SVM
- J-LIS uses “negative” examples

**Compared to Contrastive Estimation**

- Conceptually related. [More discussion]
Conclusions

- **It is possible to use binary labeled data for learning structures!**
- **J-LIS**: gains from **both** direct and indirect supervision
- Similarly, structured labeled data can help the binary task
- Allows the use of constraints on structures
Conclusions

- It is possible to use binary labeled data for learning structures!
- J-LIS: gains from both direct and indirect supervision
- Similarly, structured labeled data can help the binary task
- Allows the use of constraints on structures

Many exciting new directions!

- Using existing labeled dataset as structured task supervisions
- How to generate good “negative” examples?
- Other forms of indirect supervision?
Thank you!

Our learning code is available: the JLIS package

http://l2r.cs.uiuc.edu/~cogcomp/software.php
Compared to Contrastive Estimation: I

Contrastive Estimation

- Performing unsupervised learning with log-linear models
- Maximize log $P(x)$
- Model 1

$$P(x) = \frac{\sum_h \exp(\mathbf{w}^T \Phi(x, h))}{\sum_{h, \hat{x}} \exp(\mathbf{w}^T \Phi(\hat{x}, h))}$$

- CE

$$P(x) = \frac{\sum_h \exp(\mathbf{w}^T \Phi(x, h))}{\sum_{h, \hat{x} \in \mathcal{N}(x)} \exp(\mathbf{w}^T \Phi(\hat{x}, h))}$$
Compared to Contrastive Estimation: II

\[ P(x) = \frac{\sum_h \exp(w^T \Phi(x, h))}{\sum_{h, \hat{x} \in \mathcal{N}(x)} \exp(w^T \Phi(\hat{x}, h))} \]

<table>
<thead>
<tr>
<th></th>
<th>CE</th>
<th>J-LIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervision type</td>
<td>Neighbors</td>
<td>Structured + Binary</td>
</tr>
<tr>
<td>Inference Problem</td>
<td>sum max</td>
<td></td>
</tr>
<tr>
<td>Property</td>
<td>Can use existing data</td>
<td>J-LIS can use existing binary labeled data.</td>
</tr>
</tbody>
</table>

Compared J-LIS and CE without using labeled data

Jump Back

Part-of-speech tags experiments. Same features and dataset.

Random Base line: 35%
EM: 60.9% (62.1%), CE: 74.7% (79.0%)
J-LIS: 70.1%. J-LIS + 5 labeled example: 79.1%
Compared to Contrastive Estimation: II

\[ P(x) = \frac{\sum_h \exp(w^T \Phi(x, h))}{\sum_{h, \hat{x} \in \mathcal{N}(x)} \exp(w^T \Phi(\hat{x}, h))} \]

<table>
<thead>
<tr>
<th>Supervision type</th>
<th>CE</th>
<th>J-LIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Neighbors&quot;</td>
<td></td>
<td>Structured + Binary</td>
</tr>
</tbody>
</table>

Inference Problem

Supervision type

Can use existing data
CE needs to know the relationship between "neighbors" of the input \( x \).
J-LIS can use existing binary labeled data.

Compared J-LIS and CE

Without using labeled data

Random Baseline: 35%
EM: 60.9% (62.1%), CE: 74.7% (79.0%)
J-LIS: 70.1%
J-LIS + 5 labeled example: 79.1%
Compared to Contrastive Estimation: II

\[ P(x) = \frac{\sum_h \exp(w^T \Phi(x, h))}{\sum_{h, \hat{x} \in \mathcal{N}(x)} \exp(w^T \Phi(\hat{x}, h))} \]

<table>
<thead>
<tr>
<th>Supervision type</th>
<th>CE</th>
<th>J-LIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Neighbors”</td>
<td>sum</td>
<td>Structured + Binary</td>
</tr>
</tbody>
</table>

Jump Back

Part-of-speech tags experiments. Same features and dataset.

Random Base line: 35%
EM: 60.9% (62.1%), CE: 74.7% (79.0%), J-LIS: 70.1%, J-LIS + 5 labeled example: 79.1%
Compared to Contrastive Estimation: II

\[ P(x) = \frac{\sum_h \exp(w^T \Phi(x, h))}{\sum_{h, \hat{x} \in N(x)} \exp(w^T \Phi(\hat{x}, h))} \]

<table>
<thead>
<tr>
<th>Supervision type</th>
<th>CE</th>
<th>J-LIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Neighbors”</td>
<td>Structured + Binary</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference Problem</th>
<th>CE</th>
<th>J-LIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>sum</td>
<td></td>
<td>max</td>
</tr>
</tbody>
</table>

- CE needs to know the relationship between “neighbors” of the input \( x \).
- J-LIS can use existing binary labeled data.
Compared to Contrastive Estimation: II

\[ P(x) = \frac{\sum_h \exp(w^T \Phi(x, h))}{\sum_{h, \hat{x} \in \mathcal{N}(x)} \exp(w^T \Phi(\hat{x}, h))} \]

<table>
<thead>
<tr>
<th>Supervision type</th>
<th>CE</th>
<th>J-LIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inference Problem</td>
<td>sum</td>
<td>max</td>
</tr>
<tr>
<td>Property</td>
<td></td>
<td>Can use existing data</td>
</tr>
</tbody>
</table>

- CE needs to know the relationship between “neighbors” of the input \( x \). J-LIS can use existing binary labeled data.

### Compared J-LIS and CE without using labeled data

- Part-of-speech tags experiments. Same features and dataset.
- Random Base line: 35%
- EM: 60.9% (62.1%), CE: 74.7% (79.0%)
- J-LIS: 70.1%. J-LIS + 5 labeled example: 79.1%
Impact of structure labeled data when binary classification is our target. Results (for transliteration identification) show that joint training of direct and indirect supervision significantly improves performance, especially when direct supervision is scarce.
