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

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

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

Notation

OUTPUT: $\mathbf{h}$	JJ	NN	NN	VBZ	ADJ
INPUT: $\mathbf{x}$	Natural	language	processing	is	fun

Training

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

Notation

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

Training

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

Inference based prediction

- Given $\mathbf{x}$, find $\mathbf{h}$ that maximizes the score

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

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

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

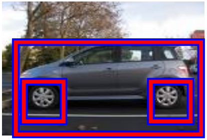
Is it possible to use (and gain from) **additional cheap** sources of supervision?

Supervising structured output problems

Task

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

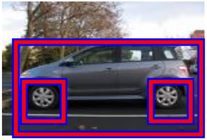
Supervising structured output problems

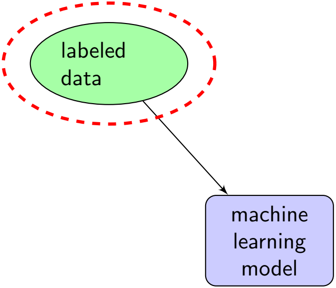


Task

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

Supervising structured output problems



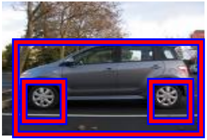


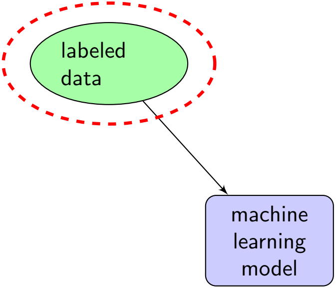
Task

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

- Supervised Approach

Supervising structured output problems



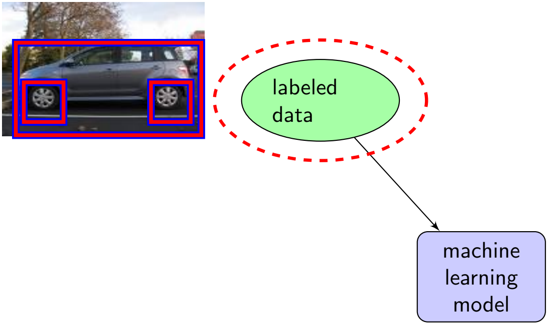


Task

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

- Supervised Approach is **Expensive!**

Supervising structured output problems

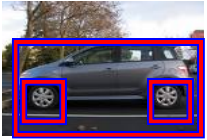


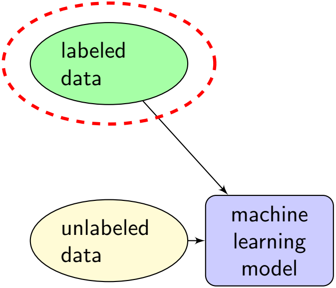
Task

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

- Supervised Approach is **Expensive!**
- Semi-Supervised Approach

Supervising structured output problems



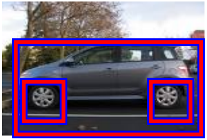


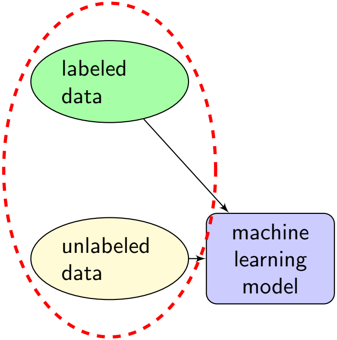
Task

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

- Supervised Approach is **Expensive!**
- Semi-Supervised Approach

Supervising structured output problems



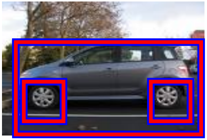


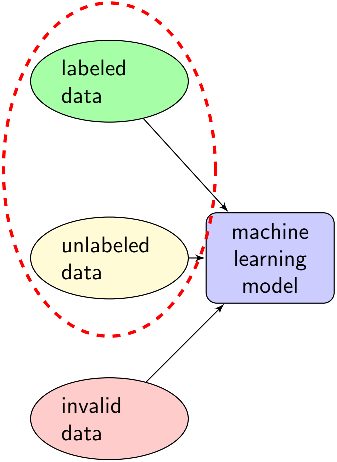
Task

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

- Supervised Approach is **Expensive!**
- Semi-Supervised Approach

Supervising structured output problems



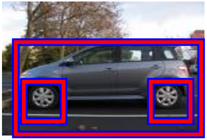


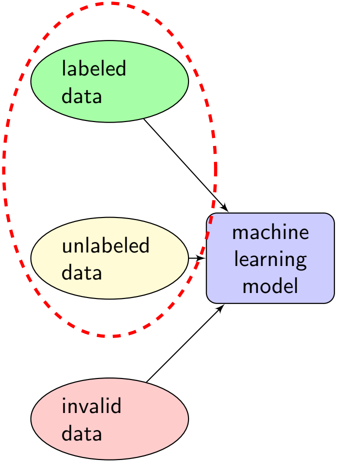
Task

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

- Supervised Approach is **Expensive!**
- Semi-Supervised Approach

Supervising structured output problems



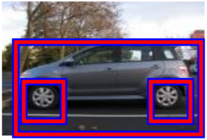


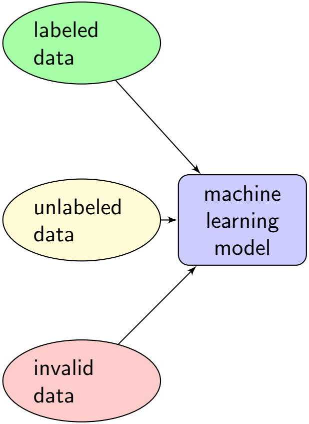
Task

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





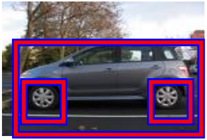
Task

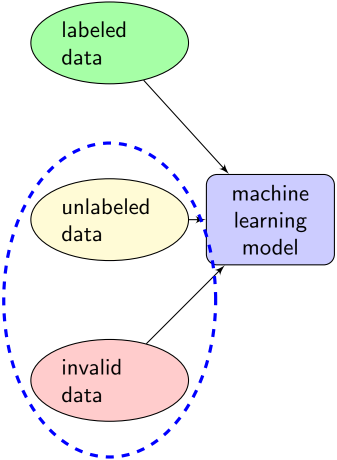
Given a car image, where are the body, windows and wheels?

- 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

Outline

1. Motivation
2. Structured Output Prediction and Its Companion Task
3. **J**oint **L**earning with **I**ndirect **S**upervision
4. Optimization
5. Experiments

Example: Object Part Recognition

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

I t a l y

איטליה

Example: Phonetic Alignment

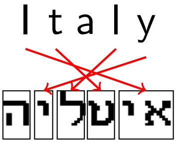
I t a l y

איטליה

Structured Output Learning

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

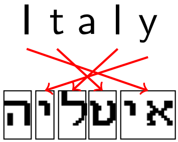
Example: Phonetic Alignment



Structured Output Learning

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

Example: Phonetic Alignment



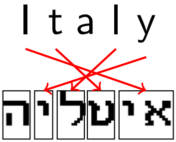
Israel

אילינוי

Structured Output Learning

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

Example: Phonetic Alignment



Israel

Yes/No

אילינוי

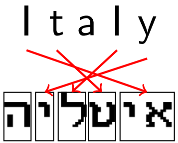
Structured Output Learning

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



Israel

Yes/No

אילינוי

Structured Output Learning

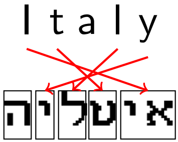
Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?

Companion Binary Output Problem

Are these two NEs a transliteration pair?

Is there any connection between these two problems?

Example: Phonetic Alignment



Israel

Yes/No

אילינוי

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

Key Intuition

Structured Output Task

Companion Binary Task

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.

Key Intuition

Structured Output Task

Companion Binary Task

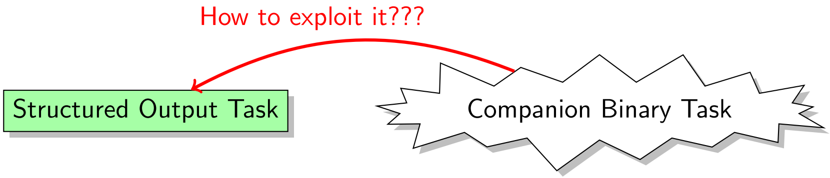
Observation

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



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

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

Training: Intuition

Given an example $(\mathbf{x}_i, \mathbf{h}_i)$, find a $\mathbf{w}$ such that the gold structure $\mathbf{h}_i$ has the highest score!

Geometric Interpretation for SSVM

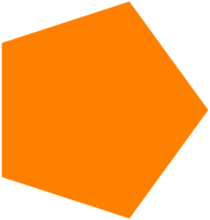
Decision Function

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

Training: Intuition

Given an example $(\mathbf{x}_i, \mathbf{h}_i)$, find a $\mathbf{w}$ such that the gold structure $\mathbf{h}_i$ has the highest score!

$\{\Phi(\mathbf{x}_1,\mathbf{h}) \mid \mathbf{h} \in \mathcal{H}(\mathbf{x}_1)\}$



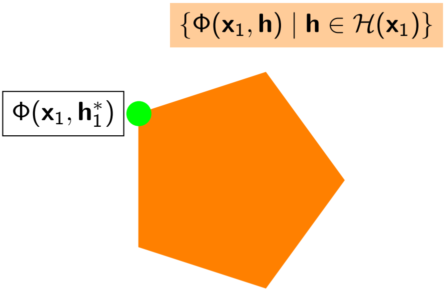
Geometric Interpretation for SSVM

Decision Function

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

Training: Intuition

Given an example $(\mathbf{x}_i, \mathbf{h}_i)$, find a $\mathbf{w}$ such that the gold structure $\mathbf{h}_i$ has the highest score!



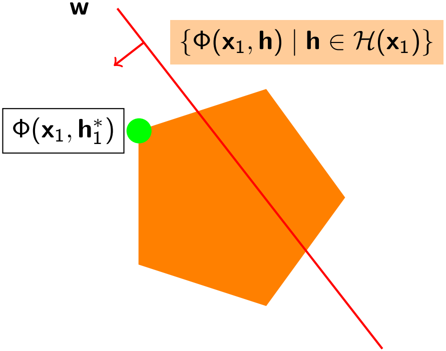
Geometric Interpretation for SSVM

Decision Function

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

Training: Intuition

Given an example $(\mathbf{x}_i, \mathbf{h}_i)$, find a $\mathbf{w}$ such that the gold structure $\mathbf{h}_i$ has the highest score!



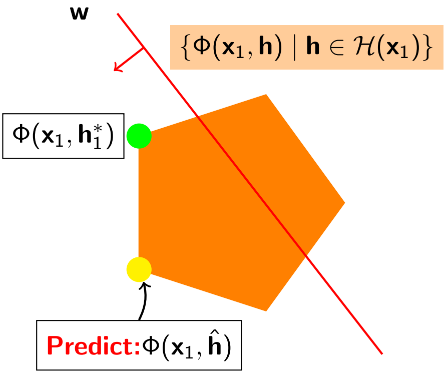
Geometric Interpretation for SSVM

Decision Function

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

Training: Intuition

Given an example $(\mathbf{x}_i,\mathbf{h}_i)$, find a $\mathbf{w}$ such that the gold structure $\mathbf{h}_i$ has the highest score!



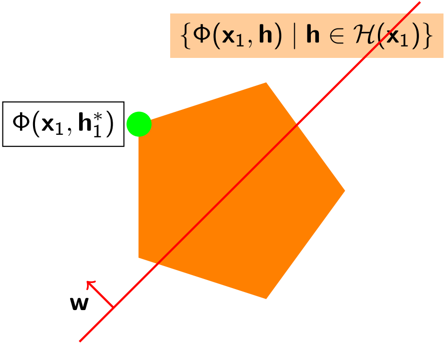
Geometric Interpretation for SSVM

Decision Function

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

Training: Intuition

Given an example $(\mathbf{x}_i, \mathbf{h}_i)$, find a $\mathbf{w}$ such that the gold structure $\mathbf{h}_i$ has the highest score!



Structural SVM

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})$$

- Regularization
  - Measures the model complexity
- Structural Loss :
  - S is the set of *structured* labeled examples:
  - $L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})$: Measures "the distance" between the current best prediction and the gold structure $\mathbf{h}_i$
- $L_{\mathcal{S}}$ can use hinge or square hinge functions or others
- A convex optimization problem

Structural SVM

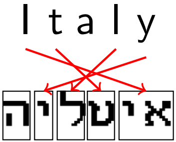
$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})$$

- Regularization
  - Measures the model complexity
- Structural Loss :
  - S is the set of *structured* labeled examples:
  - $L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})$: Measures "the distance" between the current best prediction and the gold structure $\mathbf{h}_i$
- $L_{\mathcal{S}}$ 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



Companion Binary Output Problem

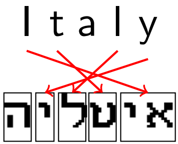
Israel

Yes/No

אילינוי

The role of binary labeled data

Structured Output Problem



Companion Binary Output Problem

Israel

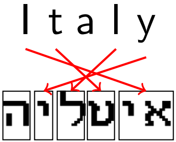
Yes/No

אילינוי

Companion Task: Does this example possess a good structure?

The role of binary labeled data

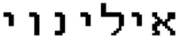
Structured Output Problem



Companion Binary Output Problem

Israel

Yes/No

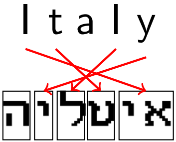


Companion Task: Does this example possess a good structure?

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

The role of binary labeled data

Structured Output Problem



Companion Binary Output Problem

Israel

Yes/No

אילינוי

Companion Task: Does this example possess a good structure?

- $\mathbf{x}_1$ is positive.
  - There must exist a good structure that justifies the positive label
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative.
  - No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Why is binary labeled data useful?

- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

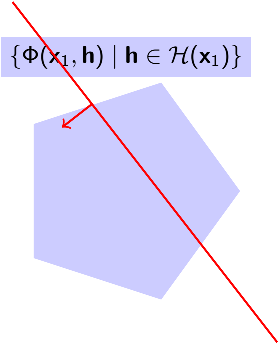
Why is binary labeled data useful?

$\{\Phi(\mathbf{x}_1, \mathbf{h}) \mid \mathbf{h} \in \mathcal{H}(\mathbf{x}_1)\}$

- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

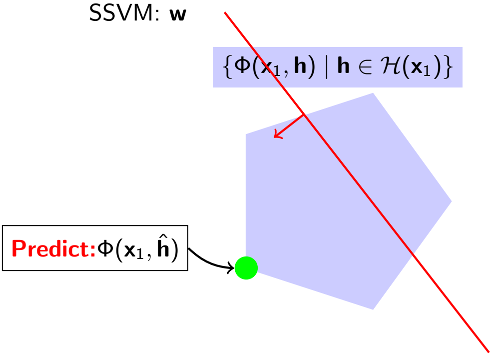
Why is binary labeled data useful?

SSVM: $\mathbf{w}$



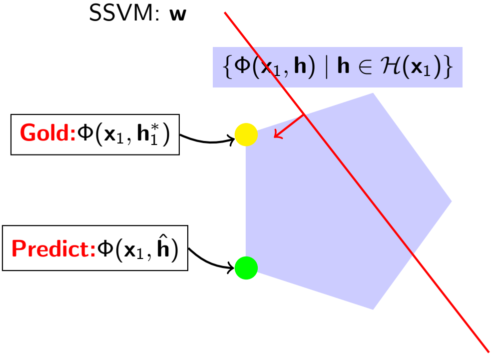
- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Why is binary labeled data useful?



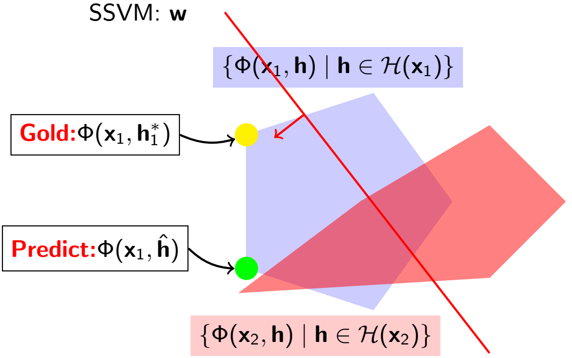
- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Why is binary labeled data useful?



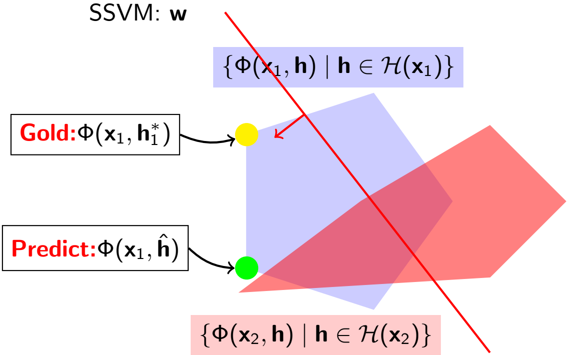
- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Why is binary labeled data useful?



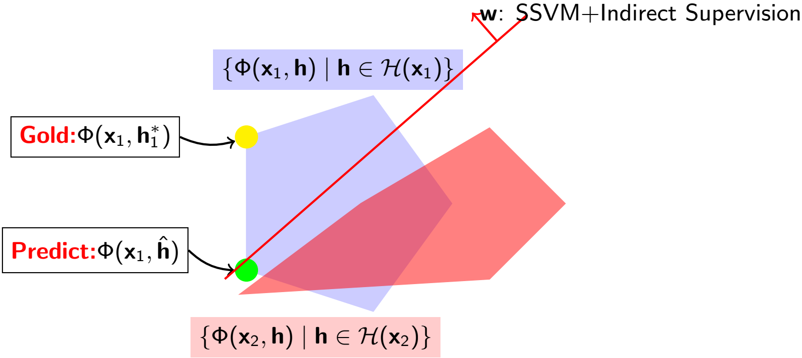
- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Why is binary labeled data useful?



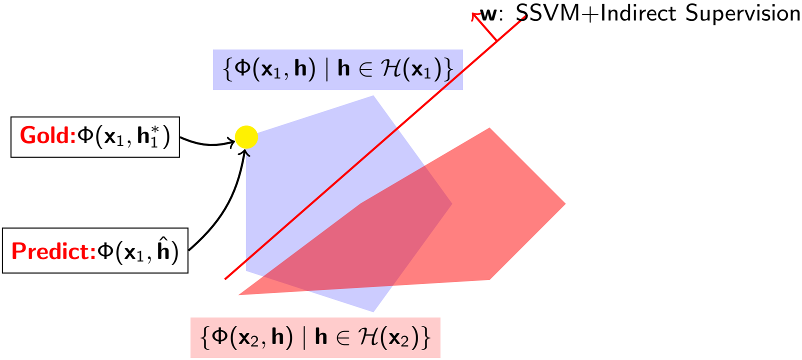
- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Why is binary labeled data useful?



- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Why is binary labeled data useful?



- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

Outline

Binary and structured labeled data

Direct Supervision: S

- **Target Task**

Indirect Supervision: B

- **Companion Task**

Binary and structured labeled data

Direct Supervision: S

- **Target Task**
- An example: $(\mathbf{x}_i, \mathbf{h}_i)$

Indirect Supervision: B

- **Companion Task**
- An example: $(\mathbf{x}_i, y_i)$

Binary and structured labeled data

Direct Supervision: S

- **Target Task**
- An example: $(\mathbf{x}_i, \mathbf{h}_i)$
- Goal:

$$\mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h}_i) \geq \max_{\mathbf{h} \in \mathcal{H}(\mathbf{x}_i)} \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h}).$$

Indirect Supervision: B

- **Companion Task**
- An example: $(\mathbf{x}_i, y_i)$
- Goal:

$$y_i \max_{\mathbf{h} \in \mathcal{H}(\mathbf{x}_i)} \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h}) \geq 0$$

Binary and structured labeled data

Direct Supervision: S

- **Target Task**
- An example: $(\mathbf{x}_i, \mathbf{h}_i)$
- Goal:

$$\mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h}_i) \geq \max_{\mathbf{h}\in\mathcal{H}(\mathbf{x}_i)} \mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h}).$$

- Structural Loss: $L_{\mathcal{S}}$

Indirect Supervision: B

- **Companion Task**
- An example: $(\mathbf{x}_i, y_i)$
- Goal:

$$y_i \max_{\mathbf{h}\in\mathcal{H}(\mathbf{x}_i)} \mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h}) \geq 0$$

- Binary Loss: $L_{\mathcal{B}}$

Binary and structured labeled data

Direct Supervision: S

- **Target Task**
- An example: $(\mathbf{x}_i, \mathbf{h}_i)$
- Goal:

$$\mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h}_i) \geq \max_{\mathbf{h} \in \mathcal{H}(\mathbf{x}_i)} \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h}).$$

- Structural Loss: $L_{\mathcal{S}}$

Indirect Supervision: B

- **Companion Task**
- An example: $(\mathbf{x}_i, y_i)$
- Goal:

$$y_i \max_{\mathbf{h} \in \mathcal{H}(\mathbf{x}_i)} \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h}) \geq 0$$

- Binary Loss: $L_{\mathcal{B}}$

Both $L_{\mathcal{S}}$ and $L_{\mathcal{B}}$ can use hinge, square-hinge, logistic, ...

Joint Learning with Indirect Supervision

$$\min_{\mathbf{w}} \boxed{\frac{\|\mathbf{w}\|^2}{2}} + \boxed{C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})} + \boxed{C_2 \sum_{i \in B} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})},$$

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

Joint Learning with Indirect Supervision

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in B} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}),$$

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

Joint Learning with Indirect Supervision

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in B} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}),$$

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

Share weight vector w

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

Outline

Convexity Properties

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in B} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}),$$

$$L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) = \ell\left(\max_{\mathbf{h}} \left(\Delta(\mathbf{h}, \mathbf{h}_i) - \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h}_i) + \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h})\right)\right) \quad (1)$$

$$L_B(\mathbf{x}_i, y_i, \mathbf{w}) = \ell\left(1 - y_i \max_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} (\mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}))\right) \quad (2)$$

Convexity Properties

Regularization, Direct Supervision, Negative Data B^-

Convex Parts

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in \mathbf{B}^-} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$$

$$+ C_2 \sum_{i \in \mathbf{B}^+} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$$

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 ℓ 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 ℓ 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

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in \mathbf{B}^-} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$$
$$+ C_2 \sum_{i \in \mathbf{B}^+} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$$

Solving the convex sub-problem

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in \mathbf{B}^-} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$$

$$+ C_2 \sum_{i \in \mathbf{B}^+} \boxed{L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}) \textbf{ with fixed structures}}$$

Solving the convex sub-problem

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in \mathbf{B}^-} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$$
$$+ C_2 \sum_{i \in \mathbf{B}^+} \boxed{L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}) \textbf{ with fixed structures}}$$

Cutting plane method

- Find the "best structure" for examples in S and B^- with the current $\mathbf{w}$
- Add chosen structure into the cache and solve it again!

Solving the convex sub-problem

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} + C_1 \sum_{i \in S} L_{\mathcal{S}}(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in \mathbf{B}^-} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$$

$$+ C_2 \sum_{i \in \mathbf{B}^+} \boxed{L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}) \textbf{ with fixed structures}}$$

Cutting plane method

- Find the "best structure" for examples in S and B^- with the current $\mathbf{w}$
- Add chosen structure into the cache and solve it again!

Dual coordinate descent method

- Simple implementation with square (L2) hinge loss

Outline

Experimental Setting

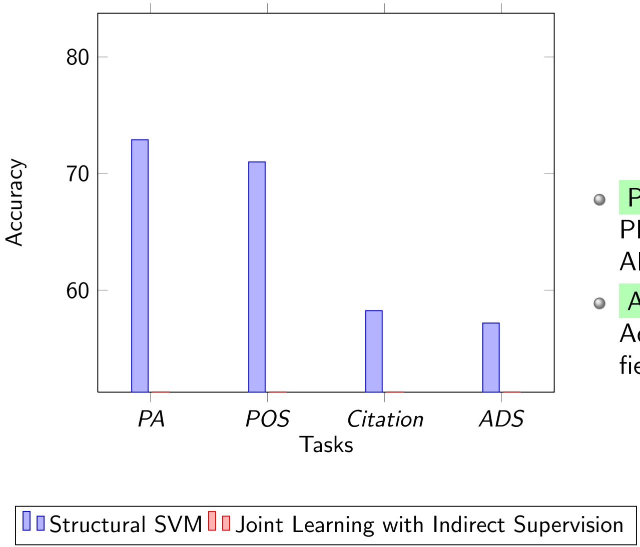
Tasks

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

Companion Tasks

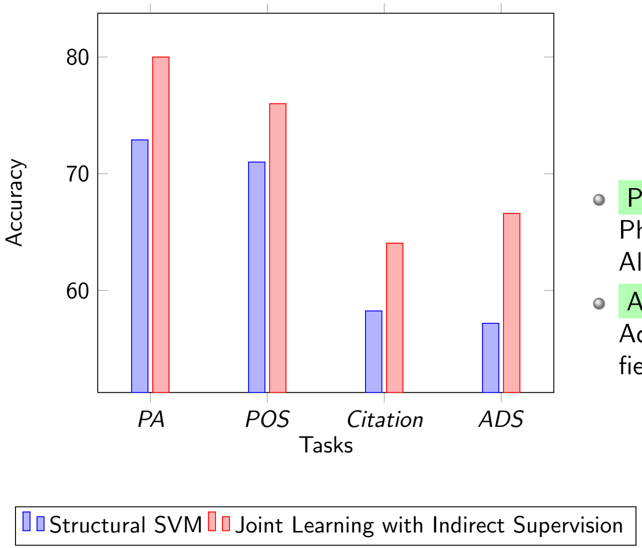
- **Phonetic alignment**: Transliteration pair or **not**
- **POS Tagging**: Has a legitimate POS tag sequence or **not**
- **IE**: Is a legitimate Citation/Advertisement or **not**

Experimental Results



- PA : Phonetic Alignment
- ADS : Advertisement field recognition

Experimental Results



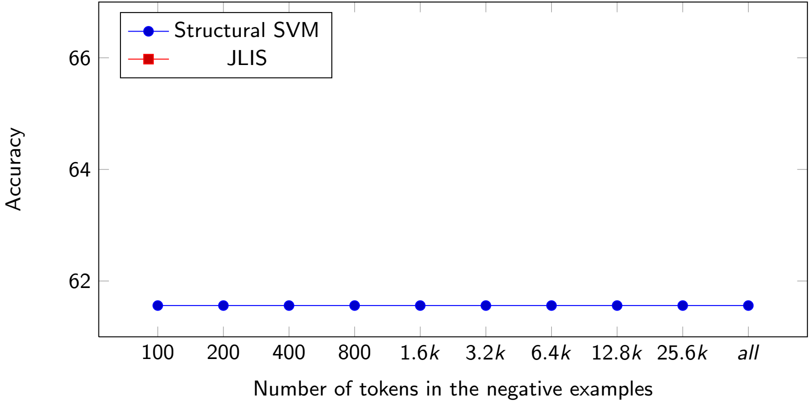
- PA : Phonetic Alignment
- ADS : Advertisement field recognition

Impact of negative examples

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

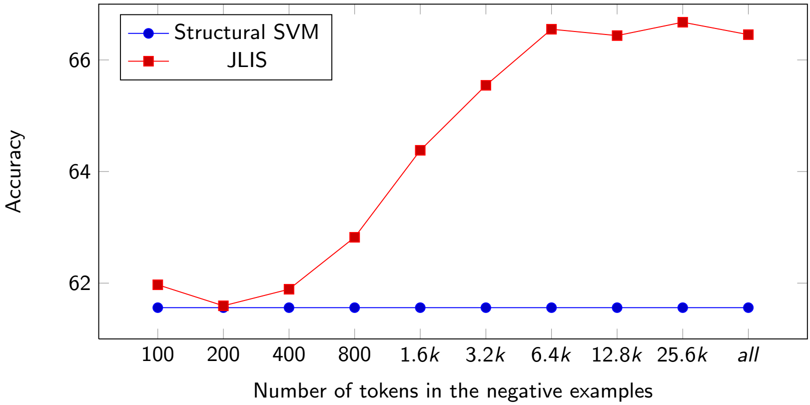
Impact of negative examples

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



Impact of negative examples

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



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)

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 Jump
- 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 Jump
- 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!

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 $\log P(\mathbf{x})$
- Model 1

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

- CE

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

Compared to Contrastive Estimation: II

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

	CE	J-LIS

Compared to Contrastive Estimation: II

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

	CE	J-LIS
Supervision type	"Neighbors"	Structured + Binary

Compared to Contrastive Estimation: II

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

	CE	J-LIS
Supervision type	"Neighbors"	Structured + Binary
Inference Problem	sum	max

Compared to Contrastive Estimation: II

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

	CE	J-LIS
Supervision type	"Neighbors"	Structured + Binary
Inference Problem	sum	max
Property		Can use existing data

- CE needs to know the relationship between "neighbors" of the input $\mathbf{x}$. J-LIS can use existing binary labeled data.

Compared to Contrastive Estimation: II

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

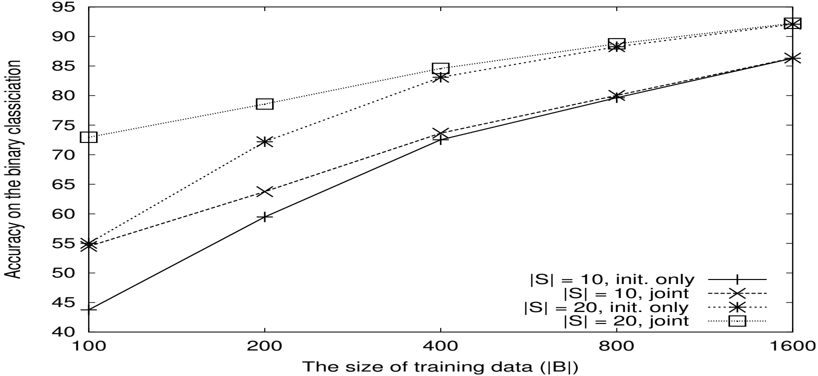
	CE	J-LIS
Supervision type	"Neighbors"	Structured + Binary
Inference Problem	sum	max
Property		Can use existing data

- CE needs to know the relationship between "neighbors" of the input $\mathbf{x}$. J-LIS can use existing binary labeled data.

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%

Joint learning: Results



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.

Brefeld, U. and T. Scheffer (2006).
Semi-supervised learning for structured output variables.
In *ICML*.

Tsochantaridis, I., T. Hofmann, T. Joachims, and Y. Altun (2004).
Support vector machine learning for interdependent and structured output spaces.
In *ICML*.

Zien, A., U. Brefeld, and T. Scheffer (2007).
Transductive support vector machines for structured variables.
In *ICML*.