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 `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 `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(·))`

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))))`

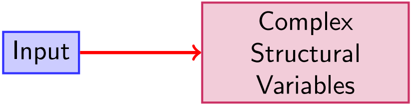
A structured task: multiple interdependent decisions

- `city(NY)` or `state(NY)`?
- `state(next_to(`$\cdot$`))` $\neq$ `next_to(state(`$\cdot$`))`

Supervision cost

- Labeling data is **very expensive!**
- The annotators need to know how to write meaning representation

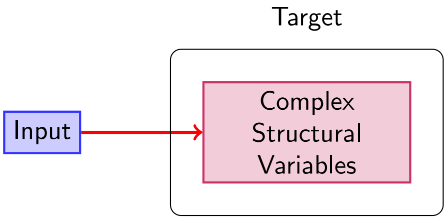
Main Idea: Indirect Supervision



Example

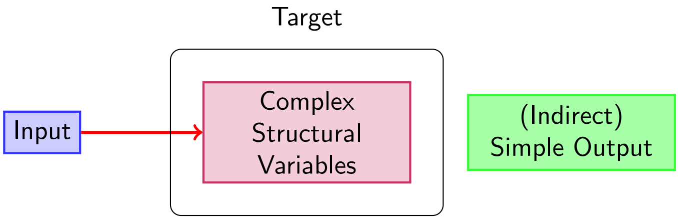
- Input Human Query, Output Meaning Representation

Main Idea: Indirect Supervision



Example

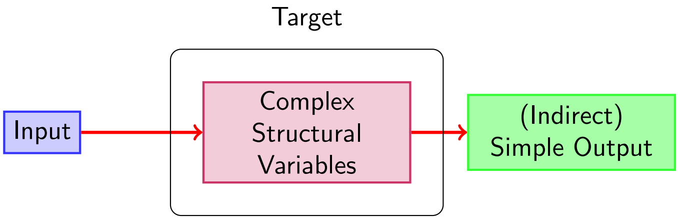
- Input Human Query, Output Meaning Representation



Example

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

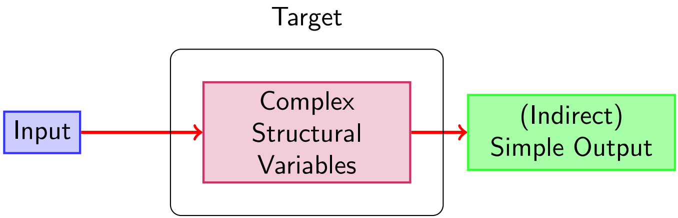
Main Idea: Indirect Supervision



Example

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

Main Idea: Indirect Supervision



Example

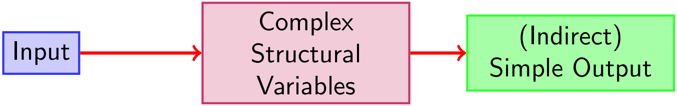
- Input Human Query, Output Meaning Representation
- (Indirect) Simple Output : Is the answer correct?

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

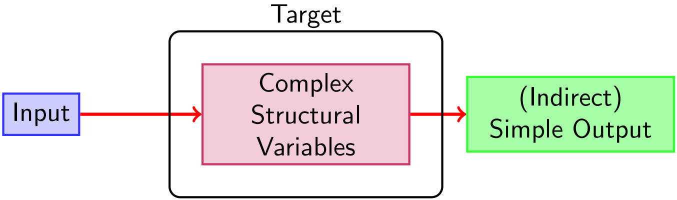
Part II: Learning with Indirect Supervision



Outline

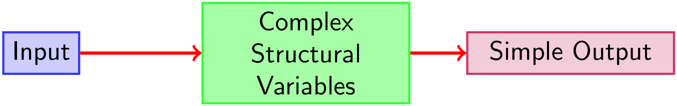
Part I: Learning with Latent Structure

Part II: Learning with Indirect Supervision

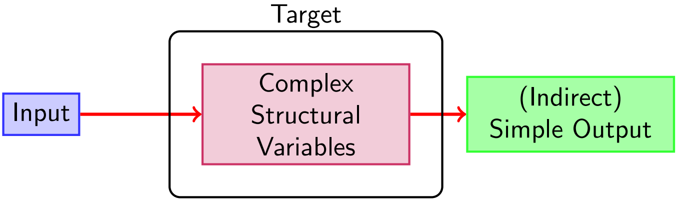


Outline

Part I: Learning with Latent Structure

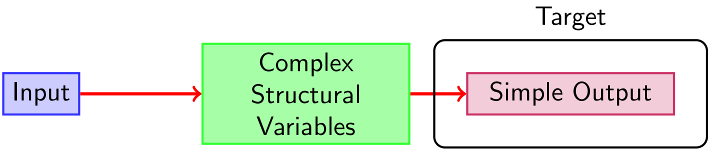


Part II: Learning with Indirect Supervision

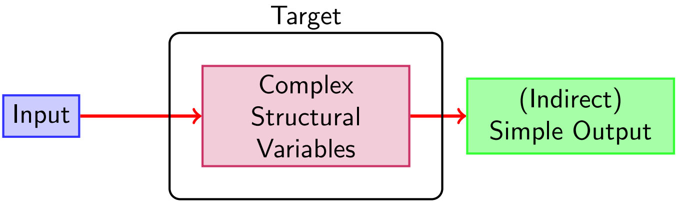


Outline

Part I: Learning with Latent Structure

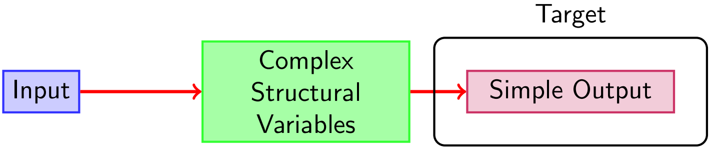


Part II: Learning with Indirect Supervision

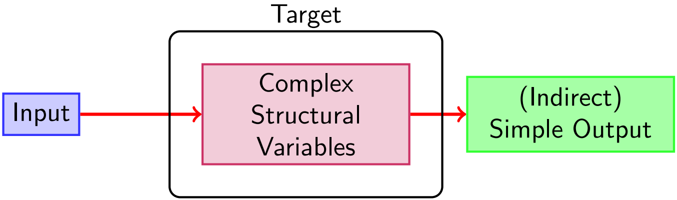


Outline

Part I: Learning with Latent Structure



Part II: Learning with Indirect Supervision



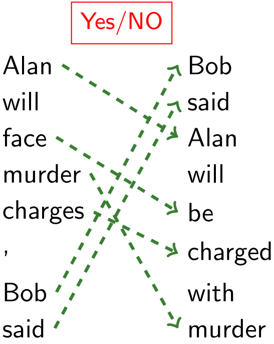
Example task: Paraphrase Identification

Yes/NO

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

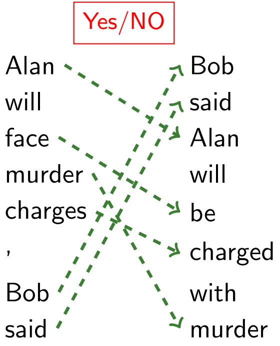
- Q: Are sentence 1 and sentence 2 paraphrases of each other?

Example task: Paraphrase Identification



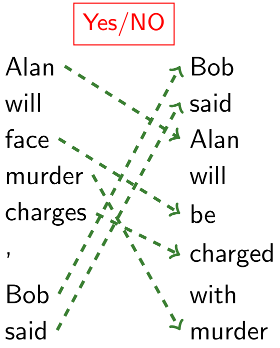
- Q: Are sentence 1 and sentence 2 paraphrases of each other?
  - Yes, but why?
  - They carry the same information!
- Justifying the decision requires **an intermediate representation**

Example task: Paraphrase Identification



- 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



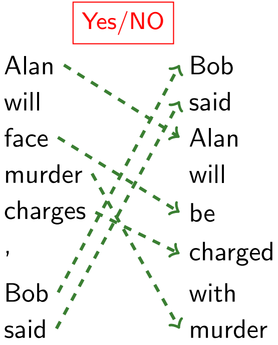
- Q: Are sentence 1 and sentence 2 paraphrases of each other?
  - Yes, but why?
  - They carry the same information!
- Justifying the decision requires **an intermediate representation**
- Just an example; the real intermediate representation is more complicated

Problem of interests

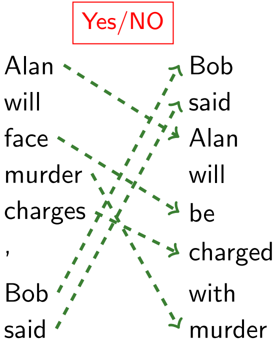
- Binary output problem: $z \in \{-1, 1\}$
- Intermediate representation: h
  - **Some structure that justifies the positive label**
  - The intermediate representation is **latent** (not present in the data)

Cognitive Computation Group
University of Illinois at Urbana-Champaign

The intuition behind the joint approach



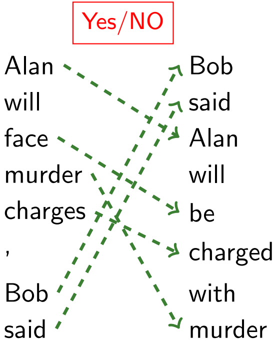
The intuition behind the joint approach



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

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

The intuition behind the joint approach

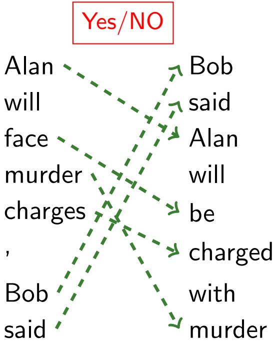


intermediate representation $\Leftrightarrow \{1, -1\}$

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

The intuition behind the joint approach

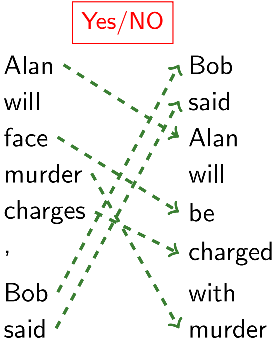


intermediate representation $\Leftrightarrow \{1, -1\}$

- 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
$\mathcal{H}(\mathbf{x})$: 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

$\mathbf{x}$: a sentence pair, **weight vector**: $\mathbf{w}$
$\mathbf{h}$: an alignment between two sentences
$\mathcal{H}(\mathbf{x})$: all possible alignments for $\mathbf{x}$

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

Geometric interpretation: **the case of two examples**

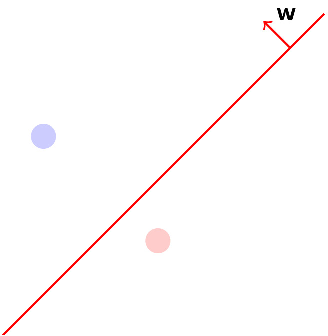
- Pair $\mathbf{x}_1$ is positive
  - There must exist a good explanation that justifies the positive label
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- Pair $\mathbf{x}_2$ is negative
  - No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



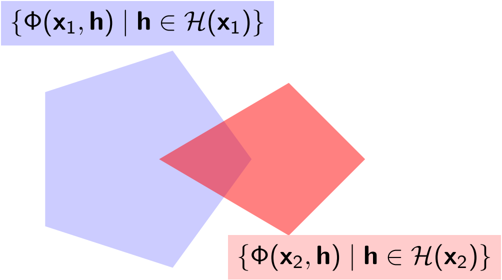
Geometric interpretation: **the case of two examples**

- Pair $\mathbf{x}_1$ 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 $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

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

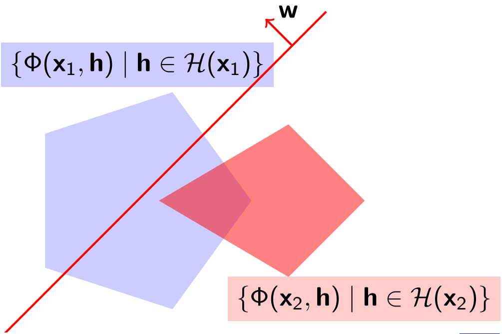
Geometric interpretation: **the case of two examples**

- 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 $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$



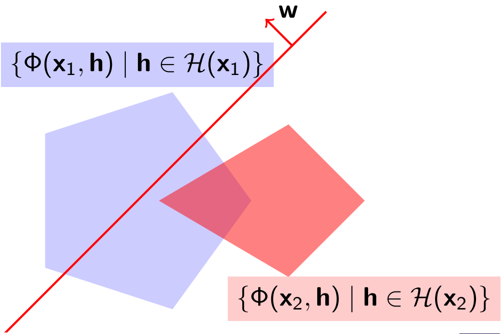
Geometric interpretation: **the case of two examples**

- Pair $\mathbf{x}_1$ 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 $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$



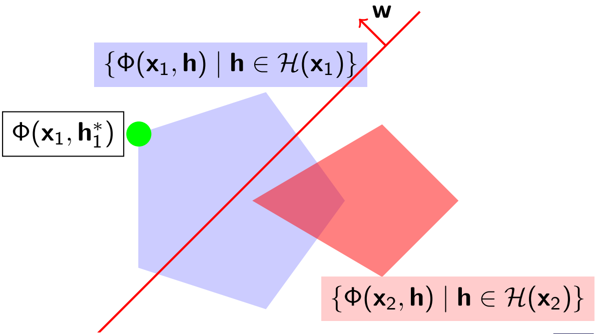
Geometric interpretation: **the case of two examples**

- Pair $\mathbf{x}_1$ 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 $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$



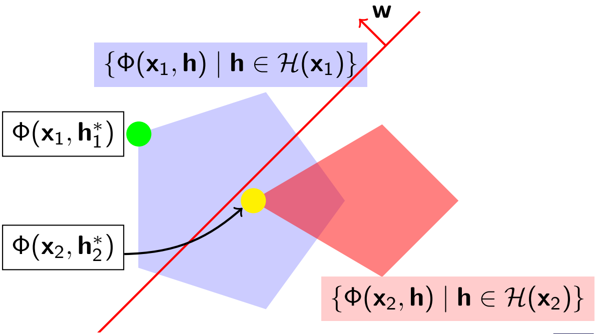
Geometric interpretation: **the case of two examples**

- Pair $\mathbf{x}_1$ 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 $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$



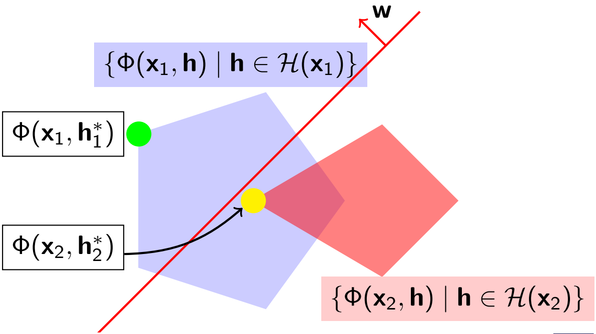
Geometric interpretation: **the case of two examples**

- Pair $\mathbf{x}_1$ 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 $\mathbf{x}_1$ 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 $\max_{\mathbf{h}} \mathbf{w}^T\Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$
- The prediction function: $\max_{\mathbf{h}} \mathbf{w}^T\Phi(\mathbf{x}, \mathbf{h})$



Finding Good Intermediate Representation

Find Structures

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

Constraint Based Declarative Framework

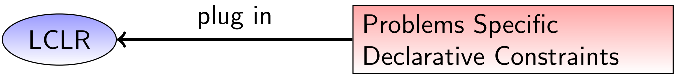
- We formulate this problem as an Integer Linear Programming problem (Roth and Yih 2004)
  1. Allow one to define the knowledge necessary for the problem declaratively
  2. Avoid designing a special purpose inference algorithm for each problem.
- Final System: Learning Constrained Latent Representation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

Find Structures

- In the learning algorithm, we need to solve $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}, \mathbf{h})$
- 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

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}) =$$
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \ell(-z_i \boxed{\max_{\mathbf{h}\in\mathcal{H}} \mathbf{w}^T \sum_{s\in\Gamma(\mathbf{x})} h_s \Phi_s(\mathbf{x})})$$

- **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

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w}) =$$
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \ell(-z_i \boxed{\max_{\mathbf{h}\in\mathcal{H}} \mathbf{w}^T \sum_{s\in\Gamma(\mathbf{x})} h_s \Phi_s(\mathbf{x})})$$

- **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

Our solution

- A new optimization algorithm: Focus on square-hinge loss
  - EM-like procedure + Cutting plane methods + Dual coordinate descent
  - $\min_{\mathbf{w}} \boxed{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{z_i=-1} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})} + C\sum_{z_i=+1} L_{\mathcal{B}}(\mathbf{x}_i, y_i, \mathbf{w})$
- Code available: `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

Experimental results

Transliteration System	Joint	ILP	Acc	MRR
(Goldwasser and Roth 2008)	$\star$		N/A	89.4
Our two-stage		$\star$	80.0	85.7
Our **LCLR**	$\star$	$\star$	**92.3**	**95.4**

Entailment System	Joint	ILP	Acc
Median of TAC 2009 systems			61.5
Our two-stage		$\star$	65.0
Our **LCLR**	$\star$	$\star$	**66.8**

Paraphrase Identification

Paraphrase System	Joint	ILP	**Acc**
Experiments using (Dolan, Quirk, and Brockett 2004)			
(Qiu, Kan, and Chua 2006)			72.00
(Das and Smith 2009)	$\star$		73.86
(Wan, Dras, Dale, and Paris 2006)			75.60
Our two-stage		$\star$	76.23
Our **LCLR**	$\star$	$\star$	**76.41**

Paraphrase Identification

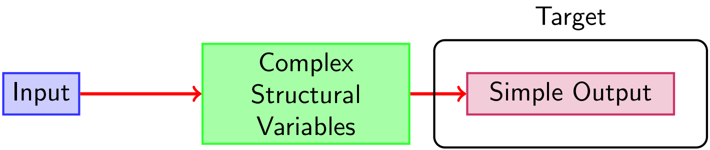
Paraphrase System	Joint	ILP	**Acc**
Experiments using (Dolan, Quirk, and Brockett 2004)			
(Qiu, Kan, and Chua 2006)			72.00
(Das and Smith 2009)	$\star$		73.86
(Wan, Dras, Dale, and Paris 2006)			75.60
Our two-stage		$\star$	76.23
Our **LCLR**	$\star$	$\star$	**76.41**

Paraphrase Identification

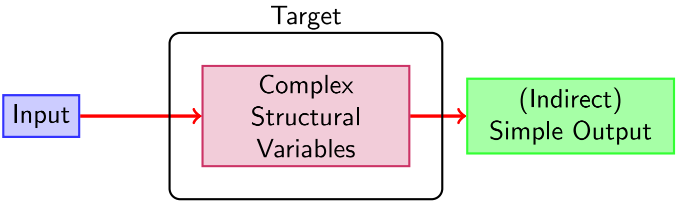
Paraphrase System	Joint	ILP	**Acc**
Experiments using (Dolan, Quirk, and Brockett 2004)			
(Qiu, Kan, and Chua 2006)			72.00
(Das and Smith 2009)	$\star$		73.86
(Wan, Dras, Dale, and Paris 2006)			75.60
Our two-stage		$\star$	76.23
Our **LCLR**	$\star$	$\star$	**76.41**
Experiments using **Noisy data set**			
Our two-stage		$\star$	72.00
Our **LCLR**	$\star$	$\star$	**72.75**

Outline

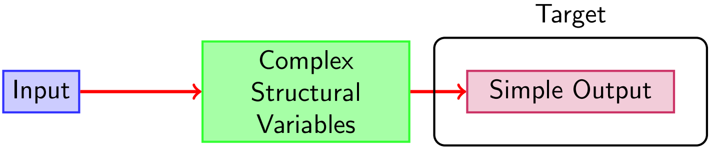
Part I: Learning with Latent Structure



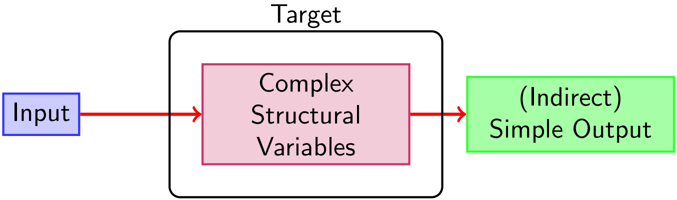
Part II: Learning with Indirect Supervision



Part I: Learning with Latent Structure



Part II: Learning with Indirect Supervision



Back to structured output learning

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?

Object Part Recognition

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

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?

Citation Recognition

Lars Ole Andersen . Program analysis and specialization for the C Programming language. PhD thesis , DIKU , University of Copenhagen , May 1994 .

Example structured output problems

Object Part Recognition

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

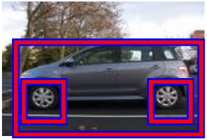
Citation Recognition

Lars Ole Andersen . Program analysis and specialization for the C Programming language. PhD thesis , DIKU , University of Copenhagen , May 1994 .

OUTPUT: **h**	Author	Author	Author	Author	Title	Title
INPUT: **x**	Lars	Ole	Andersen	.	Program	⋯

Task

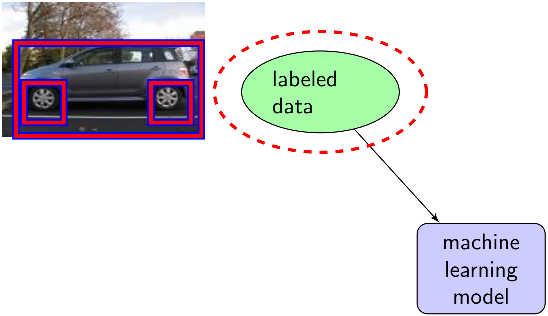
Given a car image, where are the body, windows and wheels?



Task

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

Supervising structured output problems

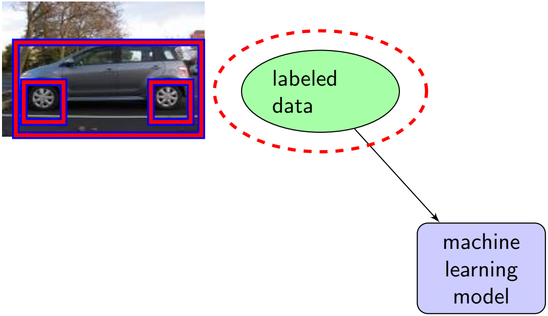


Task

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

- Supervised Approach

Supervising structured output problems

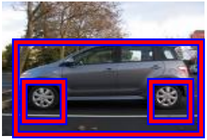


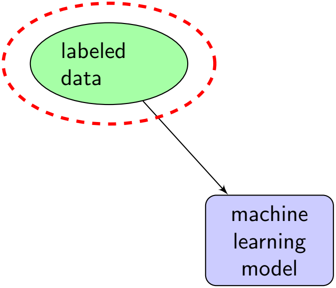
Task

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

- Supervised Approach is **Expensive!**

Supervising structured output problems

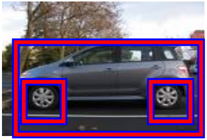


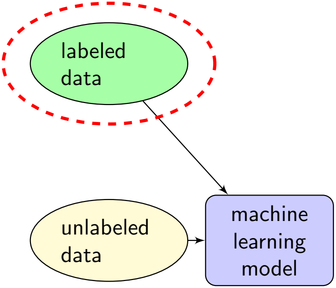


Task

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

- Supervised Approach is **Expensive!**



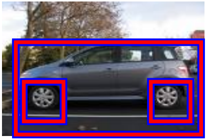


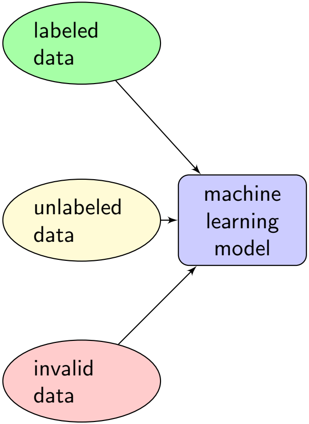
Task

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

- Supervised Approach is **Expensive!**

Supervising structured output problems



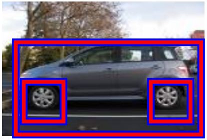


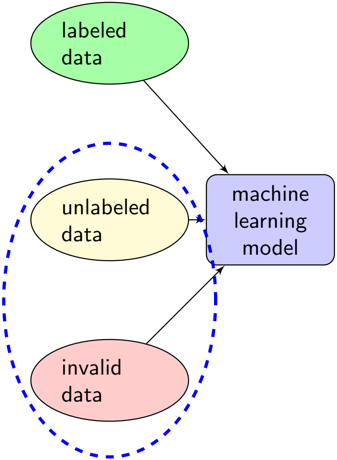
Task

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

- Supervised Approach is **Expensive!**

Supervising structured output problems



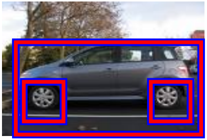


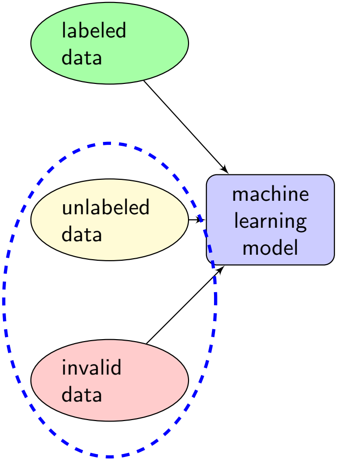
Task

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

- Supervised Approach is **Expensive!**

Supervising structured output problems





Task

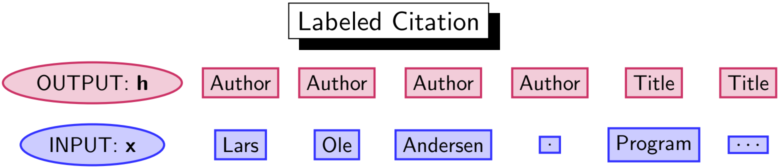
Given a car image, where are the body, windows and wheels?

- Supervised Approach is **Expensive!**

Indirect Supervision

Use binary labeled data as indirect supervisions

Supervised Learning algorithms



Semi-Supervised Learning algorithms

Labeled Citation

OUTPUT: **h**	Author	Author	Author	Author	Title	Title
INPUT: **x**	Lars	Ole	Andersen	.	Program	...

Unlabeled Citation: Positive Examples

INPUT: **x**	Ming	Wei	Chang	.	Structured	...

Supervising structured output problems: Citation

Indirect Supervision algorithm

Labeled Citation

OUTPUT: **h**	Author	Author	Author	Author	Title	Title
INPUT: **x**	Lars	Ole	Andersen	.	Program	...

Unlabeled Citation: Positive Examples

INPUT: **x**	Ming	Wei	Chang	.	Structured	...

Not a Citation: Negative Examples

- Shuffling tokens of a citation entry

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 possesses a good structure or not.

Key Intuition

Structured Output Task

Companion Binary Task

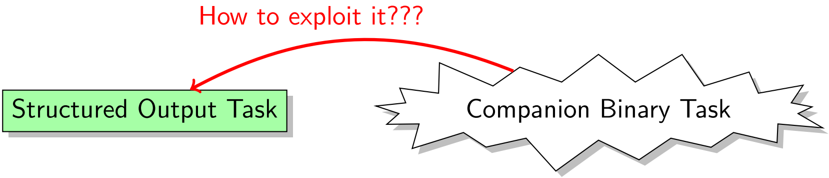
Observation

Many structured output prediction problems have a **companion** binary decision problem: predicting whether an input 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.

Why is this important

Binary labeled data is very easy to obtain

The role of binary labeled data

The role of binary labeled data

Structured Output Learning

- Recognize Car parts

Companion Binary Output Problem

- Is there a car in this image?

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?

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?

- $\mathbf{x}_1$ 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 Learning

- Recognize Car parts

Companion Binary Output Problem

- Is there a car in this image?

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?

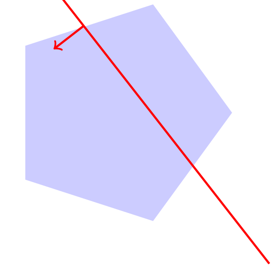
$\{\Phi(\mathbf{x}_1, \mathbf{h}) \mid \mathbf{h} \in \mathcal{H}(\mathbf{x}_1)\}$

- $\mathbf{x}_1$ 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?

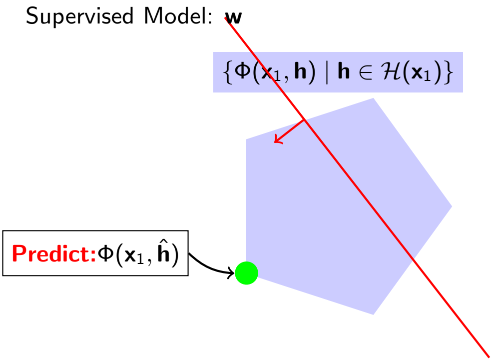
Supervised Model: $\mathbf{w}$

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



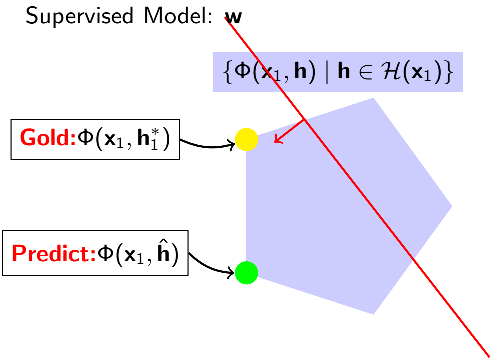
- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T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

Why is binary labeled data useful?



- $\mathbf{x}_1$ 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 $\max_{\mathbf{h}} \mathbf{w}^T\Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

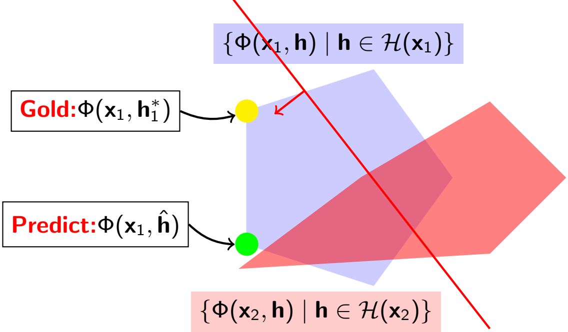
Why is binary labeled data useful?



- $\mathbf{x}_1$ 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?

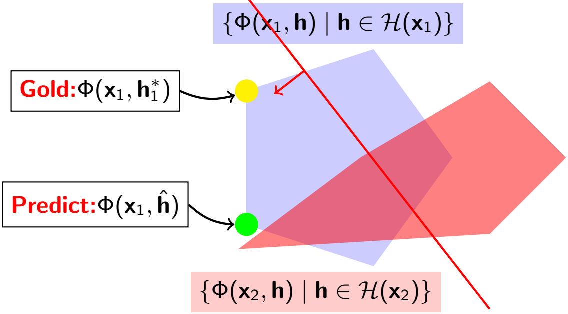
Supervised Model: $\mathbf{w}$



- $\mathbf{x}_1$ 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 $\max_{\mathbf{h}} \mathbf{w}^T \Phi(\mathbf{x}_2, \mathbf{h}) \leq 0$

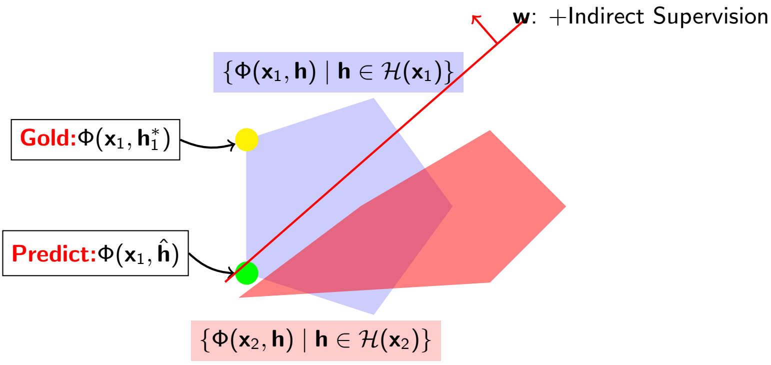
Why is binary labeled data useful?

Supervised Model: $\mathbf{w}$



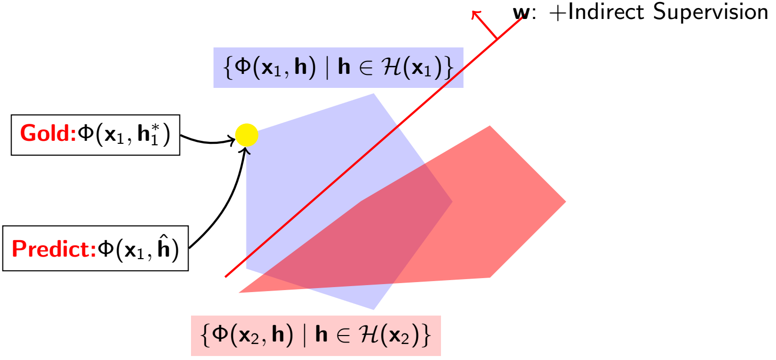
- $\mathbf{x}_1$ is positive : There exists a good structure
  - $\exists \mathbf{h}, \mathbf{w}^T\Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$, or $\max_{\mathbf{h}} \mathbf{w}^T\Phi(\mathbf{x}_1, \mathbf{h}) \geq 0$
- $\mathbf{x}_2$ is negative : No structure is good enough
  - $\forall \mathbf{h}, \mathbf{w}^T\Phi(\mathbf{x}_2,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

Joint Learning with Indirect Supervision [ICML'10]

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

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

Joint Learning with Indirect Supervision [ICML'10]

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

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

Joint Learning with Indirect Supervision [ICML'10]

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

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

Share weight vector w

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

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

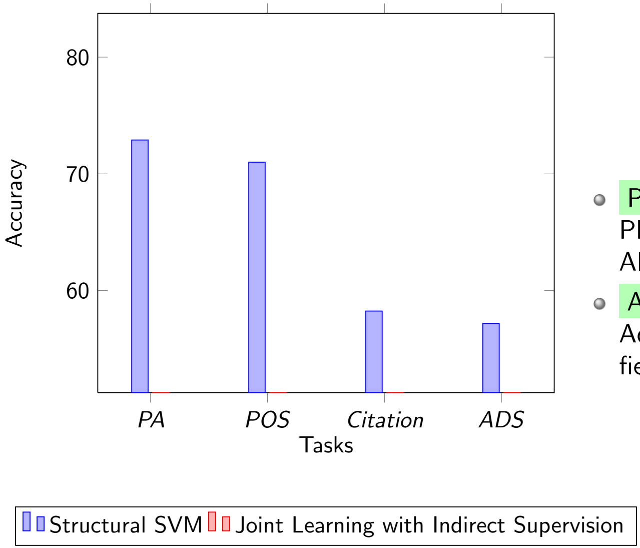
- Regularization : measures the model complexity
- Direct Supervision : structured labeled data $S = \{(\mathbf{x}, \mathbf{h})\}$
- Indirect Supervision : binary labeled data $B = \{(\mathbf{x}, z)\}$

Share weight vector w

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

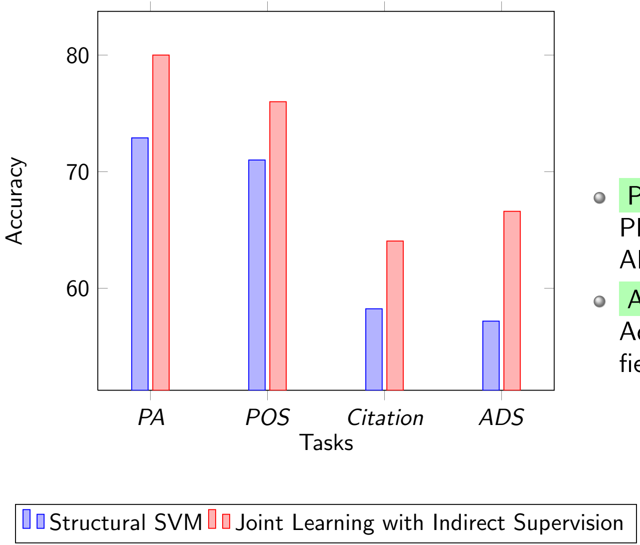
Support Structured SVM

Experimental Results



- PA : Phonetic Alignment
- ADS : Advertisement field recognition

Experimental Results



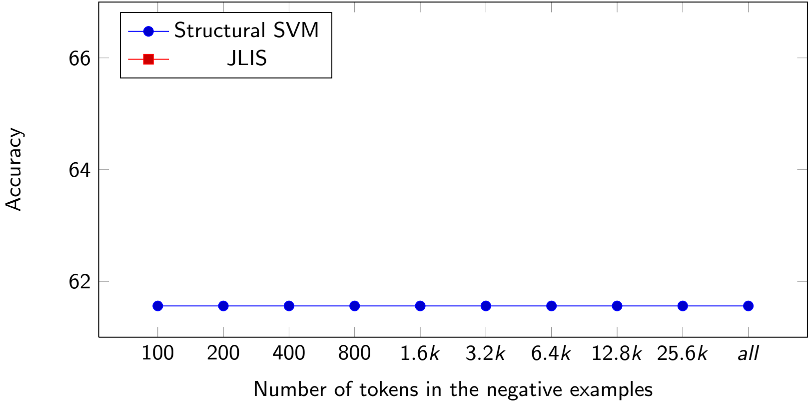
- PA : Phonetic Alignment
- ADS : Advertisement field recognition

Impact of negative examples

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

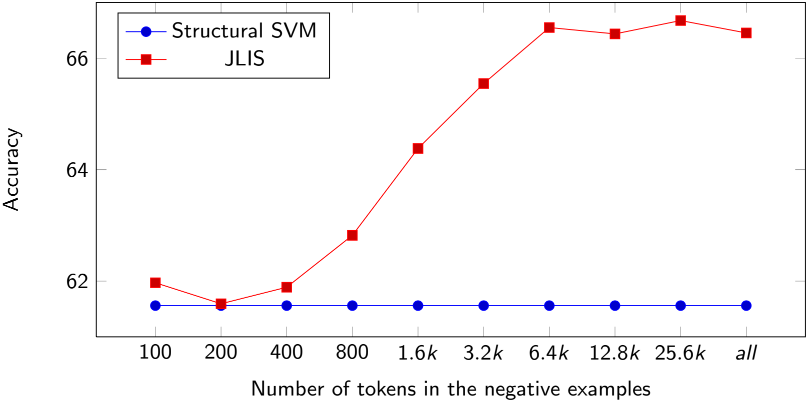
Impact of negative examples

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

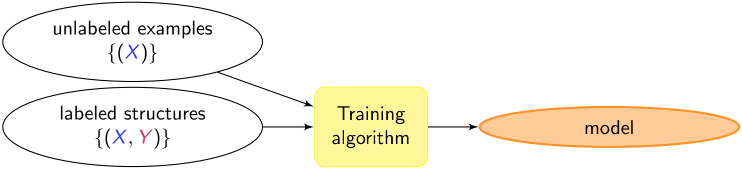


Impact of negative examples

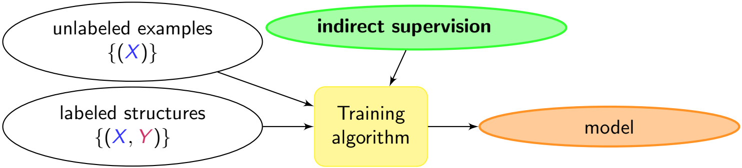
- J-LIS: takes advantage of *both* positively and negatively labeled data



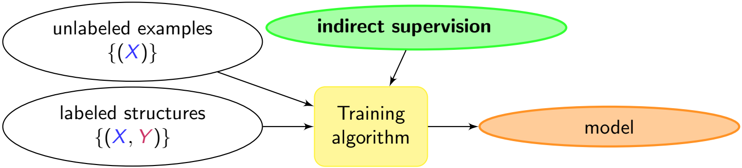
Recent publications about indirect supervisions



Recent publications about indirect supervisions



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

I t a l y

איטליה

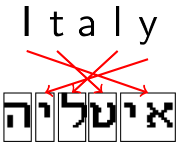
I t a l y

איטליה

Structured Output Learning

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

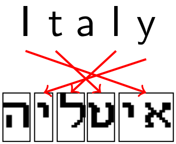
Example: Transliteration



Structured Output Learning

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

Example: Transliteration



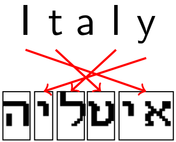
Israel

אילינוי

Structured Output Learning

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

Example: Transliteration



Israel

Yes/No

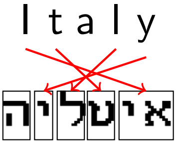
אילינוי

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?



Israel

Yes/No

אילינוי

Structured Output Learning

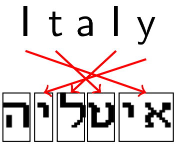
Given one English NE and its Hebrew transliteration, tell me what are the phonetic alignments?

Companion Binary Output Problem

Are these two NEs a transliteration pair?

Is there any connection between these two problems?

Example: Transliteration



Israel

Yes/No

אילינוי

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!

Carlson, A., J. Betteridge, R. C. Wang, E. R. H. Jr., and T. M. Mitchell (2010).
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*.

Clarke, J., D. Goldwasser, M. Chang, and D. Roth (2010).
Driving semantic parsing from the world's response.
In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010)*.

Das, D. and N. A. Smith (2009).
Paraphrase identification as probabilistic quasi-synchronous recognition.
In *ACL*.

Dolan, W., C. Quirk, and C. Brockett (2004).
Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.

In *COLING*.

Ganchev, K., J. Graça, J. Gillenwater, and B. Taskar (2010).
Posterior regularization for structured latent variable models.
Journal of Machine Learning Research.

Goldwasser, D. and D. Roth (2008).
Active sample selection for named entity transliteration.
In *ACL*.
Short Paper.

Liang, P., M. I. Jordan, and D. Klein (2011).
Learning dependency-based compositional semantics.
In *ACL*.

Mann, G. and A. McCallum (2008).
Generalized expectation criteria for semi-supervised learning of conditional random fields.
In *ACL*.

Qiu, L., M.-Y. Kan, and T.-S. Chua (2006).
Paraphrase recognition via dissimilarity significance classification.
In *EMNLP*.

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

Wan, S., M. Dras, R. Dale, and C. Paris (2006).
Using dependency-based features to take the p̈ara-farceöut of paraphrase.
In *Proc. of the Australasian Language Technology Workshop (ALTW)*.