Discriminative Learning over Constrained Latent Representations

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

Computer Science Department, University of Illinois at Urbana-Champaign
An one minute version of the talk

<table>
<thead>
<tr>
<th>What we did</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Provide a <em>general recipe</em> for many important NLP problems</td>
</tr>
<tr>
<td>- Our algorithm: <strong>Learning over Constrained Latent Representations</strong></td>
</tr>
</tbody>
</table>
An one minute version of the talk

**What we did**

- Provide a *general recipe* for many important NLP problems
- Our algorithm: Learning over Constrained Latent Representations

**Example NLP problems**

- Transliteration (Klementiev and Roth 2008),
- Textual entailment (RTE) (Dagan, Glickman, and Magnini 2006)
- Paraphrase identification (Dolan, Quirk, and Brockett 2004)
- Question Answering, and many more!
An one minute version of the talk

What we did
- Provide a general recipe for many important NLP problems
- Our algorithm: Learning over Constrained Latent Representations

Example NLP problems
- Transliteration (Klementiev and Roth 2008),
- Textual entailment (RTE) (Dagan, Glickman, and Magnini 2006)
- Paraphrase identification (Dolan, Quirk, and Brockett 2004)
- Question Answering, and many more!

Problems of Interests
Binary classification tasks that require an intermediate representation
Example task: Paraphrase Identification

Yes/NO

Alan will face murder charges, Bob said Alan will be charged with 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?
Example task: Paraphrase Identification

Yes/NO

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

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

Alan will face murder charges, Bob said Alan will be charged with murder.
Example task: Paraphrase Identification

Yes/NO

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

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

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

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

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

Yes/NO

Alan will face murder charges,

Bob said Alan will be charged with murder

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

Justifying the decision requires an intermediate representation

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

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

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

- Justifying the decision requires an intermediate representation
- Just an example; the real intermediate representation is more complicated

Problem of interests

- Binary output problem: $y \in \{-1, 1\}$
- Intermediate representation: $h$

- Some structure that justifies the positive label
- The intermediate representation is latent (not present in the data)
Limitations of existing approaches: two-stage approach

Most systems: a two-stage approach

Stage 1: Generate the intermediate representation

Obtain intermediate representation → Fix it (ignore the second stage)!

\[ X \rightarrow H \]
Limitations of existing approaches: two-stage approach

Most systems: a two-stage approach

Stage 1: Generate the intermediate representation
Obtain intermediate representation → Fix it (ignore the second stage)!
\( X \rightarrow H \)

Stage 2: Classification based on the intermediate representation
Extract features using the fixed representation and learn:
\( \Phi(X, H) \rightarrow Y \)
Limitations of existing approaches: two-stage approach

Most systems: a two-stage approach

Stage 1: Generate the intermediate representation
Obtain intermediate representation → Fix it (ignore the second stage)!

\[ X \rightarrow H \]

Stage 2: Classification based on the intermediate representation

Extract features using the fixed representation and learn:

\[ \Phi(X, H) \rightarrow Y \]

Problem: the intermediate representation ignores the binary task
Limitations of existing approaches: two-stage approach

Most systems: a two-stage approach

Stage 1: Generate the intermediate representation

Obtain intermediate representation → Fix it (ignore the second stage)!

\[ X \rightarrow H \]

Stage 2: Classification based on the intermediate representation

Extract features using the fixed representation and learn:

\[ \Phi(X, H) \rightarrow Y \]

Problem: the intermediate representation ignores the binary task
Limitations of existing approaches: inference

- Observation: decisions on intermediate representation are interdependent

Alan will face murder charges, said Bob.

Bob will be charged with murder.

Alan said.
Limitations of existing approaches: inference

- Observation: decisions on intermediate representation are interdependent

Alan will face murder charges, Bob said, Alan will be charged with murder.
Limitations of existing approaches: inference

- Observation: decisions on intermediate representation are interdependent

Alan will face murder charges, Bob said Alan will be charged with murder.
Limitations of existing approaches: inference

- Observation: decisions on intermediate representation are interdependent

- Many frameworks use custom designed inference procedures
- Difficult to add linguistic intuition/constraints on the intermediate representation
- Difficult to generalize to other tasks
• **Property 1:** Jointly learn intermediate representations and labels

\[ X \rightarrow H \rightarrow \Phi(X, H) \rightarrow Y \]
**Property 1:** Jointly learn intermediate representations and labels

\[ X \rightarrow H \quad \Phi(X, H) \rightarrow Y \]
Property 1: Jointly learn intermediate representations and labels.

\[ \Phi(X, H) \rightarrow Y \]

- **input**
- **intermediate representation**
Learning Constrained Latent Representation (LCLR)

- **Property 1:** Jointly learn intermediate representations and labels

![Diagram showing the process of learning constrained latent representation](Diagram)
Learning Constrained Latent Representation (LCLR)

- **Property 1:** Jointly learn intermediate representations and labels

\[ X \rightarrow H \rightarrow \Phi(X, H) \rightarrow Y \]

- input
- intermediate representation
- features
- binary label
**Property 1:** Jointly learn intermediate representations and labels

![Diagram](image_url)
Learning Constrained Latent Representation (LCLR)

- **Property 1:** Jointly learn intermediate representations and labels

  ![Diagram]

  - **input** → **intermediate representation** → **feedback** → **features** → **binary label**

  - Find an intermediate representation that helps the binary task
Property 1: Jointly learn intermediate representations and labels

Find an intermediate representation that helps the binary task

Property 2: Constraint-based inference for the intermediate representation
  - Uses integer linear programming on latent variables
  - Easy to inject constraints on latent variables
  - Easy to generalize to other tasks
Outline

1 Motivation and Contribution

2 Property 1: Jointly learn intermediate representations and labels

3 Property 2: Constraint-based inference for the intermediate representation

4 LCLR: Putting Everything Together

5 Experiments
Outline

1 Motivation and Contribution

2 Property 1: Jointly learn intermediate representations and labels

3 Property 2: Constraint-based inference for the intermediate representation

4 LCLR: Putting Everything Together

5 Experiments
The intuition behind the joint approach

Yes/NO

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

---

Alan will face murder charges, Bob said Alan will be charged with murder.
The intuition behind the joint approach

Yes/NO

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

intermediate representation ⇔ \{1, -1\}

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

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

Yes/NO

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

intermediate representation ⇔ \{1, -1\}

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

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

**Yes/NO**

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

**intermediate representation ⇔ \{1, -1\}**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The prediction function:

$$\max_h u^T \Phi(x, h) \begin{cases} \Phi(x_1, h) & | h \in \mathcal{H}(x_1) \\ \Phi(x_2, h) & | h \in \mathcal{H}(x_2) \end{cases}$$
Geometric interpretation: the case of two examples

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

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

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

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

**The prediction function:**
\[
\max_h u^T \Phi(x, h)
\]
Outline

1. Motivation and Contribution
2. Property 1: Jointly learn intermediate representations and labels
3. Property 2: Constraint-based inference for the intermediate representation
4. LCLR: Putting Everything Together
5. Experiments
Why is a declarative framework important?

- No more custom-designed inference procedures
- Easy to generalize to other tasks
- Easy to inject constraints and linguistic intuition

Declarative Framework
Why is a declarative framework important?
- No more custom-designed inference procedures
- Easy to generalize to other tasks
- Easy to inject constraints and linguistic intuition
Why is a declarative framework important?
- No more custom-designed inference procedures
- Easy to generalize to other tasks
- Easy to inject constraints and linguistic intuition

Paraphrasing
Model input as graphs. \( G_a \): the first sentence. \( G_b \): the second sentence.
- Each vertex in \( G_a \) can be mapped to at most one vertex in \( G_b \) (vice versa)
- Each edge in \( G_a \) can be mapped to at most one edge in \( G_b \) (vice versa)
- Edge mapping is active iff the corresponding node mappings are active
Why is a declarative framework important?
- No more custom-designed inference procedures
- Easy to generalize to other tasks
- Easy to inject constraints and linguistic intuition
- Check out the CCM tutorial!

Paraphrasing
Model input as graphs. \( G_a \): the first sentence. \( G_b \): the second sentence.
- Each vertex in \( G_a \) can be mapped to at most one vertex in \( G_b \) (vice versa)
- Each edge in \( G_a \) can be mapped to at most one edge in \( G_b \) (vice versa)
- Edge mapping is active iff the corresponding node mappings are active
<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alan will face murder charges</td>
<td>Bob said Alan will be charged with murder</td>
</tr>
</tbody>
</table>

We need this because of the formulation. You do not need to parse the symbols in this page.
We need this because of the formulation. You do not need to parse the symbols in this page.

\[ \Gamma(x), \text{ the set of all "parts" that } x \text{ can generate } \]

\[ |\Gamma(x)| = 8 \times 8 = 64 \]
Finding intermediate representation using ILP

- We need this because of the formulation. You do not need to parse the symbols in this page.
- \( \Gamma(x) \), the set of all “parts” that \( x \) can generate \( |\Gamma(x)| = 8 \times 8 = 64 \)
- Rewrite \( h \in \{0, 1\}^{64} \) as a binary vector \( h = \{0, 0, 0, \ldots, 1, 0, 0, 1, 1\} \)
We need this because of the formulation. You do not need to parse the symbols in this page.

- $\Gamma(x)$, the set of all “parts” that $x$ can generate $|\Gamma(x)| = 8 \times 8 = 64$
- Rewrite $h \in \{0, 1\}^{64}$ as a binary vector $h = \{0, 0, 0, \ldots, 1, 0, 0, 1, 1\}$
- A feature vector $\Phi_s(x)$ for every part $h_s$
Finding intermediate representation using ILP

Sentence 1
Alan will face murder charges.

Sentence 2
Bob said, "Alan will be charged with murder."

- We need this because of the formulation. You do not need to parse the symbols in this page.
- \( \Gamma(x) \), the set of all “parts” that \( x \) can generate \( |\Gamma(x)| = 8 \times 8 = 64 \)
- Rewrite \( h \in \{0, 1\}^{64} \) as a binary vector \( h = \{0, 0, 0, \ldots, 1, 0, 0, 1, 1\} \)
- A feature vector \( \Phi_s(x) \) for every part \( h_s \)

"Inference Problem = ILP formulation (pink box)"

\[
\max_{h \in \mathcal{H}} \mathbf{u}^T \Phi(x, h) = \max_{h \in \mathcal{H}} \mathbf{u}^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x)
\]
Outline

1 Motivation and Contribution

2 Property 1: Jointly learn intermediate representations and labels

3 Property 2: Constraint-based inference for the intermediate representation

4 LCLR: Putting Everything Together

5 Experiments
Review: Logistic Regression and Support Vector Machine
Decision Function: $f(x, u) \geq 0$
Review: Logistic Regression and Support Vector Machine

- Decision Function: \( f(x, u) \geq 0 \)
- Objective Function:

\[
\min_u \frac{1}{2} \|u\|^2 + C \sum_{i=1}^{l} \ell(-y_i f(x, u))
\]
Review: Logistic Regression and Support Vector Machine

- Decision Function: \( u^T \Phi(x) \geq 0 \)
- Objective Function:

\[
\min_u \frac{1}{2} \|u\|^2 + C \sum_{i=1}^{l} \ell(-y_i; u^T \Phi(x_i))
\]
Review: Logistic Regression and Support Vector Machine

- Decision Function: $\mathbf{u}^T \Phi(\mathbf{x}) \geq 0$
- Objective Function:

$$\min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u}\|^2 + C \sum_{i=1}^{l} \ell(-y_i; \mathbf{u}^T \Phi(\mathbf{x}_i))$$
LCLR: The objective function

- **Learning over Constrained Latent Representations**
- Decision Function (ILP): $f(x, u) \geq 0$
LCLR: The objective function

- **Learning over Constrained Latent Representations**
  - Decision Function (ILP): \( f(x, u) \geq 0 \)
  - Objective Function

\[
\min_u \frac{1}{2} \|u\|^2 + C \sum_{i=1}^{l} \ell(-y_i; f(x, u))
\]
LCLR: The objective function

- **Learning over Constrained Latent Representations**
  - Decision Function (ILP):
    \[
    \max_{h \in \mathcal{H}} \ u^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x) \geq 0
    \]
  - Objective Function
    \[
    \min_u \ \frac{1}{2} \|u\|^2 + C \sum_{i=1}^{l} \ell(-y_i) \max_{h \in \mathcal{H}} \ u^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x)
    \]
LCLR: The objective function

Learning over Constrained Latent Representations

- Decision Function (ILP):
  \[ \max_{h \in \mathcal{H}} \ u^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x) \geq 0 \]

- Objective Function
  \[ \min_{u} \frac{1}{2} \|u\|^2 + C \sum_{i=1}^{I} \ell(-y_i) \max_{h \in \mathcal{H}} u^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x) \]

Beyond standard LR/SVM
Solves an inference problem (max) to select \( h \) (also affect features)
Challenges in optimizing the objective function

\[
\min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u}\|^2 + C \sum_{i=1}^{l} \ell(-y_i \max_{\mathbf{h} \in \mathcal{H}} \mathbf{u}^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x))
\]

- Not a regular LR/SVM
- LCLR has an inference procedure inside the minimization problem
Challenges in optimizing the objective function

\[
\min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u}\|^2 + C \sum_{i=1}^{I} \ell(-y_i \max_{h \in \mathcal{H}} \mathbf{u}^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x))
\]

Not a regular LR/SVM

- LCLR has an inference procedure inside the minimization problem
  - No shortcut

- Find the best representation for all examples
- Obtain a new weight vector using a LR/SVM package with the updated representations.
- Repeat.
Challenges in optimizing the objective function

\[
\min_u \frac{1}{2} \|u\|^2 + C \sum_{i=1}^{\ell} \ell(-y_i \max_{h \in H} u^T \sum_{s \in \Gamma(x)} h_s \Phi_s(x))
\]

- Not a regular LR/SVM

- LCLLR has an inference procedure inside the minimization problem

- Find the best representation for all examples
- Obtain a new weight vector using a LR/SVM package with the updated representations.
- Repeat.

Does not minimize the objective function
LCLR: optimization procedure

Algorithm

1: Find the best intermediate representations for **positive examples**
2: Find the weight vector with this intermediate representation
   - Still need to do inference for negative examples
   - **Not a regular SVM problem even in this step!**
3: Repeat!
LCLR: optimization procedure

Algorithm

1: Find the best intermediate representations for positive examples
2: Find the weight vector with this intermediate representation
   - Still need to do inference for negative examples
   - Not a regular SVM problem even in this step!
3: Repeat!

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

1: Find the best intermediate representations for positive examples
2: Find the weight vector with this intermediate representation
   - Still need to do inference for negative examples
   - Not a regular SVM problem even in this step!
3: Repeat!

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

Properties of the algorithm: Asymmetric nature

- Asymmetry between positive and negative examples
- Converting a non-convex problem into a series of smaller convex problems
Comparison to other latent variable frameworks

Inference procedure

- Other frameworks often use application-specific inference.
- LCLR allows you to add constraints and generalize to other tasks.
## Comparison to other latent variable frameworks

<table>
<thead>
<tr>
<th>Inference procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other frameworks often use application-specific inference.</td>
</tr>
<tr>
<td>LCLR allows you to add constraints and generalize to other tasks.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not only for SVM. Many different loss functions can be used.</td>
</tr>
<tr>
<td>Dual coordinate descent methods and cutting plane method</td>
</tr>
<tr>
<td>Fewer parameters to tune. Allows parallel inference procedure.</td>
</tr>
</tbody>
</table>
Comparison to other latent variable frameworks

Inference procedure
- Other frameworks often use application-specific inference.
- LCLR allows you to add constraints and generalize to other tasks.

Learning
- Not only for SVM. Many different loss functions can be used.
- Dual coordinate descent methods and cutting plane method
  - Fewer parameters to tune. Allows parallel inference procedure.

CRF-like latent variable framework
- LCLR can use logistic regression and have a probabilistic interpretation
- LCLR solves the “max” problem. CRF-like models solves the “sum” problem. "Max" enables adding constraints.
Outline

1 Motivation and Contribution

2 Property 1: Jointly learn intermediate representations and labels

3 Property 2: Constraint-based inference for the intermediate representation

4 LCLR: Putting Everything Together

5 Experiments
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?

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

<table>
<thead>
<tr>
<th>Transliteration System</th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Goldwasser and Roth 2008)</td>
<td>⋆</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our two-stage</td>
<td>⋆</td>
<td>⋆</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Experimental results

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

Median of TAC 2009 systems

61.5

65.0

66.8
## Experimental results

<table>
<thead>
<tr>
<th>Transliteration System</th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Goldwasser and Roth 2008)</td>
<td>✯</td>
<td></td>
<td>N/A</td>
<td>89.4</td>
</tr>
<tr>
<td>Our two-stage</td>
<td></td>
<td>✯</td>
<td>80.0</td>
<td>85.7</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>✯</td>
<td>✯</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Experimental results

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

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

<table>
<thead>
<tr>
<th>Entailment System</th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median of TAC 2009 systems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our two-stage</td>
<td></td>
<td>⋆</td>
<td></td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td></td>
</tr>
</tbody>
</table>
## Experimental results

### Transliteration System

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

### Entailment System

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

### Transliteration System

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

### Entailment System

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

### Transliteration System

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

### Entailment System

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

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

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

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

<table>
<thead>
<tr>
<th>Paraphrase System</th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Experiments using Noisy data set</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our two-stage</td>
<td>⋆</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td></td>
</tr>
</tbody>
</table>
## Paraphrase Identification

### Experiments using *(Dolan, Quirk, and Brockett 2004)*

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

### Experiments using **Noisy data set**

<table>
<thead>
<tr>
<th>Paraphrase System</th>
<th>Joint</th>
<th>ILP</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our two-stage</td>
<td>⋆</td>
<td></td>
<td>72.00</td>
</tr>
<tr>
<td>Our <strong>LCLR</strong></td>
<td>⋆</td>
<td>⋆</td>
<td><strong>72.75</strong></td>
</tr>
</tbody>
</table>
Conclusions

LCLR = Constraint-based Inference + Large Margin Learning

Contributions
- LCLR joint approach is better than two-stage approaches
- LCLR allows the use of constraints on latent variables
- A novel learning framework
Conclusions

\[ \text{LCLR} = \text{Constraint-based Inference} + \text{Large Margin Learning} \]

**Contributions**

- LCLR joint approach is better than two-stage approaches
- LCLR allows the use of constraints on latent variables
- A novel learning framework

**Bonus: Learning Structures with Indirect Supervision**

- Easy to get *binary* labeled data can be used to improve learning structures!
- Check out our ICML paper this year!
Thank you!

- Our learning code is available: the JLIS package
- http://l2r.cs.uiuc.edu/~cogcomp/software.php
Main Idea: Learning with indirect supervision

Indirect supervision: the supervision form that does not tell you the target output directly

Advantages of using indirect supervision:
- Can directly use human/domain knowledge to improve the model
- Allow us to use supervision signals that are a lot easier to obtain than labeling structures
- Use existing labeled data for the related tasks

Indirect supervision greatly reduces the supervision effort!
Indirect supervision: the supervision form that does not tell you the target output directly

Main Idea: Learning with indirect supervision

Indirect supervision greatly reduces supervision effort!
Main Idea: Learning with indirect supervision

Indirect supervision: the supervision form that does not tell you the target output directly

Advantage of using indirect supervision

- Can directly use human/domain knowledge to improve the model
- Allow us to use supervision signals that are a lot easier to obtain than labeling structures
- Use existing labeled data for the related tasks
Main Idea: Learning with indirect supervision

Indirect supervision: the supervision form that does not tell you the target output directly

Advantage of using indirect supervision
- Can directly use human/domain knowledge to improve the model
- Allow us to use supervision signals that are a lot easier to obtain than labeling structures
- Use *existing* labeled data for the related tasks

Indirect supervision greatly reduce the supervision effort!
Compared to CRF-like latent variable framework

### CRF-like latent variable framework

\[
P(y = 1|x) = \sum_h P(y = 1, h|x) = \frac{\sum_h \exp(u^T \phi(x, h, y = 1))}{\sum_h, y \exp(u^T \phi(x, h, y))}
\]

### LCLR with logistic loss

\[
P(y = 1|x) = \frac{\max_h \exp(u^T \phi(x, h))}{1 + \max_h \exp(u^T \phi(x, h))}
\]

- **Difference 1**: LCLR only models the “goodness”
  - This is important for many NLP problems, where only positive examples have good representations.
- **Difference 2**: LCLR only need to solve the max inference
  - Sometimes calculating sum is a lot harder!!
Paraphrase Identification: Revisited

Sentence 1
- Alan
- will
- face
- murder
- charges
- Bob
- said

Sentence 2
- Bob
- said
- Alan
- will
- be
- charged
- with
- murder

- **Left**: The intermediate representation is not expressive enough
  - For example, “word ordering” is a problem
- **The real setting**
  - Input: two word sequence → two graphs.
  - We used Stanford Parser to construct dependency parse trees for each sentence

*Integer Linear Programming to solve the graph matching problem*

- Four types of sub-structure: node matching, node-deletion, edge matching, edge-deletion
- Add constraints to enforce consistency
  - edge matching if and only if the corresponding nodes are matched

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

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

Active sample selection for named entity transliteration.
In *ACL.*
Short Paper.

Named entity transliteration and discovery in multilingual corpora.
In C. Goutte, N. Cancedda, M. Dymetman, and G. Foster (Eds.), *Learning Machine Translation.*