Practical Learning Algorithms for Structured Prediction Models

Kai-Wei Chang

University of Illinois at Urbana-Champaign
Dream:
Intelligent systems that are able to read, to see, to talk, and to answer questions.
“What’s my schedule for today”

OK, Kevin, I found 4 appointments for today:

- Yoga 3:30 PM
- Dance 6 PM

Translation system

United Arab Emirates (UAE) today ordered an F-16 fighter squadron stationed in Jordan, to support Jordan raid militant group "Islamic State" (IS).

Alābō lănghé dàgōngguó (UAE) jīnlán xiàolíng yīgè F-16 zhǎnjiē zhòngduì jìnzhòu yuèdàn, zhiyuán yuèdàn kōngxì jījīn zǔzhì “yīsīlán guó” (IS).
Carefully Slide
Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book.
Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book.
Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book
Scalability Issues

- **Large** amount of data

- Complex decision structure
Goal: Practical Machine Learning

- [Modeling] Expressive and general formulations
- [Algorithms] Principled and efficient
- [Applications] Support many applications
My Research Contributions

Limited memory linear classifier [KDD 10, 11, TKDD 12]

Latent representation for knowledge bases [EMNLP 13, 14]

Linear classification [ICML08, KDD 08, JMLR 08a, 10a, 10b, 10c]

Structured prediction models [ICML 14, ECML 13a, 13b, AAAI 15, CoNLL 11, 12]
My Research Contributions

**LIBLINEAR** [ICML08, KDD 08, JMLR 08a, 10a, 10b, 10c]

- Implements our proposed learning algorithms
- Supports *binary* and *multiclass* classification

**Impact:** > 60,000 downloads, > 2,600 citations in

- AI (AAAI, IJCAI), Data Mining (KDD, ICDM), Machine Learning (ICML, NIPS)
- Computer Vision (ICCV, CVPR), Information Retrieval (WWW, SIGIR),
- NLP (ACL, EMNLP), Multimedia (ACM-MM), HCI (UIST), System (CCS)
My Research Contributions

(Selective) Block Minimization

[KDD 10, 11, TKDD 12]

Supports learning from large data and streaming data

KDD best paper (2010), Yahoo! KSC award (2011)
My Research Contributions

Latent Representation for KBs

[EMNLP 13b,14]

Tensor methods for completing missing entries in KBs

Applications: e.g., entity relation extraction, word relation extraction.
My Research Contributions

Structured Prediction Models
[ECML 13a, 13b, ICML14, CoNLL 11,12, ECML 13a, AAAI15]

- Design tractable, principled, domain specific models
- Speedup general structured models
## Structured Prediction

Assign values to a set of interdependent output variables

<table>
<thead>
<tr>
<th>Task</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part-of-speech Tagging</td>
<td>They operate ships and banks.</td>
<td><img src="image" alt="Part-of-speech Tagging Output" /></td>
</tr>
<tr>
<td>Dependency Parsing</td>
<td>They operate ships and banks.</td>
<td><img src="image" alt="Dependency Parsing Output" /></td>
</tr>
<tr>
<td>Segmentation</td>
<td><img src="image" alt="Segmentation Input" /></td>
<td><img src="image" alt="Segmentation Output" /></td>
</tr>
</tbody>
</table>
Structured Prediction Models

- Learn a scoring function:
  \[ \text{Score} (\text{output } y \mid \text{input } x, \text{model } w) \]
- Linear model: \( S(y \mid x, w) = \sum_i w_i \phi_i(x, y) \)
- Features: e.g., Verb-Noun, Mary-Noun

Features based on both input and output
Inference

- Find the best scoring output given the model
  \[ \text{argmax } \text{Score} (output \ y \mid input \ x, \text{model } w) \]

- Output space is usually exponentially large

- Inference algorithms:
  - **Specific**: e.g., Viterbi (linear chain)
  - **General**: Integer linear programming (ILP)
  - **Approximate** inference algorithms: e.g., belief propagation, dual decomposition
Learning Structured Models

- **Online**, e.g., Structured Perceptron [Collins 02]
- **Batch** e.g., Structured SVM
  - Cutting plane: [Tsochantaridis+ 05, Joachims+ 09]
  - Dual Coordinate Descent: [Shevade+ 11, Chang+ 13]
  - Block-Coordinate Frank-Wolfe: [Lacoste-Julien+ 13]
  - Parallel Dual Coordinate Descent: [ECML 13a]

Solve inferences \hspace{2cm} Update the model
Outline

1. Applications:
   Co-reference; ESL Grammar Correction; Word Relation;

2. Modeling: Supervised Clustering Model

3. Algorithms: Learning with Amortized Inference
1. Applications:
   Co-reference; ESL Grammar Correction; Word Relation;

2. Modeling: Supervised Clustering Model

3. Algorithms: Learning with Amortized Inference
Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book
Co-reference Resolution

[EMNLP 13a, ICML14, In submission]

Proposed a novel, principled, linguistically motivated model

![Performance Graph]

*Avg (MUC, B³, CEAF)

Winner of the CoNLL ST 11

Winner of the CoNLL ST 12

Stanford
Chen+
Ours (2012)
Martschat+
Ours (2013)
Fernandes+
HOTCoref
Berkeley
Ours (2015)

Latent forest structure
Co-reference Resolution Demo

ESL Grammar Error Correction

[CoNLL 13, 14]

They believe that such situation must be avoided.

- situation
- a situation
- situations
- a situations

First place in CoNLL Shared tasks 13’ 14’
Identifying Relations between Words
[EMNLP 14]

- GRE antonym task (no context):

  Which word is the opposite of *adulterate*?
  (a) renounce  (b) forbid  (c) purify  (d) criticize  (e) correct

- Look up in a thesaurus [Encarta]: 56%

- Our tensor method [EMNLP 13b]: 77% (the best result so far)

- Why?

  - Considers multiple word relations simultaneously
    e.g., *inanimate* ← Ant → *alive* ← Syn → *living*
Antonym of *adulterate*?

(a) renounce  -0.014  
(b) forbid      0.004  
(c) purify       0.781  
(d) criticize  -0.004  
(e) correct    -0.010  

Outline

1. Applications:
   Co-reference; ESL Grammar Correction; Word Relation;

2. Modeling: Supervised Clustering Model

3. Algorithms: Learning with Amortized Inference
Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book
Co-reference Resolution

- Learn a pairwise similarity measure (local predictor)
  
  Example features:
  - same sub-string?
  - positions in the paragraph
  - other 30+ feature types

- Key questions:
  - How to learn the similarity function
  - How to do clustering

Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book.
Decoupling Approach

A heuristic to learn the model [Soon+ 01, Bengtson+ 08, CoNLL11]

- **Decouple** learning and inference:

  Learn a pairwise similarity function

  Cluster based on this function
As a boy, Chris$_1$ lived in a pretty home called Cotchfield Farm. When Chris$_2$ was three years old, his father$_3$ wrote a poem about him$_4$. The poem was printed in a magazine for others to read. Mr. Robin$_5$ then wrote a book.

Positive Samples
(Chris$_1$, him$_4$)
(Chris$_2$, him$_4$)
(Chris$_1$, Chris$_2$)
(his father$_3$, Mr. Robin$_5$)

Negative Samples
(Chris$_1$, his father$_3$)
(Chris$_2$, his father$_3$)
(him$_4$, his father$_3$)
(Chris$_1$, Mr. Robin$_5$)
(Chris$_2$, Mr. Robin$_5$)
(him$_4$, Mr. Robin$_5$)
[Bill Clinton], recently elected as the [President of the USA], has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [USA] and [Russia].
Greedy Best-Left-Link Clustering

[Bill Clinton], recently elected as the [President of the USA], has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [USA] and [Russia].
[Bill Clinton], recently elected as the [President of the USA], has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [US] and [Russia].
[Bill Clinton], recently elected as the [President of the USA], has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [USA] and [Russia].

Best Left-Linking Forest
Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book.
Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book.
Challenges

- In addition, we need world knowledge

As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him.

1. **Complexity**: need an efficient algorithm
2. **Modeling**: learn the metric while clustering
3. **Knowledge**: augment with knowledge
Structured Learning Approach

Learn the similarity function while clustering

Update the similarity function

Cluster based on this function.
Attempt: All-Links Clustering

[Mccallum+ 03, CoNLL 11]

- Define a global scoring function:
  - Attempt: using all within-cluster pairs:
  - Inference problem is too hard

Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book.
Latent Left-Linking Model (L3M)
[ICML 14, EMNLP 13]

Score (a clustering $C$)
$= \text{Score (the best left-linking forest that is consistent with } C)$
$= \sum \text{Score of edges in the forests}$

Christopher Robin is alive and well. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book
Linguistic Constraints

- Must-link constraints:
  - E.g., SameProperName, …

- Cannot-link constraints:
  - E.g., ModifierMismatch, …

[Bill Clinton], recently elected as the [President of the USA], has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [USA] and [Russia].
Inference in L3M [ICML 14, EMNLP 13]

- Solved by a **greedy** algorithm or formulated as an Integer Linear Program (ILP)

\[
\arg \max_y \sum_c S_{i,j} y_{i,j} \quad \text{s.t.} \quad A y \leq b; \quad y_{i,j} \in \{0,1\}
\]

\[y_{i,j} = 1 \iff i, j \text{ is an edge in the forest}\]

- Modeling constraints
- Linguistic constraints
[Bill Clinton], recently elected as the [President of the USA], has been invited by the [Russian President], [Vladimir Putin], to visit [Russia]. [President Clinton] said that [he] looks forward to strengthening ties between [USA] and [Russia].
Loop until stopping condition is met:

For each \((x_i, y_i)\) pair:

\[
\bar{y}, \bar{h} = \arg \max_{y, h} w^T \phi(x_i, y, h)
\]

\[
h_i = \arg \max_h w^T \phi(x_i, y_i, h)
\]

\[
w \leftarrow w + \eta(\phi(x_i, y_i, h_i) - \phi(x_i, \bar{y}, \bar{h}))\]

\(\eta\): learning rate
Define a log-linear model

\[
\Pr \left[ \text{a clustering } C \right] \\
= \sum \Pr \left[ \text{forests that are consistent with } C \right] \\
= \sum \sum \Pr \left[ \text{edges in the forest} \right]
\]

\[
\Pr \left[ \text{edge} \right] \sim \Pr \left[ \sum_{j \in e} \exp(w \cdot \phi(i, j)/\gamma) \right] \quad (\circ : \text{a parameter})
\]

Regularized Maximum Likelihood Estimation:

\[
\min_w LL(w) = \beta ||w||^2 + \sum_d \log Z_d(w) \\
- \sum_d \sum_i \log(\sum_{j<i} \exp(w \cdot \phi(i, j)/\gamma) C_d(i, j))
\]
Coreference: OntoNotes-5.0 (with gold mentions)

Performance*

Decoupled  L3M  Probabilistic L3M

*Avg (MUC, B³, CEAF)
Latent Left-Linking Model (L3M)
[ICML 14, EMNLP 13]

Advantages:
- **Complexity**: Very efficient
- **Modeling**: Learn the metric while clustering
- **Knowledge**: Easy to incorporate constraints (must-link or cannot-link)

Can be applied to other supervised clustering problems! e.g., the posts in a forum, error reports from users …
Outline

1. Applications:
   Co-reference; ESL Grammar Correction; Word Relation;

2. Modeling: Supervised Clustering Model

3. Algorithms: Learning with Amortized Inference
Learning Structured Models

- **Online**, e.g., Structured Perceptron [Collins 02]
- **Batch** e.g., Structured SVM
  - Cutting plane: [Tschantaridis+ 05, Joachims+ 09]
  - Dual Coordinate Descent: [Shevade+ 11, Chang+ 13]
  - Block-Coordinate Frank-Wolfe: [Lacoste-Julien+ 13]
  - Parallel Dual Coordinate Descent: [ECML 13a]

Solve inferences  Update the model
Redundancy in Learning Phase

Recognizing Entities and Relations Task

Counts

Inference problems

Distinct solutions

# training rounds
Redundancy of Solutions [Kundu+13]

<table>
<thead>
<tr>
<th>S1</th>
<th>POS</th>
<th>S2</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>Pronoun</td>
<td>She</td>
<td>Pronoun</td>
</tr>
<tr>
<td>is</td>
<td>VerbZ</td>
<td>is</td>
<td>VerbZ</td>
</tr>
<tr>
<td>reading</td>
<td>VerbG</td>
<td>watching</td>
<td>VerbG</td>
</tr>
<tr>
<td>a</td>
<td>Det</td>
<td>a</td>
<td>Det</td>
</tr>
<tr>
<td>book</td>
<td>Noun</td>
<td>movie</td>
<td>Noun</td>
</tr>
</tbody>
</table>

Although the inference problems are different, their solutions might be the same.
Fewer Inference Calls [AAAI 15]

Recognizing Entities and Relations Task

Obtain the same model with fewer inference calls

Performance

![Graph showing performance improvement with fewer inference calls](image-url)
Learning with Amortized Inference

[AAAI 15]

- A general inference framework ... to represent inference problems
- A condition ... to check if two problems have the same solution

If \text{CONDITION}(\text{problem cache, new problem})
then (no need to call the solver)
\text{SOLUTION}(\text{new problem}) = \text{old solution}

Else
Call \text{base solver} and update cache

End
A General Inference Framework

Integer Linear Programming (ILP)

$$\arg\max_y \sum_c S_c y_c \quad s.t \quad Ay \leq b; \quad y_c \in \{0,1\}$$

- Widely used in NLP & Vision tasks [Roth+ 04]
  - E.g., Dependency Parsing, Sentence Compression
- Any MAP problem w.r.t. any probabilistic model, can be formulated as an ILP [Roth+ 04, Sontag 10]
- Only used for verifying amortized conditions
Amortized Inference Theorem [Kundu+13]

- **Theorem 1:** If the following conditions are satisfied
  1. Same # variables & same constraints
  2. \( \forall i, \quad (2x_{p,i}^* - 1)(c_{Q,i} - c_{P,i}) \geq 0 \)

(The solution is not sensitive to the changes of the coefficients.)

Then the optimal solution of Q is \( x_p^* \)

- \( x_p^* \): the solution to P
- \( c \): the coefficients of ILPs
Amortized Inference Theorem [Kundu+13]

- **Theorem 1**: If the following conditions are satisfied
  1. Same # variables & same constraints (same equivalence class)
  2. \( \forall i, (2x^*_p,i - 1)(c_{Q,i} - c_{P,i}) \geq 0 \)

then the optimal solution of Q is \( x^*_p \)

\[
\begin{align*}
\text{P:} & \quad \begin{cases}
\max 2x_1 + 3x_2 + 2x_3 + 1x_4 \\
x_1 + x_2 \leq 1 \\
x_3 + x_4 \leq 1
\end{cases} \\
x^*_p: & \quad <0, 1, 1, 0> \\
x': & \quad <1, 0, 1, 0>
\end{align*}
\]

\[
\begin{align*}
\text{Q:} & \quad \begin{cases}
\max 2x_1 + 4x_2 + 2x_3 + 0.5x_4 \\
x_1 + x_2 \leq 1 \\
x_3 + x_4 \leq 1
\end{cases} \\
\quad & \quad 6 \\
\quad & \quad 4
\end{align*}
\]
Amortized Inference Theorem [Kundu+13]

Theorem 1: If the following conditions are satisfied

1. Same # variables & same constraints
2. \( \forall i, (2x_{p,i}^* - 1)(c_{Q,i} - c_{P,i}) \geq 0 \)

if \( x_{p,i}^* = 1 \) then \( (c_{Q,i} - c_{P,i}) \geq 0 \)
if \( x_{p,i}^* = 0 \) then \( (c_{Q,i} - c_{P,i}) \leq 0 \)

then the optimal solution of Q is \( x_p^* \)

- \( x_p^* \): the solution to P
- \( c \): the coefficients of ILPs
Approx. Amortized Inference [AAAI 15]

- **Theorem 2:** If the following conditions are satisfied
  1. Same # variables & same constraints
  2. \( \forall i, \ (2x_{p,i}^* - 1)(c_{Q,i} - c_{P,i}) \geq -\epsilon |c_{Q,i}| \)

then \( x_P^* \) is a \( \left( \frac{1}{1 + M\epsilon} \right) \)-approximate solution to Q

- \( x_P^* \): the solution to P
- \( M \): a constant
- \( c \): the coefficients of ILPs
Approx. Amortized Inference [AAAI 15]

Theorem 2: If the following conditions are satisfied

1. Same # variables & same constraints
2. \( \forall i, (2x^*_{p,i} - 1)(c_{Q,i} - c_{P,i}) \geq -\epsilon |c_{Q,i}| \)

then \( x^*_p \) is a \( \left( \frac{1}{1 + M\epsilon} \right) \)-approximate solution to Q

Corollary 1:
Learning Structured SVM with approximate amortized inference gives a model with bounded empirical risk
Approx. Amortized Inference [AAAI 15]

- **Theorem 2**: If the following conditions are satisfied
  1. Same # variables & same constraints
  2. \( \forall i, (2x^*_p,i - 1)(c_{Q,i} - c_{P,i}) \geq -\epsilon |c_{Q,i}| \)

  then \( x^*_p \) is a \( \left( \frac{1}{1+M\epsilon} \right) \)-approximate solution to \( Q \)

**Corollary 2:**
Dual coordinate descent for structured SVM can still return an exact model even if approx. amortized inference is used.
# Solver Calls (Entity-Relation Extraction)

- **Exact**: Ent F1: 87.7, Rel F1: 47.6
- **Better**: Ent F1: 87.3, Rel F1: 47.8
Outline

1. Applications:
   Co-reference; ESL Grammar Correction; Word Relation;

2. Modeling: Supervised Clustering Model

3. Algorithms: Learning with Amortized Inference
1. **Applications:** Co-reference; ESL Grammar Correction; Word Relation; Dependency Parsing [Arxiv 15 b]; Multi-label Classification [ECML13]

2. **Modeling:** Supervised Clustering Model
   - Semi-Supervised Learning [ECML 13a]
   - Search-Based Model [Arxiv 15 a]

3. **Algorithms:** Learning with Amortized Inference
   - Parallel learning algorithms [ECML 13b]
My Research Contributions

Limited memory linear classifier [KDD 10, 11, TKDD 12]

Latent representation for knowledge bases [EMNLP 13, 14]

Linear classification [ICML08, KDD 08, JMLR 08a, 10a, 10b, 10c]

Structured prediction models [ICML 14, ECML 13a, 13b, AAAI 15, CoNLL 11, 12]
Future Work: Practical Machine Learning

1. **Applications:** More applications, easy access tools

2. **Modeling:** Learning from heterogeneous information

3. **Algorithms:** Handle large & complex data
Learning From World Knowledge

- Go beyond supervised learning
  - Learning from indirect supervision signals

After the *vessel* suffered a catastrophic torpedo detonation, *Kursk* sank in the waters of Barents Sea with *all hands* lost.
Learning From World Knowledge

- Massive textual data on the Internet
  - Wikipedia: 4.7 M English articles 35M in total
  - Tweets: 500 M per day & 200 Billion per year

- Learn world knowledge to support target tasks
  - Extract knowledge from free text [EMNLP 13a, 14, ICML 14]
  - Handle large-scale data [Liblinear, KDD 12]
  - Inference on knowledge bases [EMNLP 14b, 14]
Applications & Tools

- **LIBLINEAR**: library for classification
- **Streaming Data SVM**:  
  - Support training on very large data
- **Illinois-SL**: library for structured prediction  
  - Support various algorithms; parallel $\Rightarrow$ very fast

Provide a nice platform
- for developing novel methods
- for collaboration
- for education

More easy-access tools; More collaborations
Conclusion

**Goal:** Practical Machine Learning

- **[Modeling]** Expressive and general formulations
- **[Algorithms]** Principled and efficient
- **[Applications]** Support many applications

Code and Demos:

http://www.illinois.edu/~kchang10