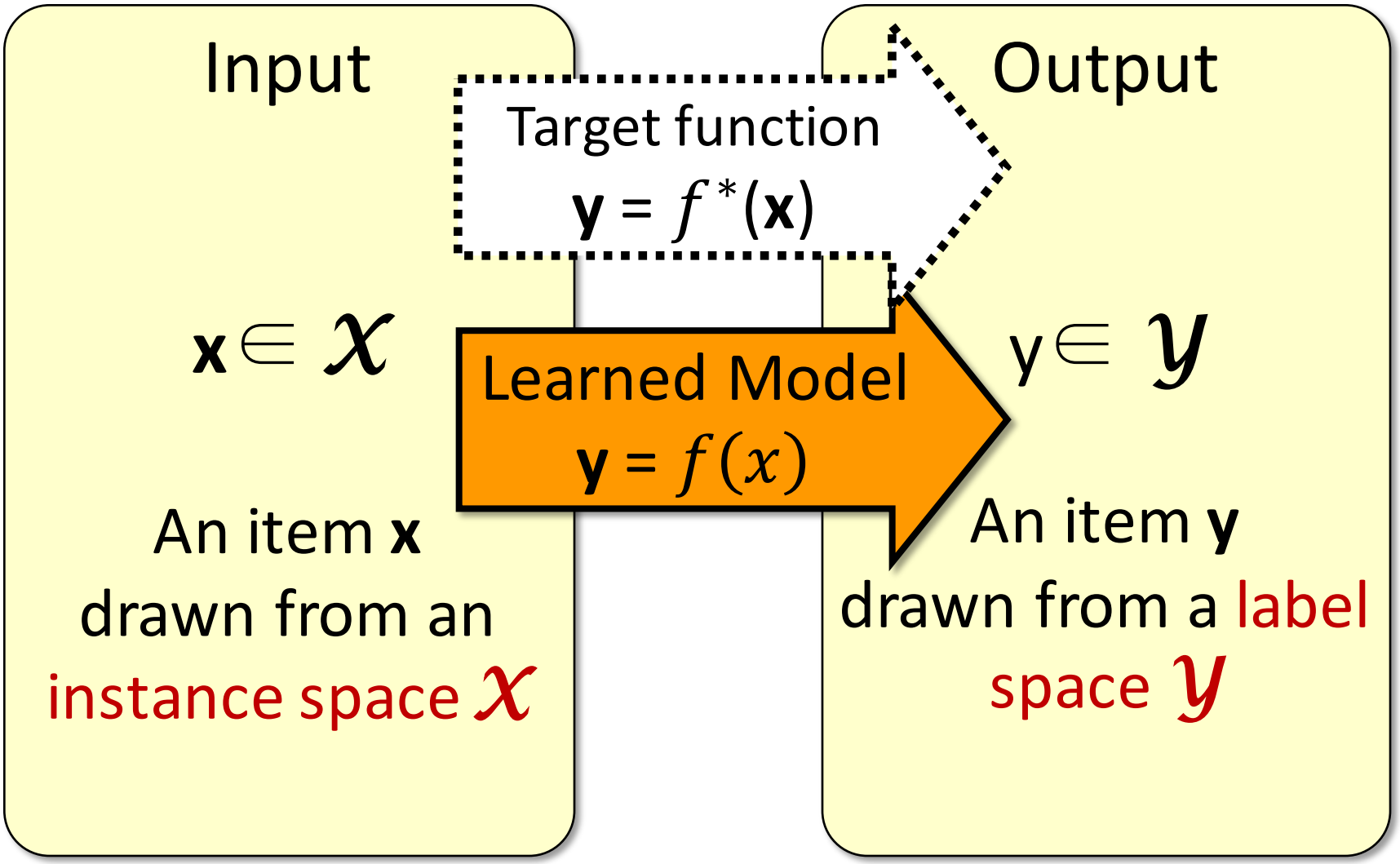
Structured Predictions:
Practical Advancements and Applications

Kai-Wei Chang

University of Virginia
Department of Computer Science

References: http://kwchang.net/talks/sp.html

Supervised learning



Q: [Chris] = [Mr. Robin] ?

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

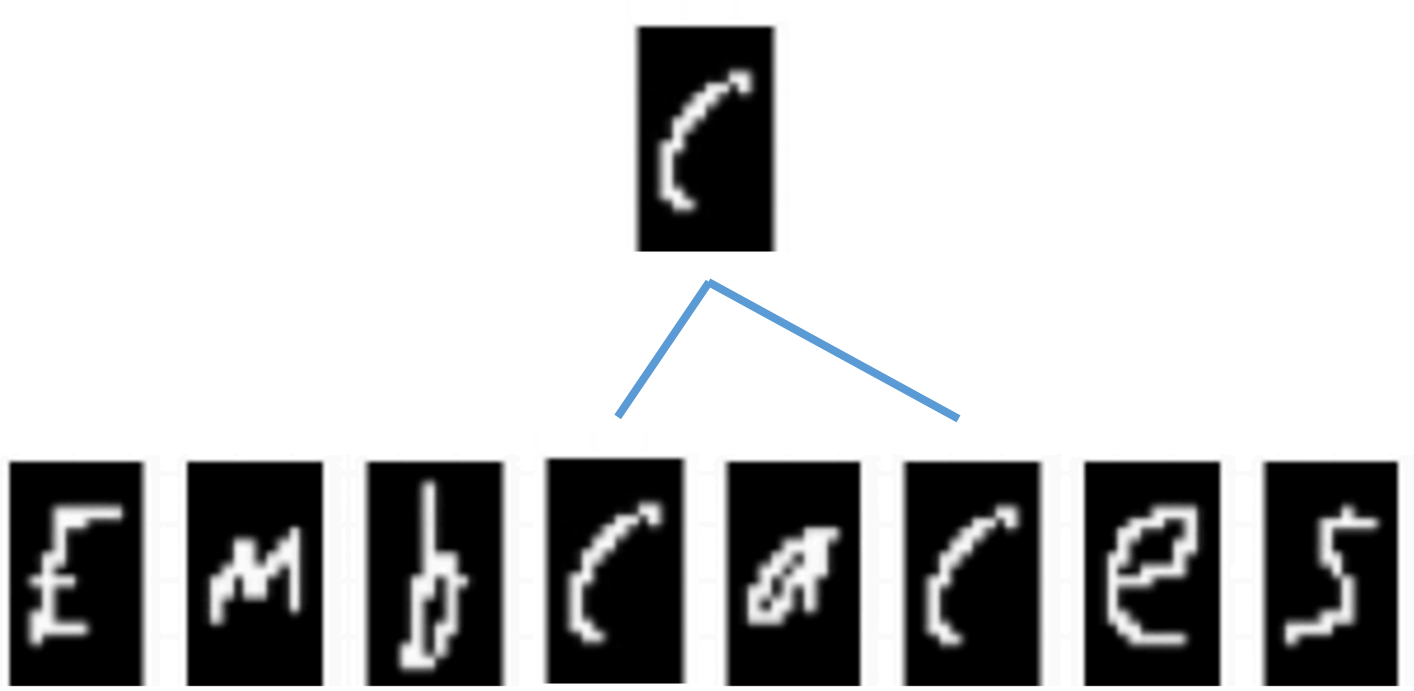
Slide modified from Dan Roth

Complex Decision Structure

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

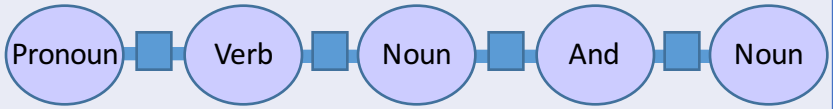
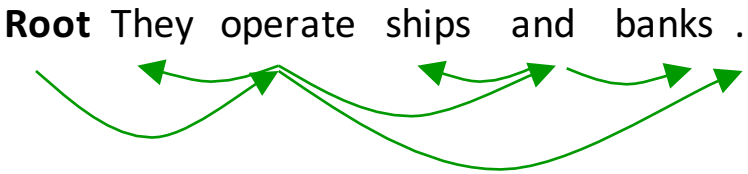
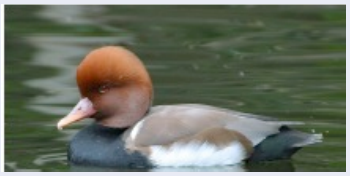
Why is structure important?
Hand written recognition example

- What is this letter?

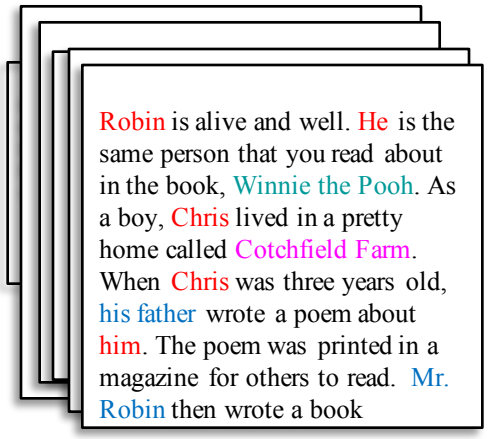


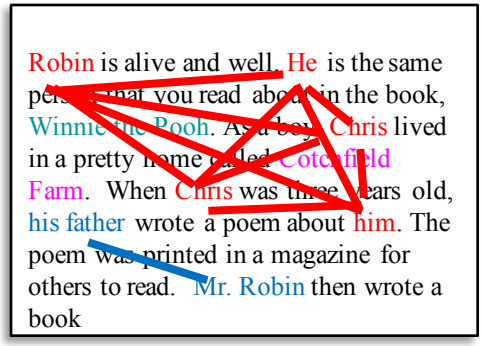
Structured Prediction

Assign values to a set of interdependent output variables

Task	Input	Output
Part-of-speech Tagging	They operate ships and banks.	
Dependency Parsing	They operate ships and banks.	
Segmentation		

Challenge: Scalability Issues





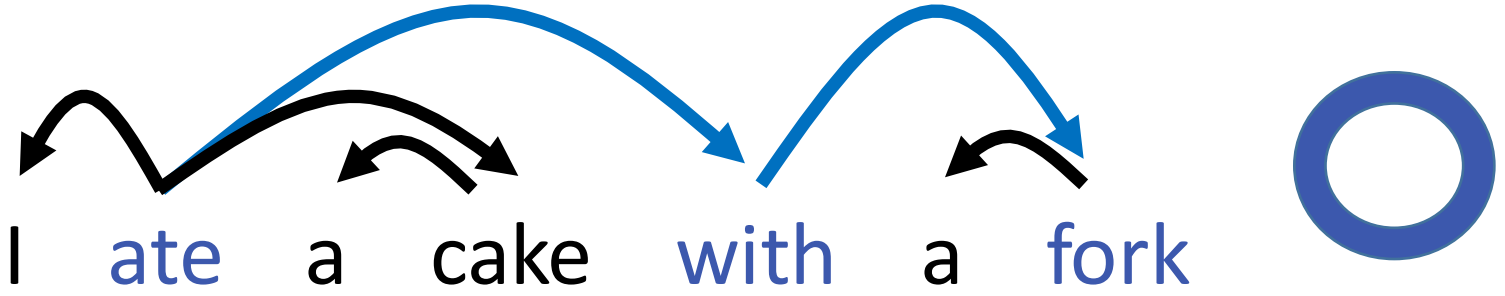
- Large amount of data
- Complex decision structure

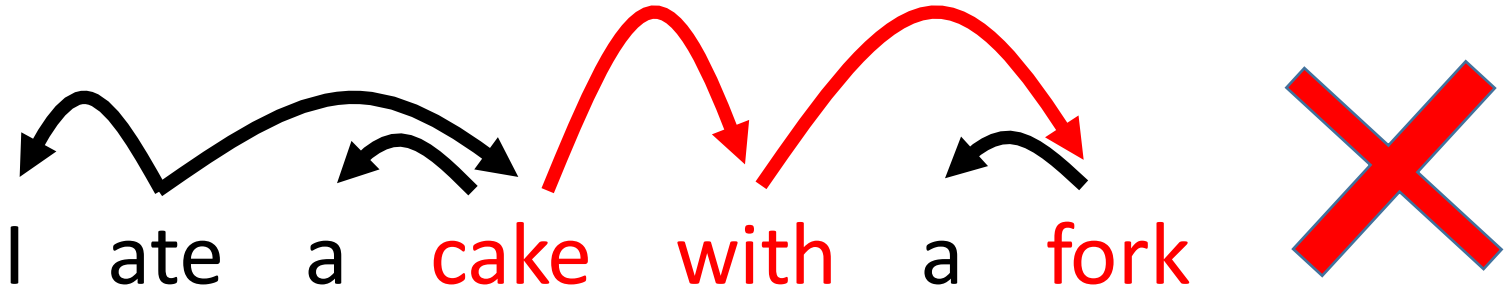
Solution Methods

- Assume a graphical structure; optimize
  - Use within various structured predictions algorithms (e.g., CRF, Structured Perceptron, M3N, Structured SVM)
    [Lafferty+ 01, Collins02, Taskar04]
  - See our AAAI16 tutorial (https://goo.gl/TF7cGj)
- Learning to search approaches
  - Assume the complex decision is incrementally constructed by a sequence of decisions
  - E.g., LASO, dagger, Searn, transition-based methods
  - See our NAACL15 tutorials (http://hunch.net/~l2s)

Example: Dependency Parsing

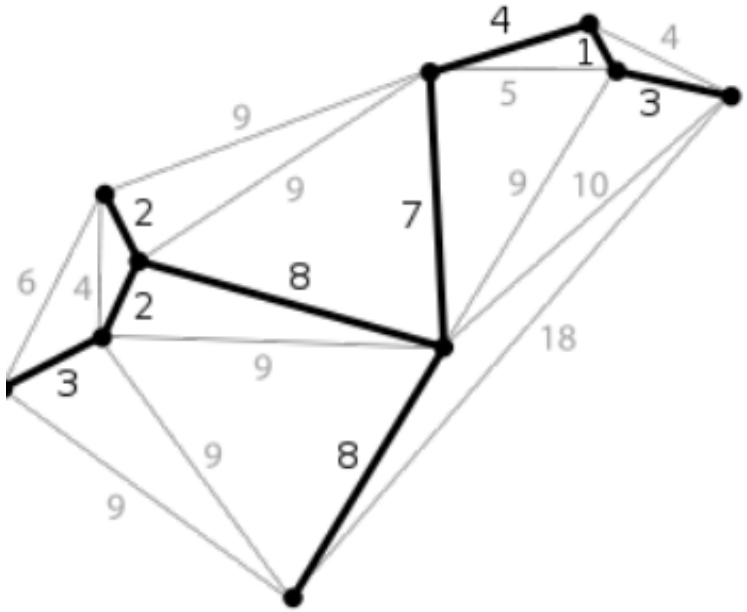
- Identifying relations between words





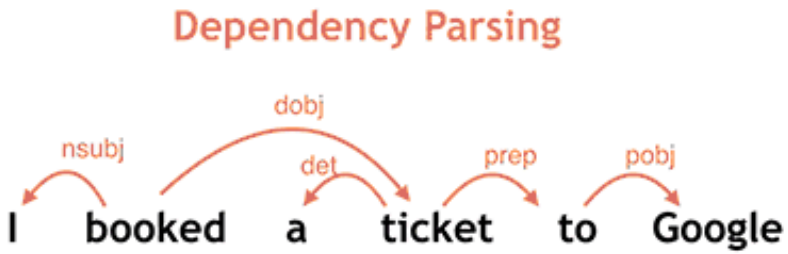
Graphical Model Approaches: Graph-Based Parser [McDonald+. 2005]

- Consider all word pairs and assign scores
- Score of a tree = sum of score of edges
- Can be formulated as a MST problem
  - Chu-Liu-Edmonds



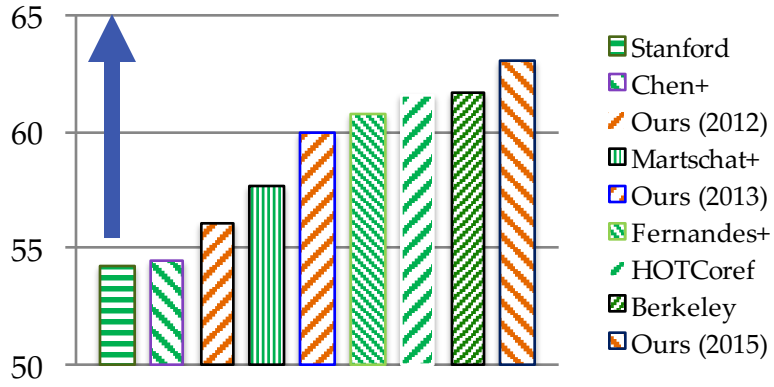
Learning to search approaches Shift-Reduce parser[Nivre03,NIPS16]

- Maintain a buffer and a stack
- Make predictions from left to right
- Three (four) types of actions:
  Shift, Reduce-Left, Reduce-Right

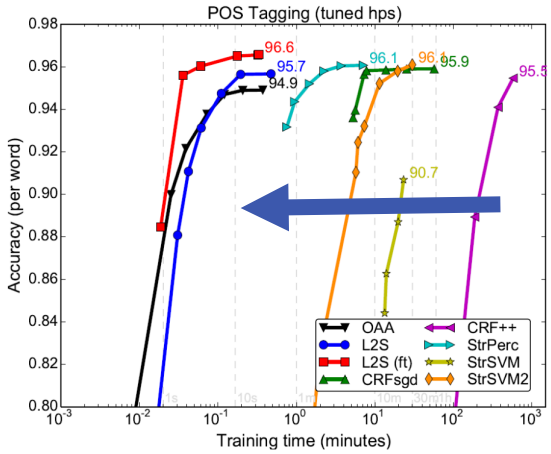


Credit: Google research blog

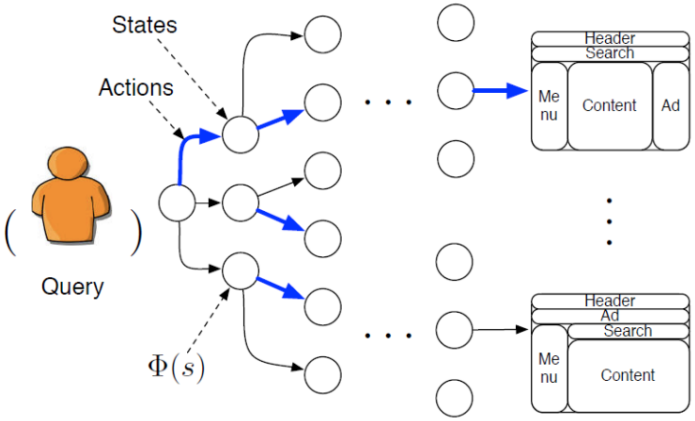
What We Care about



Prediction accuracy



Training/test/dev speed

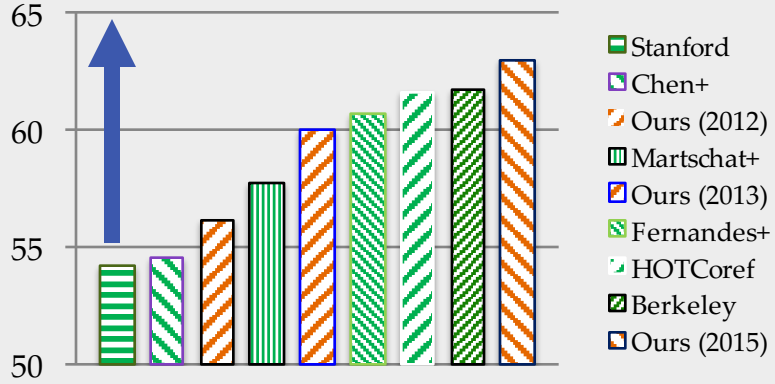


Learning signals

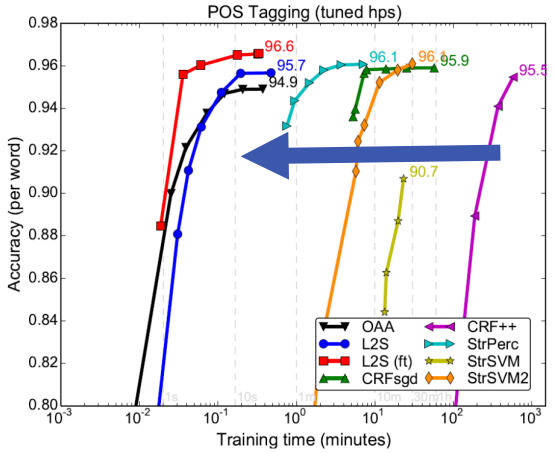
activity	**cooking**
agent	**woman**
food	**vegetable**

Fairness (data biases)

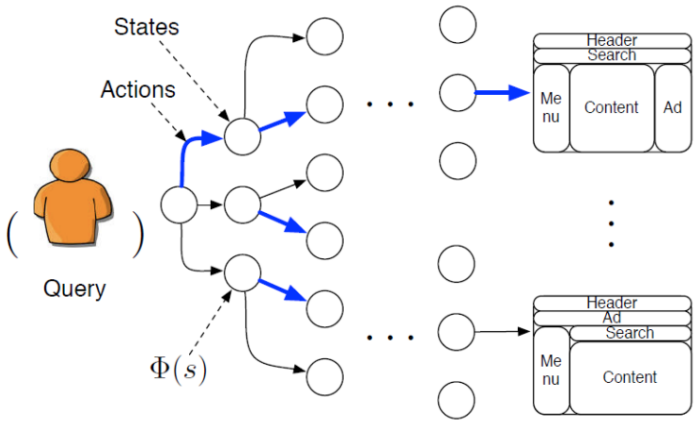
Outline



Prediction accuracy



Training/test/dev speed



Learning signals

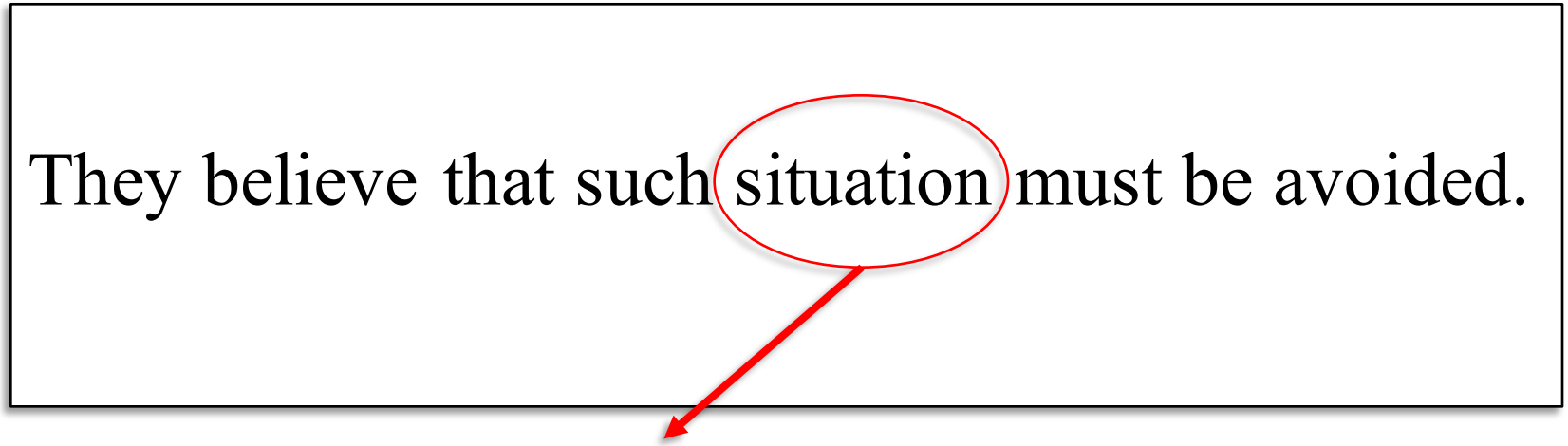
Query

activity	cooking
agent	woman
food	vegetable

Fairness (data biases)

Structured prediction application: ESL Grammar Error Correction

[CoNLL 13, 14]



- ✗ situation
- ✓ a situation
- ✓ situations
- ✗ a situations

Structured prediction application: Algebra Word Problems [EMNLP 16]

Problem: Maria is now four times as old as Kate. Four years ago, Maria was six times as old as Kate. Find their ages now.

Equations: $m = 4 \times n$ and $m - 4 = 6 \times (n - 4)$

Solution: $m = 40, n = 10$

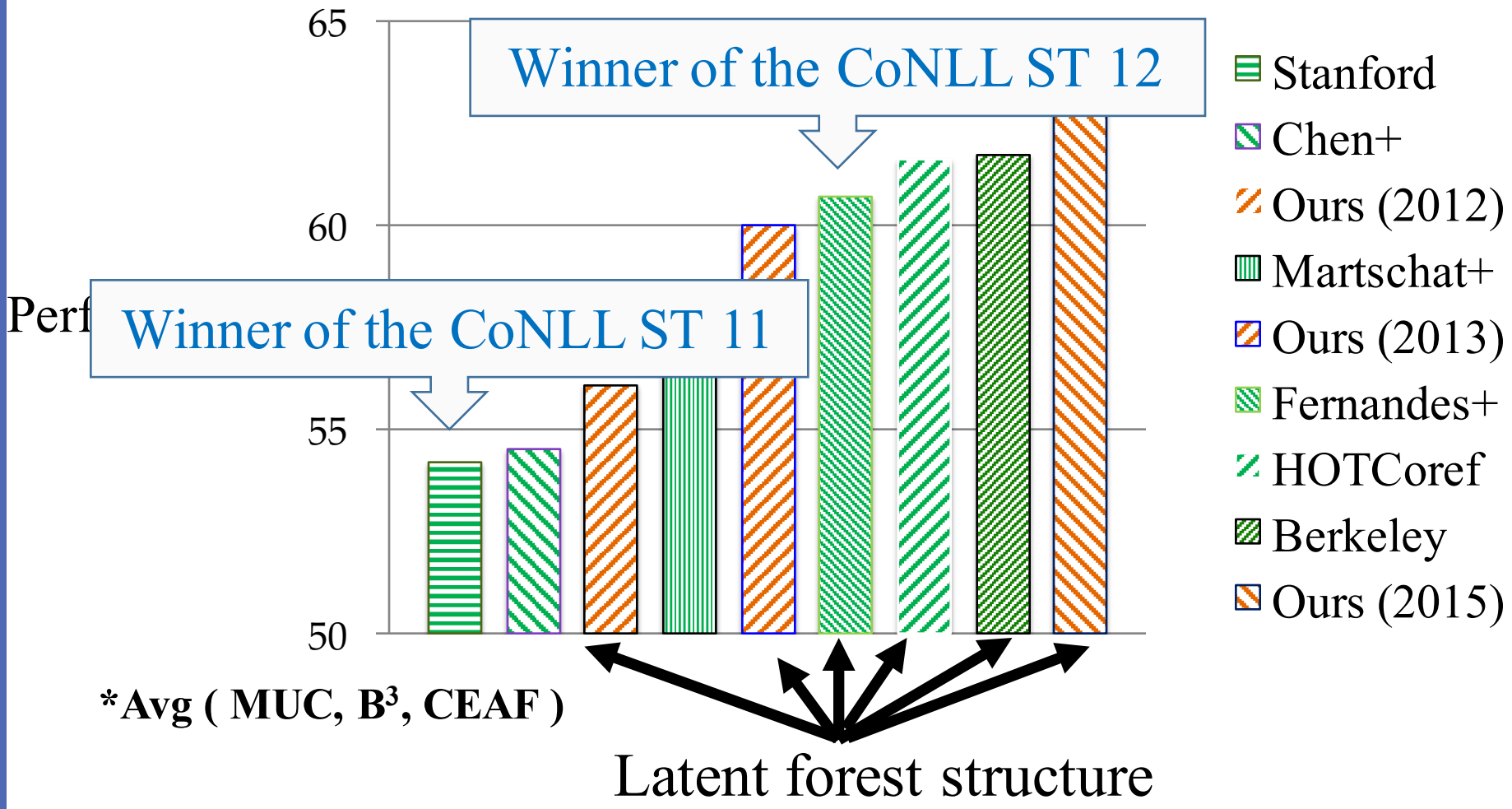
Structured prediction application: Co-reference Resolution

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

Structured prediction application:
Co-reference Resolution

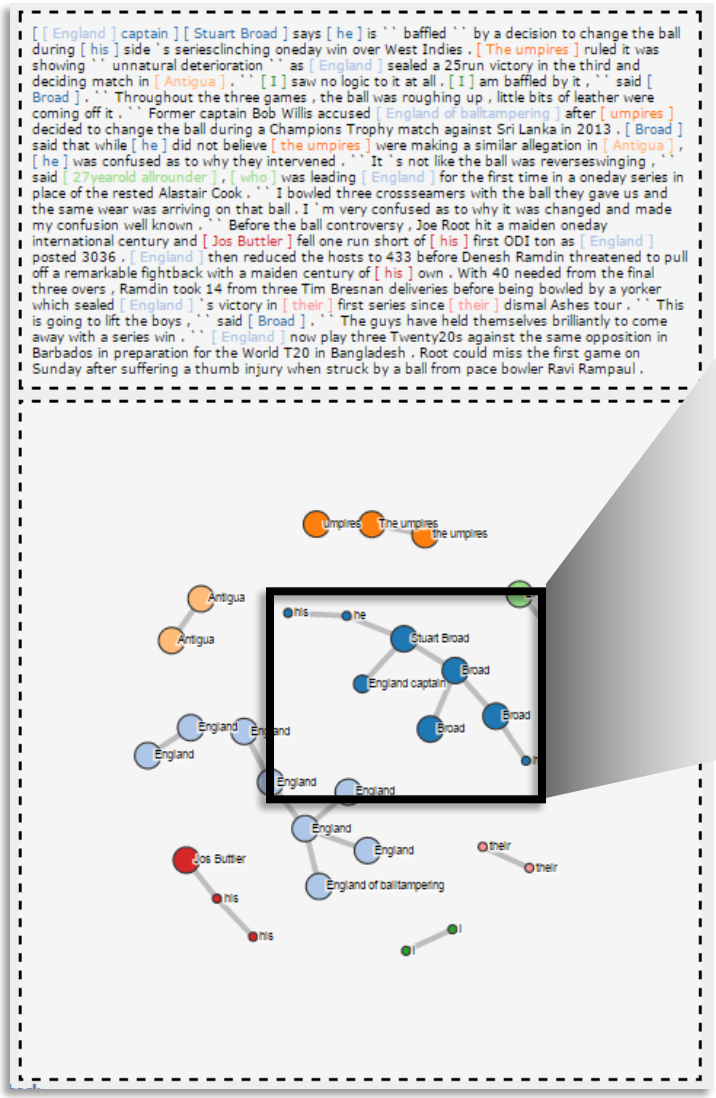
[EMNLP 13a, ICML 14, CoNLL 11,12, 15]

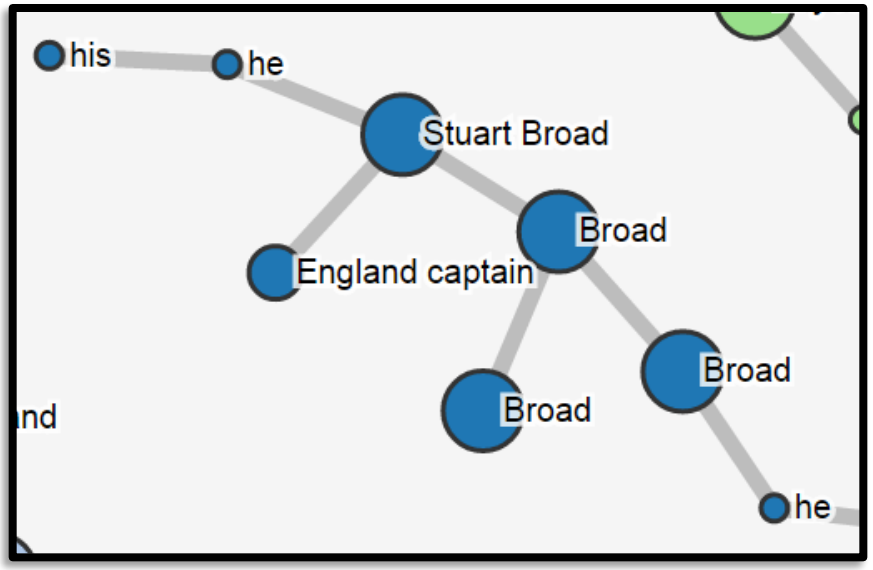
Proposed a novel, principled, linguistically motivated model



The state-of-the-art approach using NN&RL achieves 65.73(Clark+16)

Co-reference Resolution Demo





http://bit.ly/illinoisCoref

Co-reference Resolution

- Learn a pairwise similarity meas (local predictor)

  Example features:
  - same sub-string?
  - positions in the paragraph
  - other 30+ feature types
- Key components:
  - Pairwise classification
  - Clustering (jointly or not?)

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

Decoupling Approach-Learning

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$_5$)

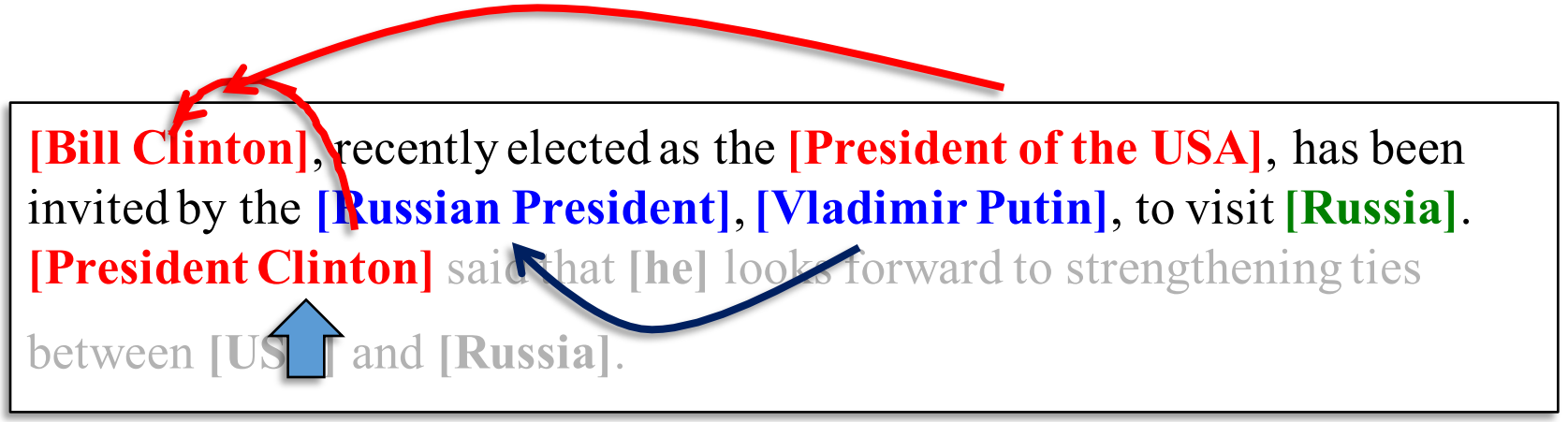
Greedy Best-Left-Link Clustering

[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].

Greedy Best-Left-Link Clustering

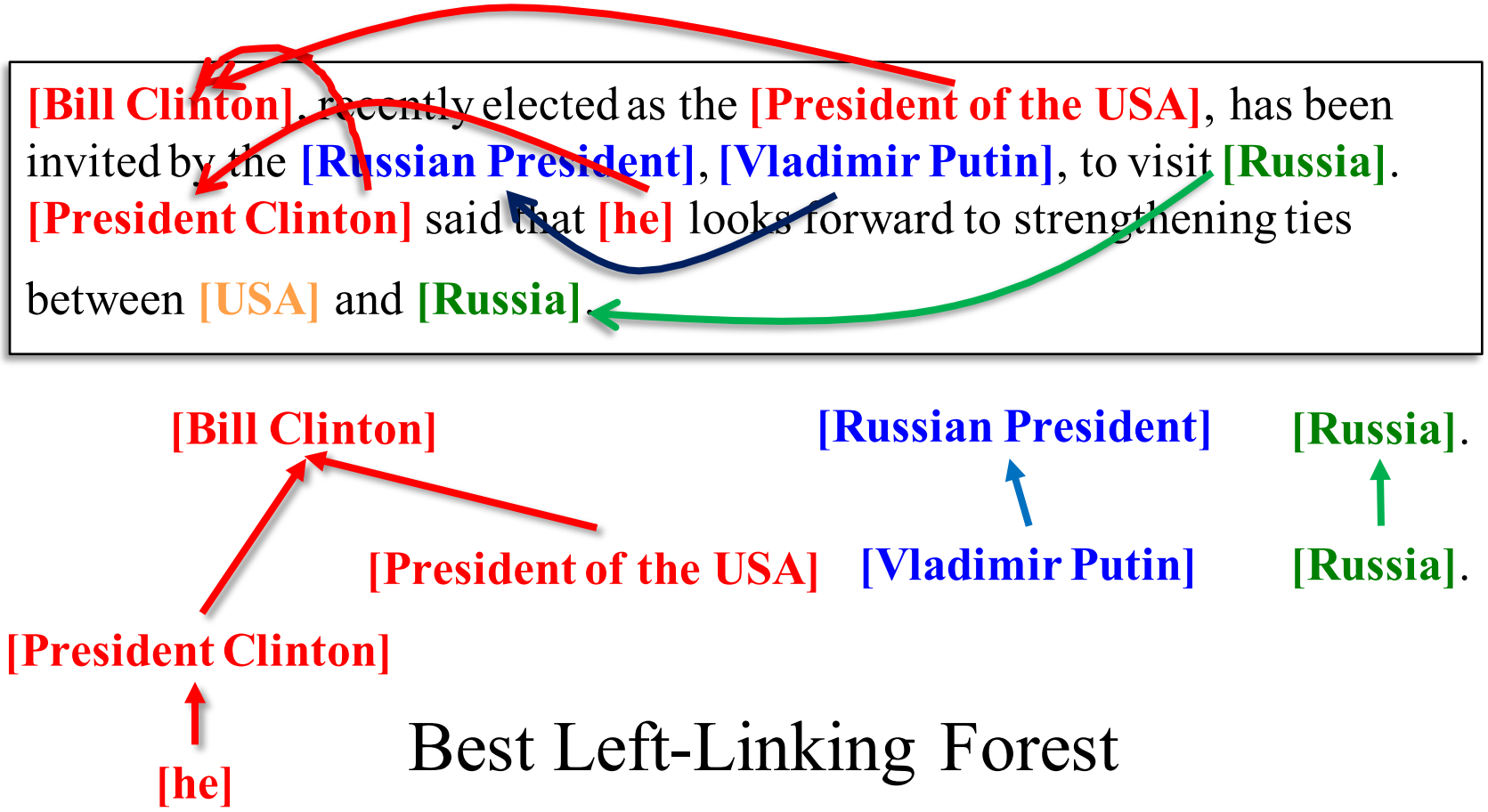
[Bill Clinton], recently elected as the **[President of the USA]**, has been invited by the **[Russian President]**, **[Vladimir Putin]**, **to visit [Russia]. [President Clinton]** said **that [he] looks forward to strengthening** ties between **[USA]** and **[Russia]**.

Greedy Best-Left-Link Clustering



Greedy Best-Left-Link Clustering

[Soon+ 01, Bengtson+ 08, CoNLL11]



Challenges

- Decoupling may lose information

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

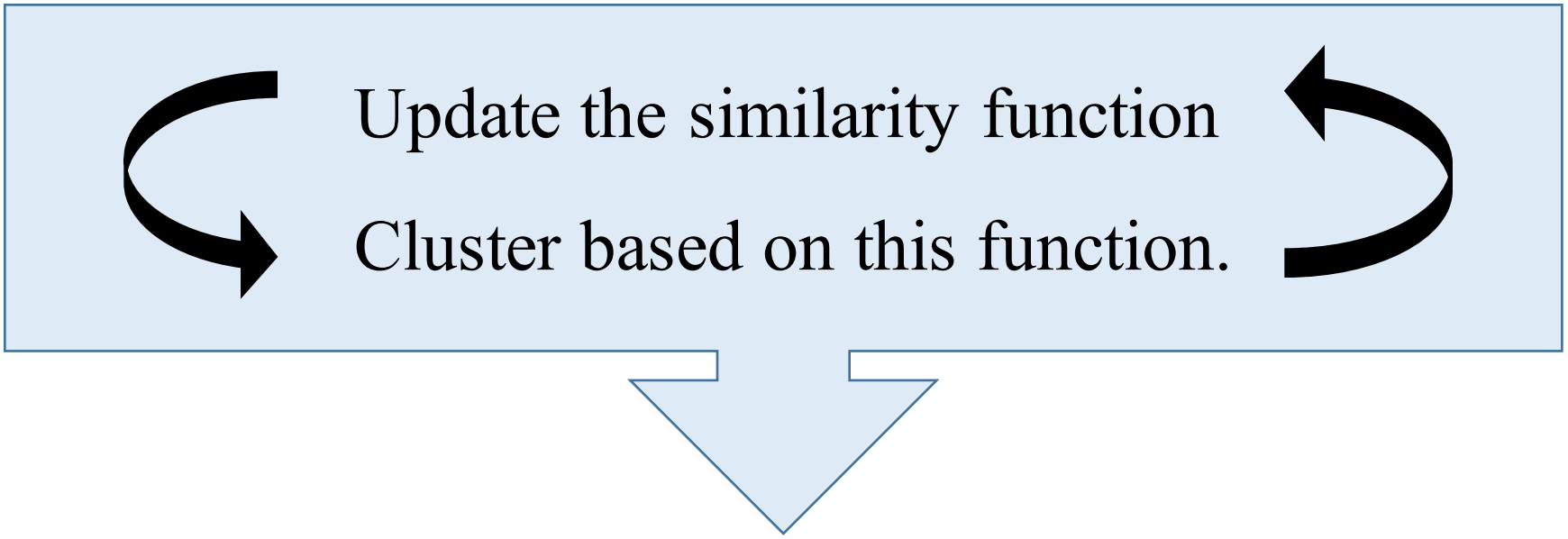
Challenges

- 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

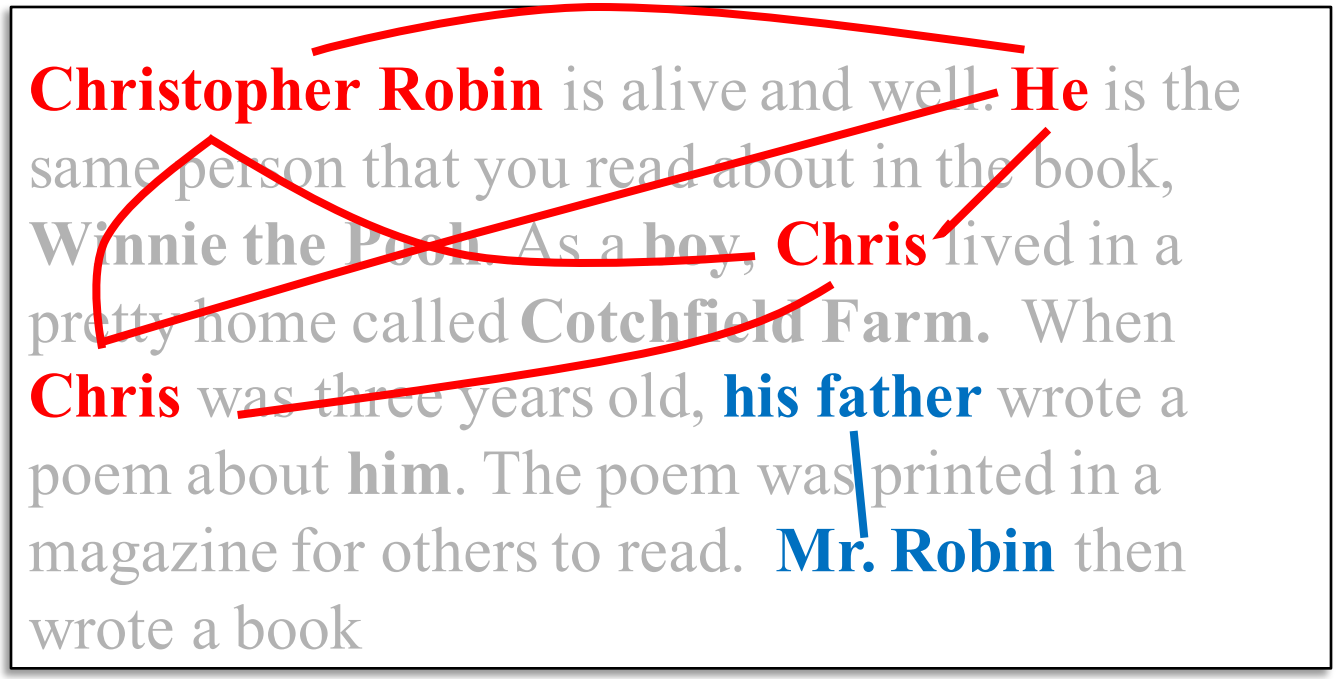
Attempt: All-Links Clustering

[Mccallum+ 04, CoNLL 11]

- Define a global scoring function:

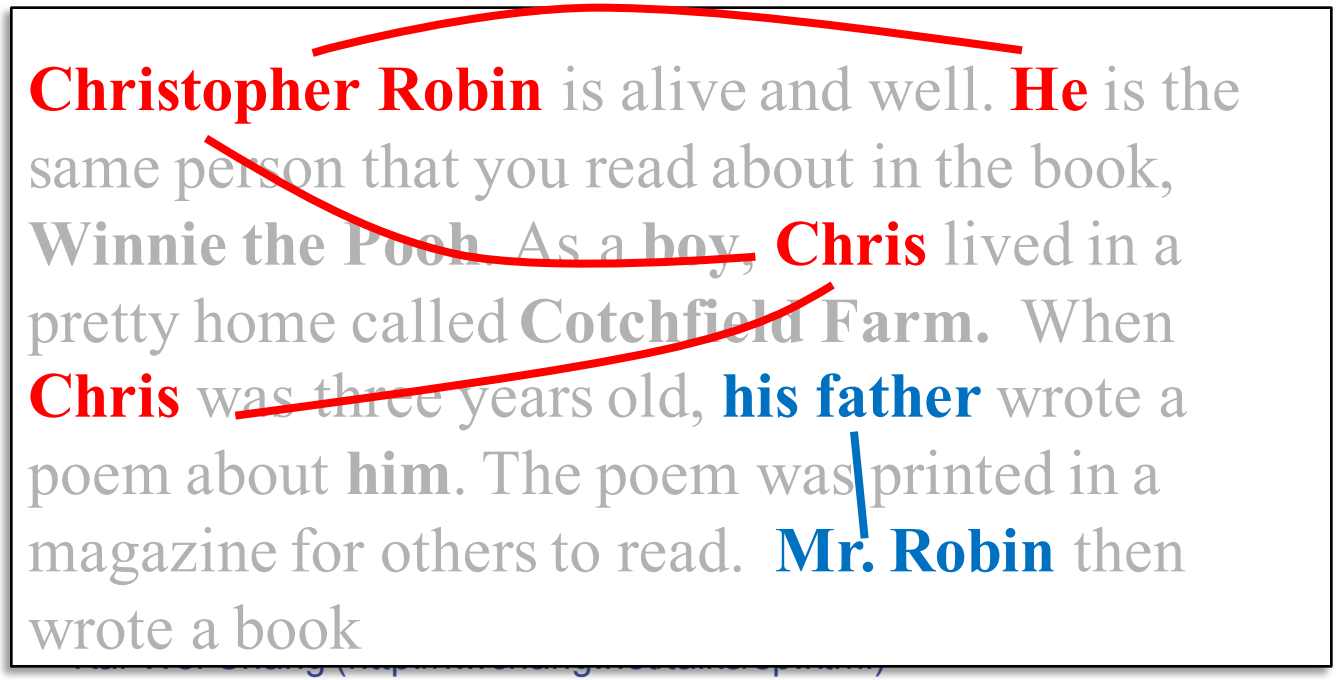
  Attempt: using all within-cluster pairs:

  - Inference problem is too hard



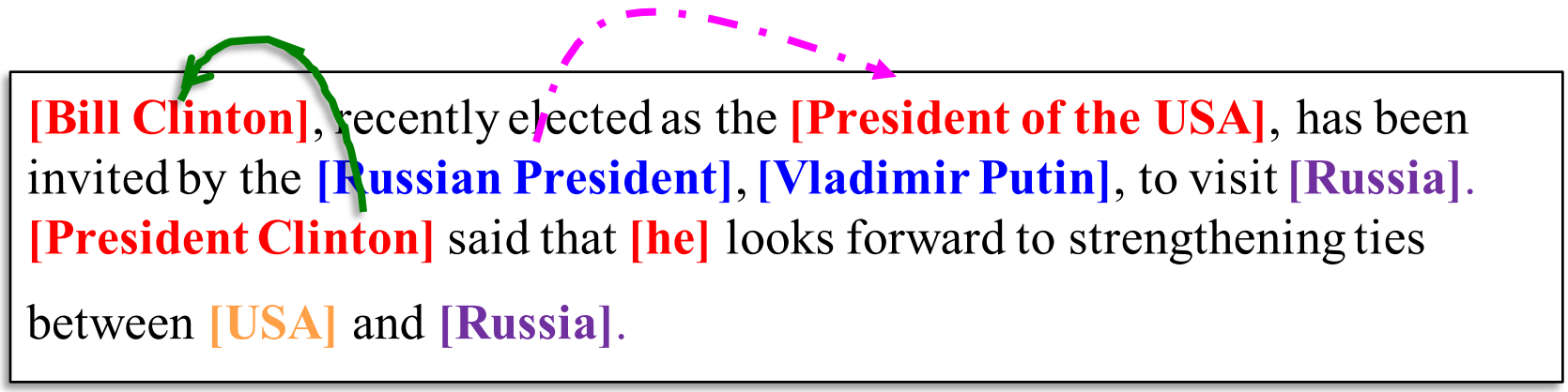
Latent Left-Linking Model (L3M)
[ICML 14, EMNLP 13]

Score (a clustering C)
= Score (the best left-linking forest that is consistent with C)
= ∑ Score of edges in the forests



Linguistic Constraints

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



- Clustering with constraints [(Basu+08, Zhi+14]

Inference in L3M [ICML 14, EMNLP 13]

- Represented using an ILP formulation [Scott+ 2004/2007]
- Inference can be done using a greedy heuristics.

$y_{i,j} = 1 \Leftrightarrow i,j$ is an edge in the forest

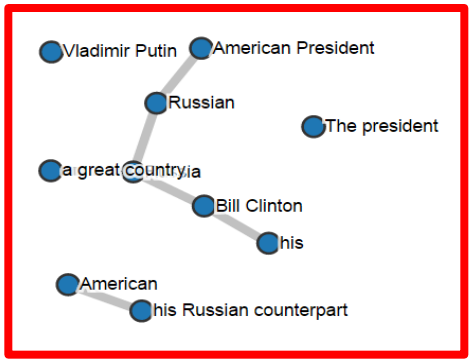
$$\arg\max_{y} \sum_c S_{i,j} y_{i,j}$$

$$s.t \quad Ay \leq b; \quad y_{i,j} \in \{0,1\}$$

- Modeling constraints
- Linguistic constraints

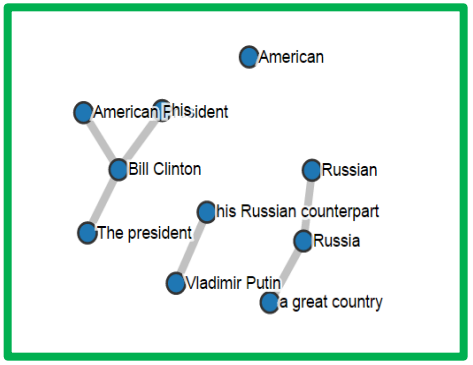
Learning L3M (simplified version)[ICML 14, EMNLP 13a]

[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]**.



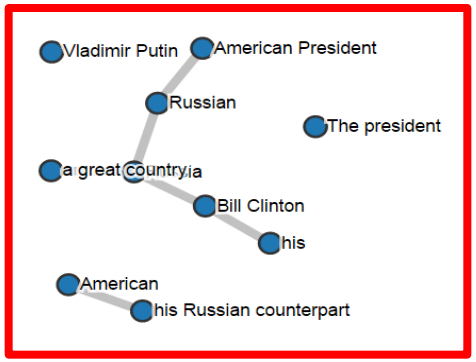
predicted forest

[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]**.

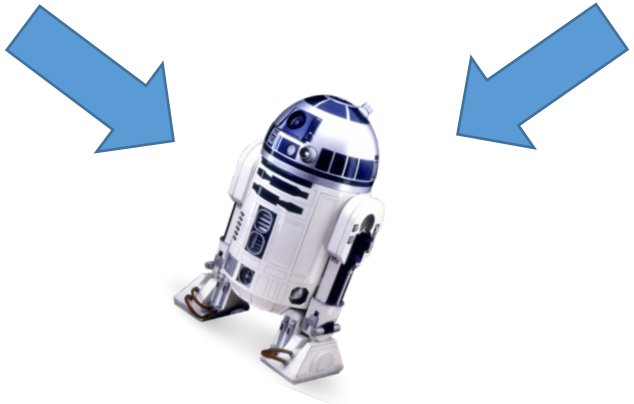


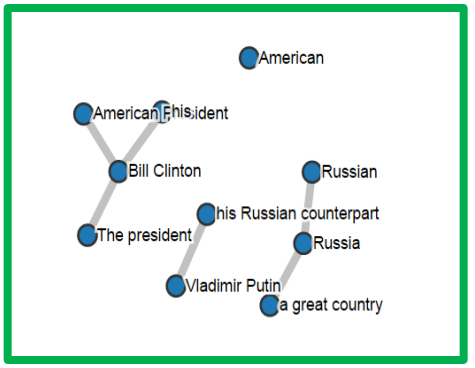
latent forest

Learning L3M (simplified version)[ICML 14, EMNLP 13a]



predicted forest





latent forest

Loop until stopping condition is met:

For each $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ pair:

$$\bar{\boldsymbol{y}}, \bar{\boldsymbol{h}} = \arg\max_{\boldsymbol{y},\boldsymbol{h}} \boldsymbol{w}^T \phi(\boldsymbol{x}_i, \boldsymbol{y}, \boldsymbol{h})$$

$$\mathbf{h_i} = \arg\max_{\boldsymbol{h}} \boldsymbol{w}^T \phi(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{h})$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta(\phi(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{h}_i) - \phi(\boldsymbol{x}_i, \bar{\boldsymbol{y}}, \bar{\boldsymbol{h}})), \ \eta\text{: learning rate}$$

Extension: Probabilistic L3M

[ICML 14, EMNLP 13a]

- Define a log-linear model

Pr [a clustering C]

Pr [a clustering C]

$= \sum$ Pr [forests that are consistent with C]

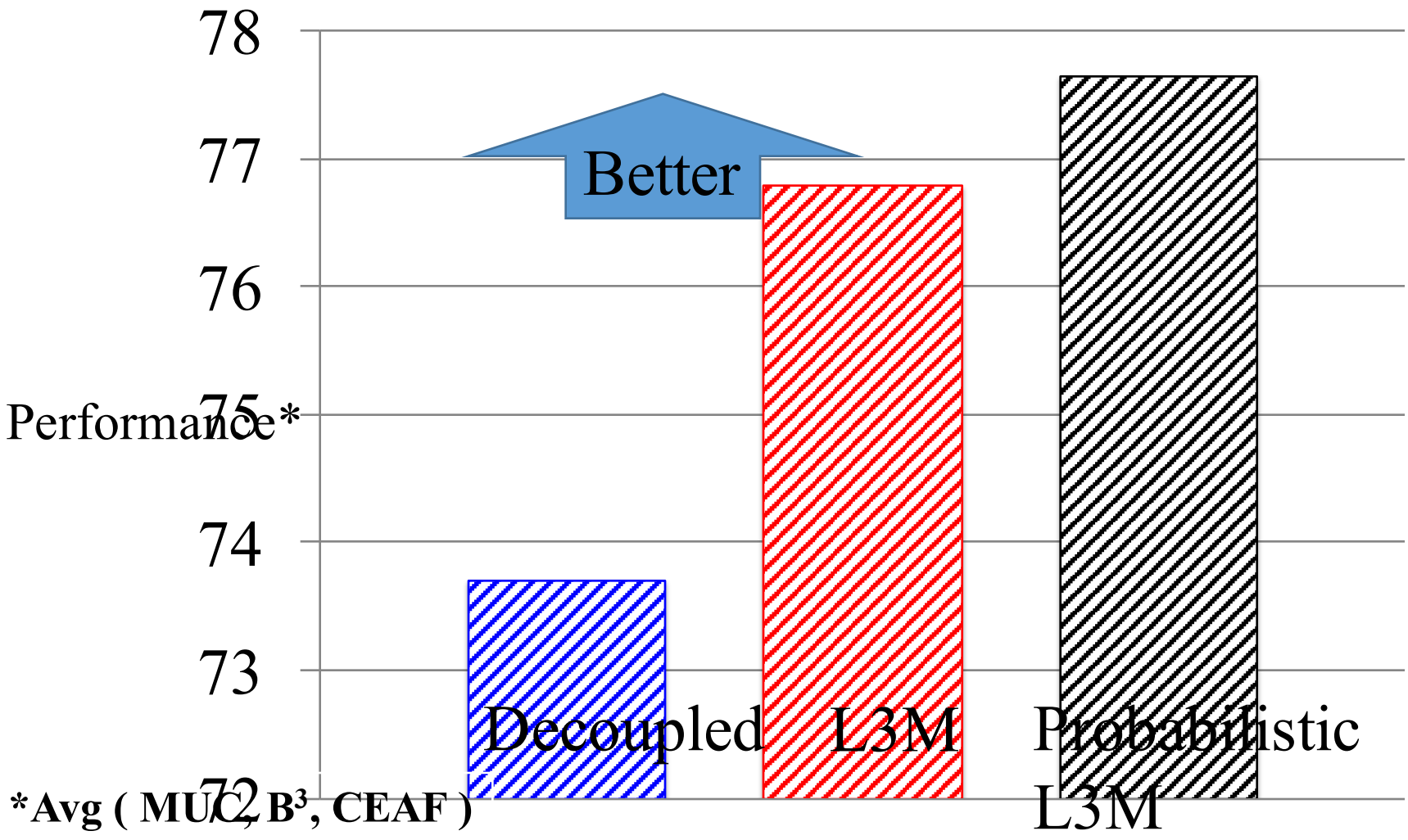
$= \sum\prod$ Pr [edges in the forest]

$= \prod_i \sum_{j \in e(i)}$ Pr [edge(j,i)]

Pr [edge(j,i)] $\sim \exp(\boldsymbol{w} \cdot \phi(j,i)/\gamma)$ (γ: a parameter)

$$\min_{\mathbf{w}} \; \mathrm{LL}(\mathbf{w}) = \beta||\mathbf{w}||^2 + \sum_d \log Z_d(\boldsymbol{w}) - \sum_d \sum_i \log\left(\sum_{j<i} \exp(\boldsymbol{w} \cdot \phi(i,j)/\gamma) C_d(i,j)\right)$$

Coreference: OntoNotes-5.0 (with gold mentions)



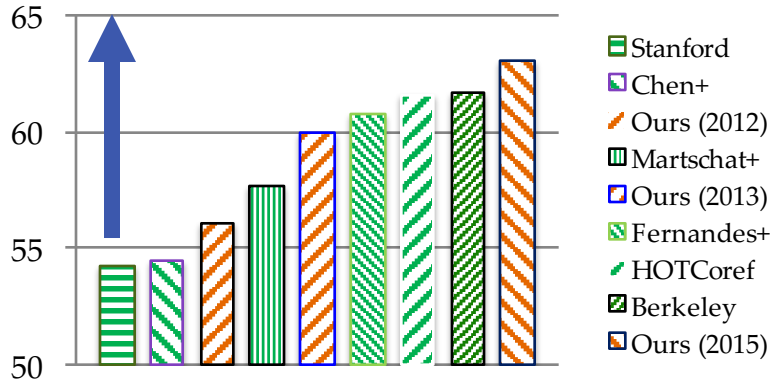
Latent Left-Linking Model (L3M)
[ICML 14, EMNLP 13]

Advantages:

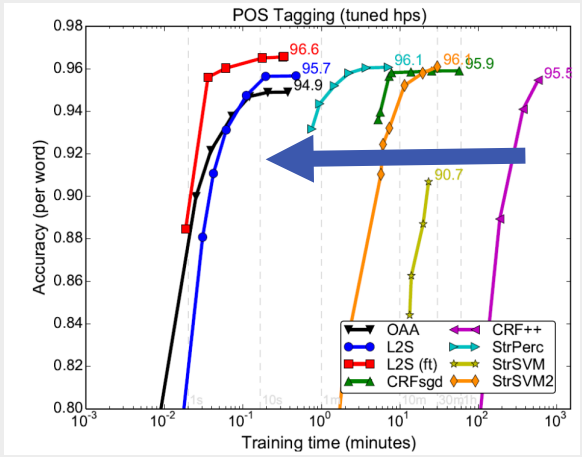
- Complexity: Very efficient
- Modeling: Learn the metric while clustering
- Knowledge: Easy to incorporate constraints (must-link or cannot-link)

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

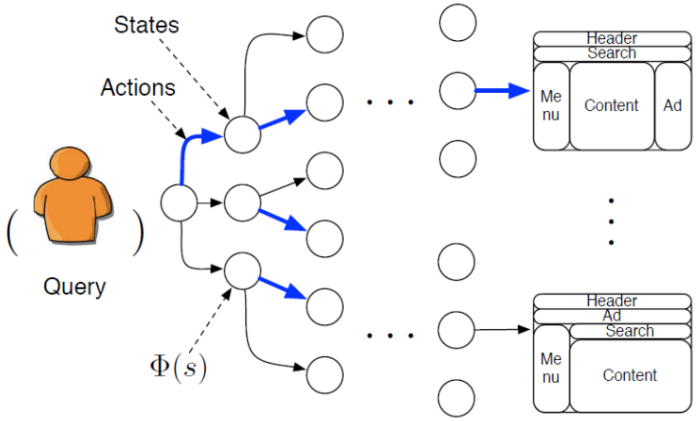
Outline



Prediction accuracy



Training/test/dev speed



Learning signals

Query

activity	**cooking**
agent	**woman**
food	**vegetable**

Fairness (data biases)

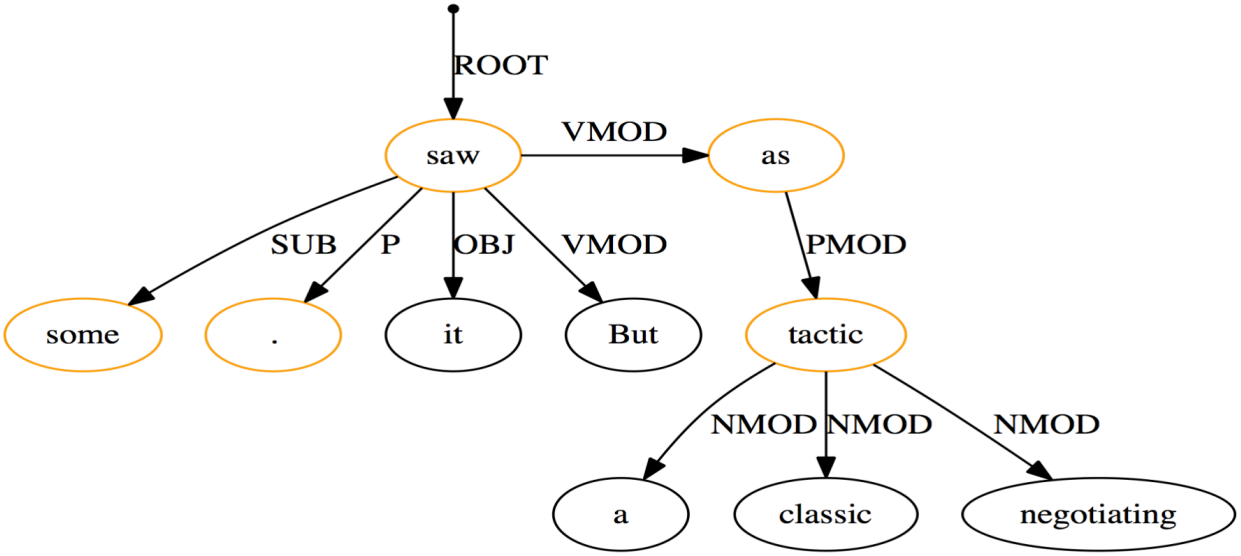
Solution Methods

- Assume a graphical structure; optimize
  - Three ideas for improving learning/inference speed
  - See our AAAI16 tutorial (https://goo.gl/TF7cGj)

- Learning to search approaches
  - A programmable framework
  - See our NAACL15 tutorials (http://hunch.net/~l2s)

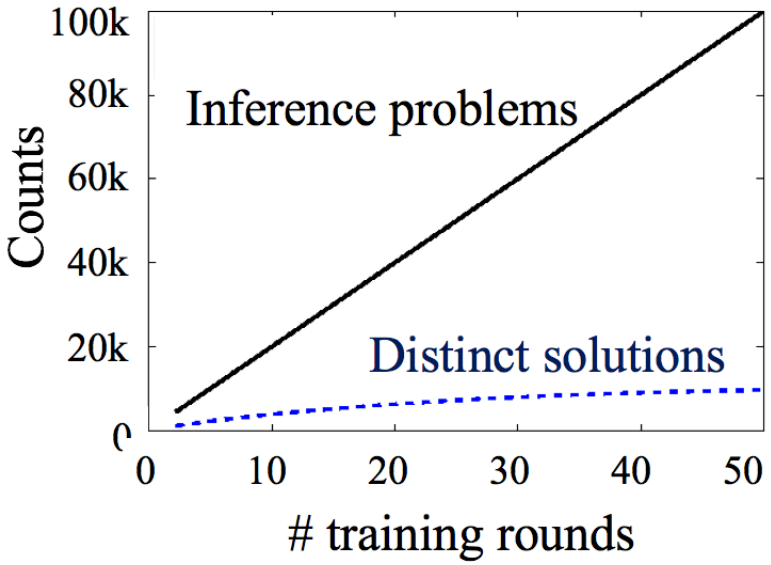
Graphical model approach: Speed up Inference/Learning

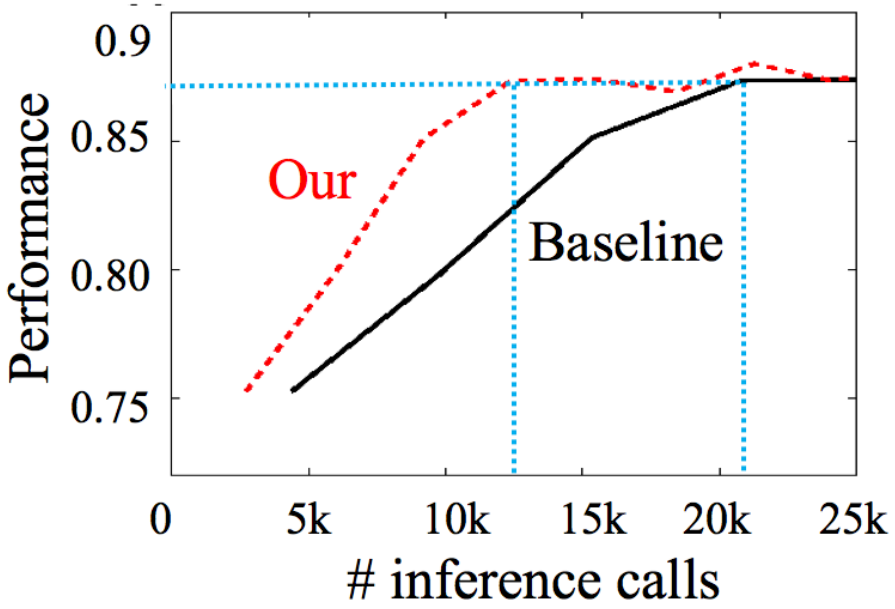
- Observation 1: some decisions are simpler than the others
- Idea: adaptively generate computationally costly features during test-time [AAAI 17]



Graphical model approach: Speed up Inference/Learning

- Observation 2: Many inference problems share the same solution
- Idea: Exploit this redundancy by caching old inference solutions [AAAI 15]





Amortized inference – key components

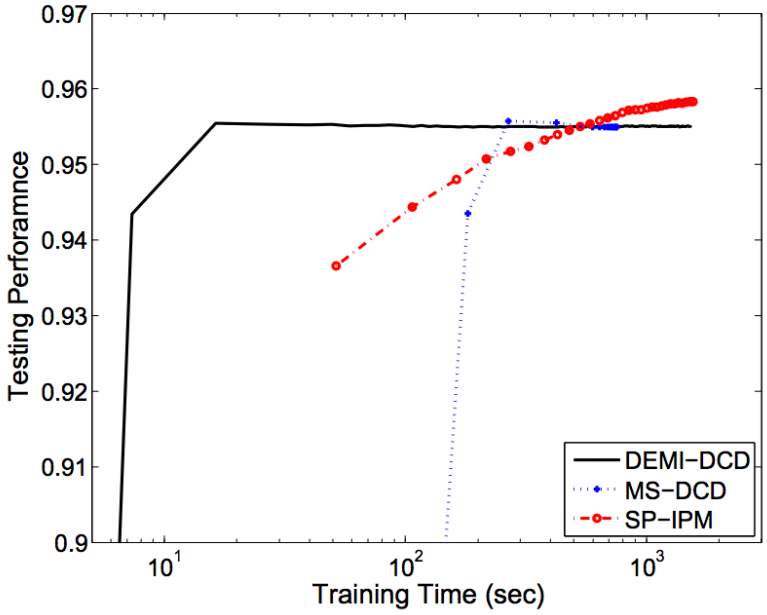
- Formulating the inference as an Integer Linear Programming

$$\arg\max_{\boldsymbol{y}\in\{0,1\}^n} \quad \sum_c S_c y_c \qquad s.t \quad A\boldsymbol{y} \le b$$

  - A very general formulation [Roth & Yih 04, Sontag 10]
  - Inference can be solved by any (exact or approximate) method
- A condition is being checked to determine if a new inference problem has the same solution as a previously observed problem. [Srikumar+ 12; Kundu+ 13]

Graphical model approach: Speed up Inference/Learning

- Observation 3: Inference can be solved in parallel
- Idea: Decouple inference and learning in the dual space
- Works both in the multi-thread [ECML13] and the multi-machines [NIPS OPT 15, journal in preparation] settings

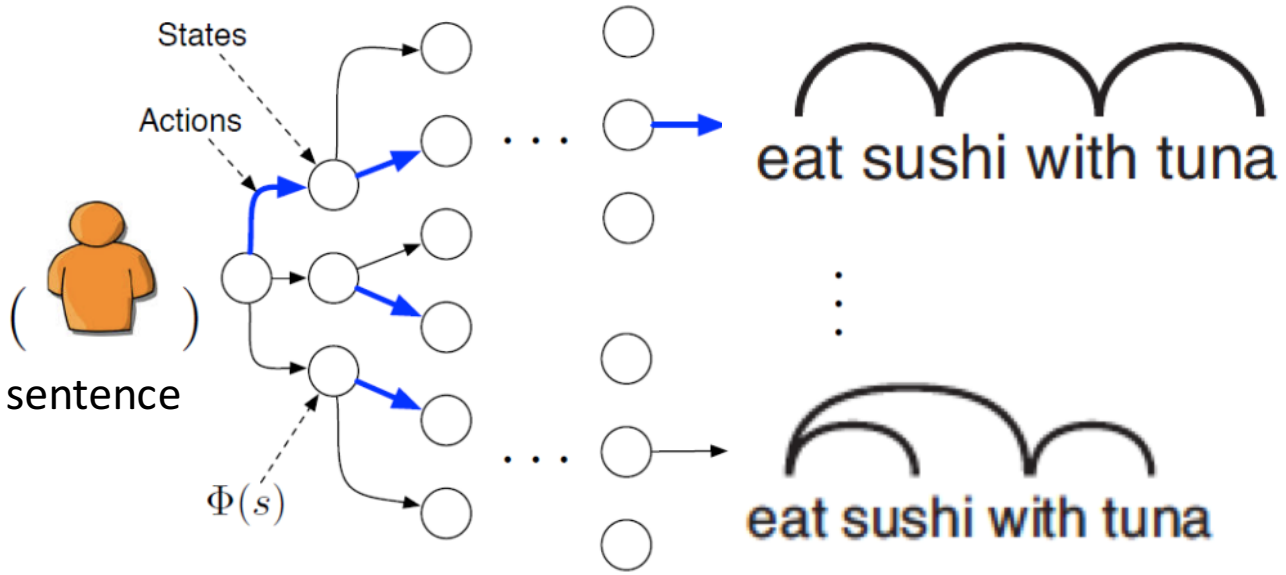


Learning to search (L2S) approaches

1. Define a search space and features
2. Construct a reference policy (Ref) based on the gold label
3. Learning a policy that imitates Ref

Credit Assignment Problem

When making a mistake, which local decision should be blamed?



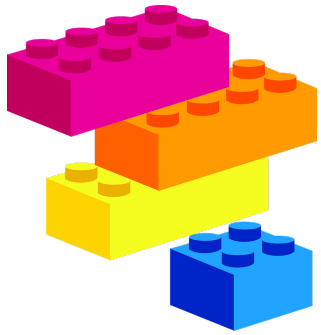
Existing L2S algorithms give

$$R(\pi) \leq R(\pi^{\text{ref}}) + o(1).$$

Learning to search approaches: Credit Assignment Compiler [NIPS16]

Sequential_RUN(*examples*)

```
1: for i = 1 to len(examples) do
2:     prediction ← predict(examples[i], examples[i].label)
3:     loss(prediction ≠ examples[i].label)
4: end for
```



- Write the decoder, providing some side information for training
- Library functions:
  - **predict**: returns individual predictions.
  - **loss**: declares the joint loss.
- An analogy to Factorie [McCallum+09]

Credit Assignment Compiler [NIPS 16]

Sequential_RUN(*examples*)

```
1: for i = 1 to len(examples) do
2:     prediction ← predict(examples[i], examples[i].label)
3:     loss(prediction ≠ examples[i].label)
4: end for
```

- Runs **Run()** many times to learn **predict()** that yields low **loss().**

  ⇨ turns **Run()** and training data into model updates

- Reduce a joint prediction problem to (cost-sensitive) multi-class problems.

Libraries for Structured Predictions

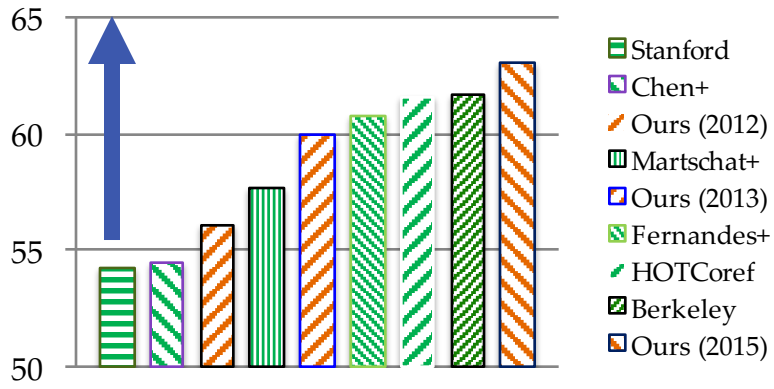
- Illinois-SL: graph-based structured prediction
  - Support various algorithms; parallel ⇒ very fast
- Vowpal-Wabbit: credit assignment compiler
  - A general online learning library
  - Support search-based structured prediction

Provide a nice platform

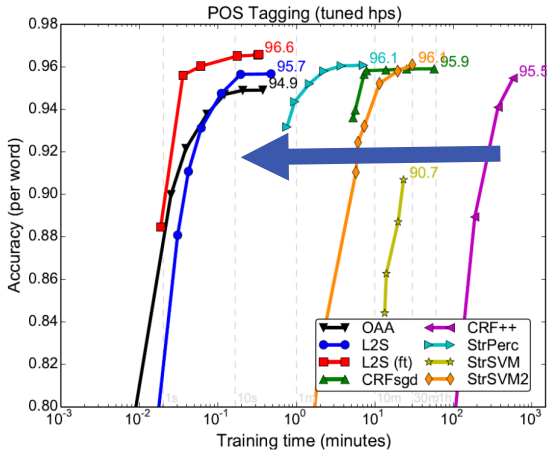
- for developing novel methods
- for collaboration
- for education

More easy-access tools; More collaborations

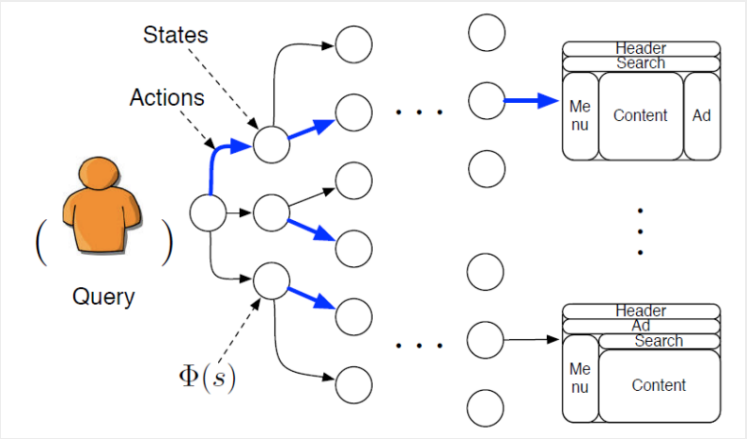
Outline



Prediction accuracy



Training/test speed



Learning signals

Query

activity	**cooking**
agent	**woman**
food	**vegetable**

Fairness (data biases)

Weak Supervision Challenges
[CRII grant]

- Implicit Supervision
  - Loss is not decomposable and can be estimated only when the entire output structure is derived

- Structured Contextual Bandit
  - Only a few (single) structured labels can be observed.

Implicit Supervision

- Consider algebra word problem

> Maria is now four times as old as Kate. Four years ago, Maria was six times as old as Kate. Find their ages now.

- Build semantic parser to translate question to an equation system

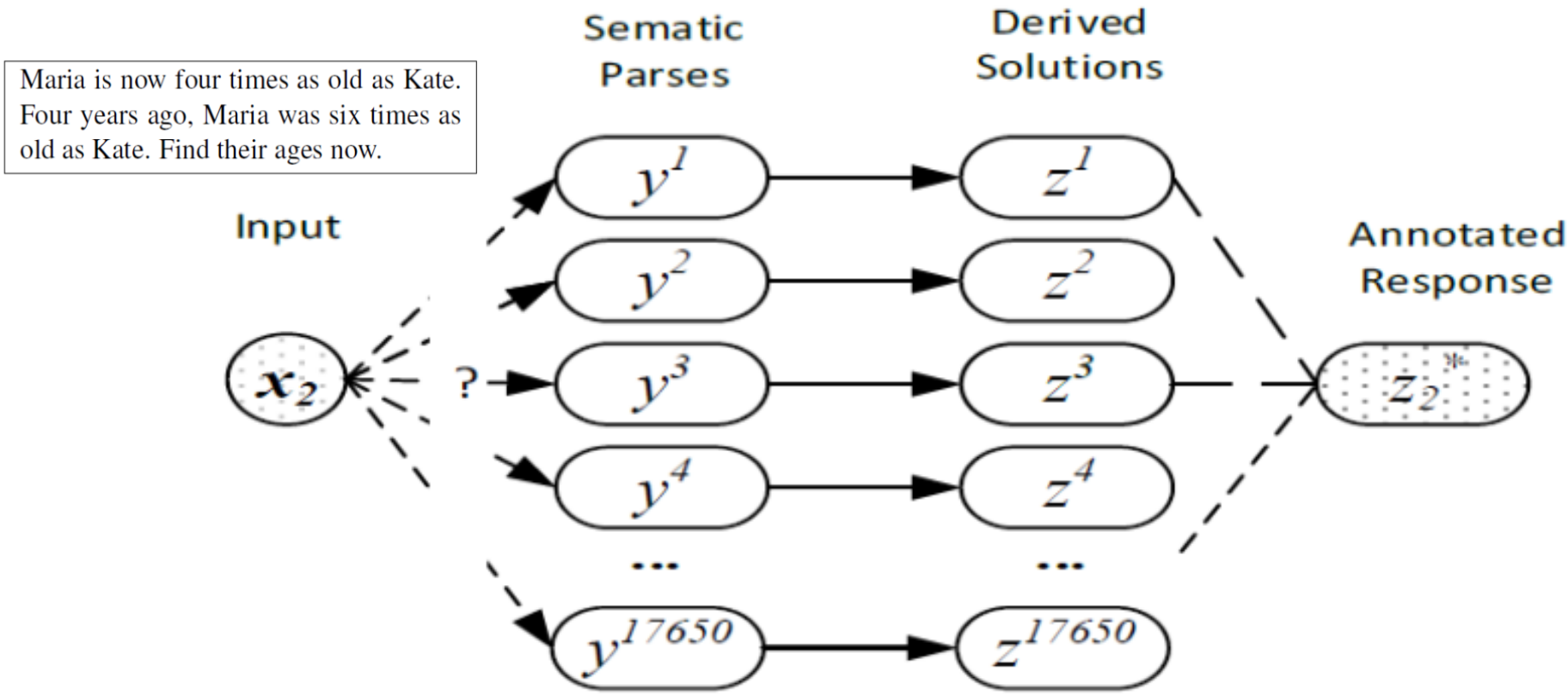
$$m = 4 \times n \text{ and } m - 4 = 6 \times (n - 4).$$

- Then answer can be derived: m=40, n=10

Implicit Supervision [EMNLP 16]

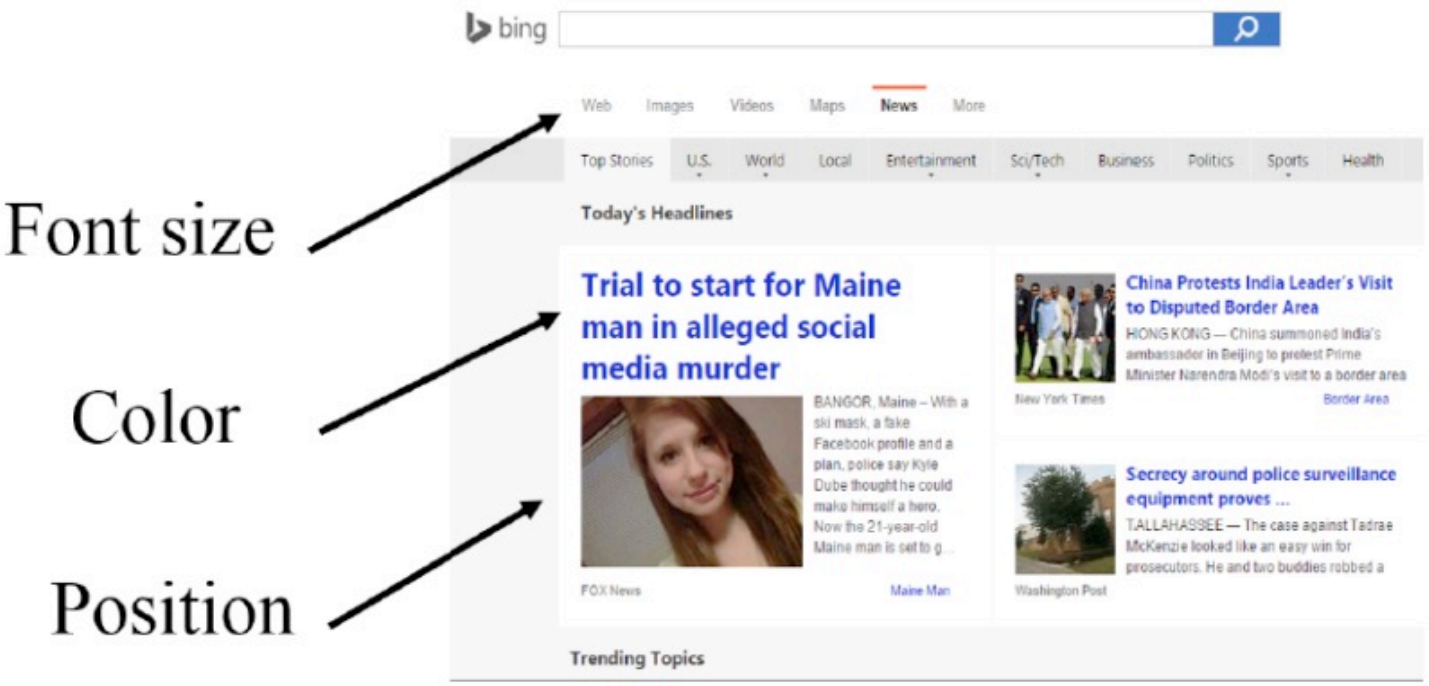
$$m = 4 \times n \text{ and}$$
$$m - 4 = 6 \times (n - 4).$$

m=40, n=10

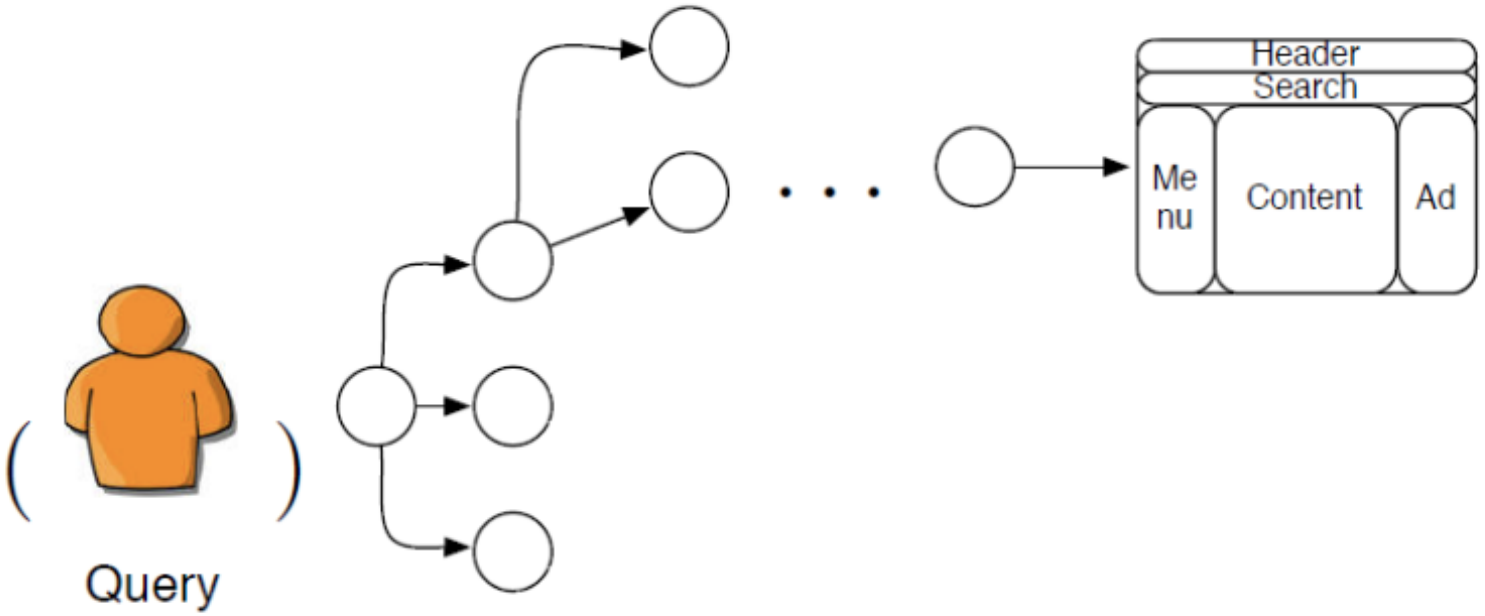


Structured Contextual Bandit Setting [ICML15]

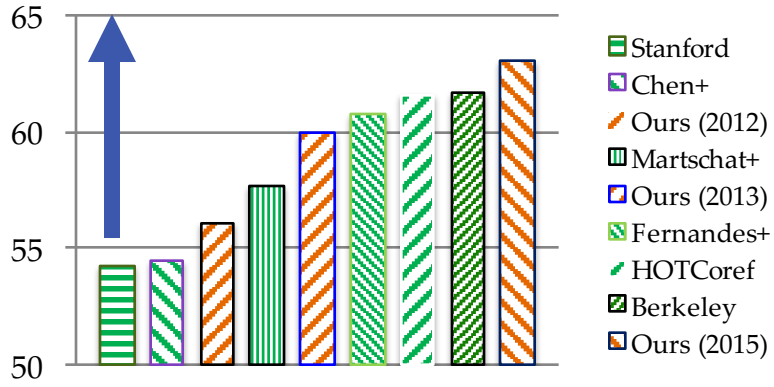
- Loss of only a single structured label can be observed



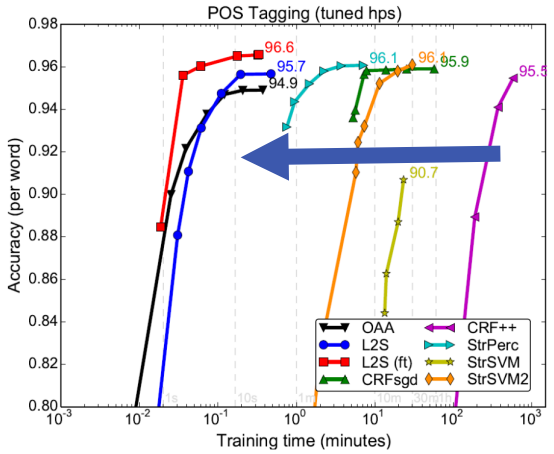
A Search Problem



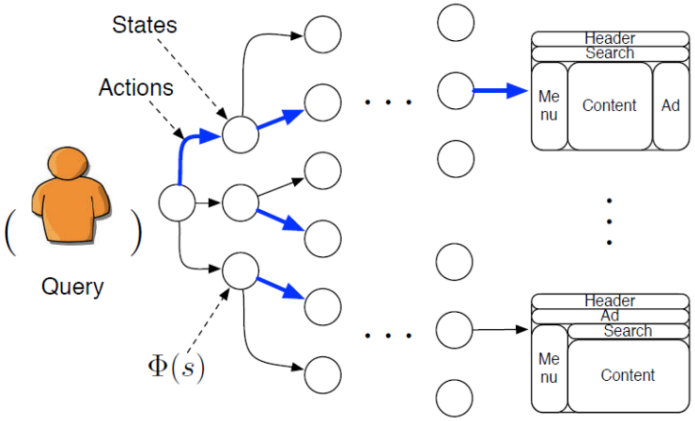
Outline



Prediction accuracy



Training/test speed



Learning signals

activity	cooking
agent	woman
food	vegetable

Fairness (data biases)

Human Bias in Structured model
[in submission]

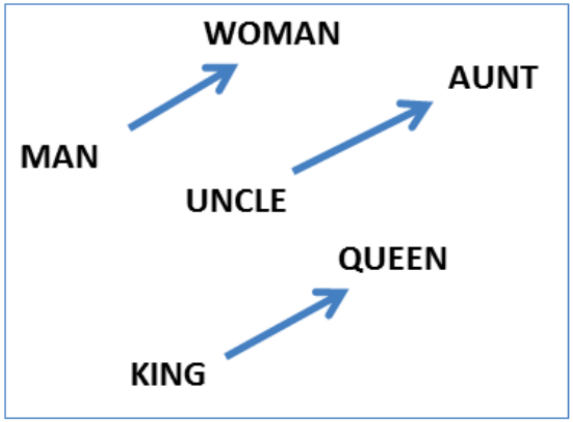
- A visual semantic role labeling system
  [Mark+16]

Query

activity	cooking
agent	woman
food	vegetable
container	bowl
tool	knife
place	kitchen

Word Embeddings can be Dreadfully Sexist [nips16]

- $v_{man} - v_{woman} + v_{uncle} \sim v_{aunt}$



he: ___	she: __
uncle	aunt
lion	
surgeon	
architect	
beer	
professor	

Debiasing Learning Models

- Idea1: Remove problematic correlation
  - E.g., remove gender bias subspace in WE
- Idea2: Set corpus-wise constraints to calibrate the gender ratios
  - Technique: Inference can be done by Lagrange relaxation

Structured Prediction – an active direction

- Landscape of methods in Deep∩Structure
  - Deep learning/hidden representation
    e.g., seq2seq, RNN, SP-energy network
  - Deep features, traditional factor graph inference
    e.g., LSTM+CRF, graph transformer networks,
- What is the right way to encode structures?
  - How to constrain the output
  - How can we leverage different learning signals?

Conclusions

Goal: Practical Structured Prediction Approaches

Tutorials/Workshops:

1. AAAI-16: Learning and Inference in SP Models
2. NAACL15: Hands-on Learning to Search for SP
3. EMNLP 16, 17: workshop SP for NLP

References/Code/Demos:

http://kwchang.net

Illinois-SL: a structured learning package

Vowpal Wabbit: an online learning library