COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

MULTIMODAL INFORMATION ACCESS & SYNTHESIS mias

Learning
from
Natural Instructions

Dan Roth

Department of Computer Science

University of Illinois at Urbana-Champaign

With thanks to:

Collaborators: **Ming-Wei Chang, Michael Connor, Dan Goldwasser, Vivek Srikumar,**

Funding:

NSF; DHS; NIH; DARPA.

DASH Optimization (Xpress-MP)

Comprehension

A process that maintains and updates a collection of propositions about the state of affairs.

(ENGLAND, June, 1989) - Christopher Robin is alive and well. He lives in England. 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. He made up a fairy tale land where Chris lived. His friends were animals. There was a bear called Winnie the Pooh. There was also an owl and a young pig, called a piglet. All the animals were stuffed toys that Chris owned. Mr. Robin made them come to life with his words. The places in the story were all near Cotchfield Farm. Winnie the Pooh was written in 1925. Children still love to read about Christopher Robin and his animal friends. Most people don't know he is a real person who is grown now. He has written two books of his own. They tell what it is like to be famous.

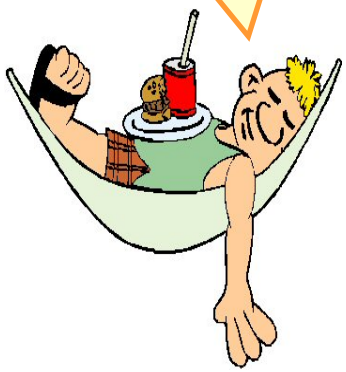
1. Christopher Robin was born in England.
2. Winnie the Pooh is a title of a book.
3. Christopher Robin's dad was a magician.
4. Christopher Robin must be at least 65 now.

This is an Inference Problem

Connecting Language to the World

Can we rely on this interaction to provide supervision (and, eventually, recover meaning) ?

Can I get a coffee with sugar and no milk



Semantic Parser

MAKE(COFFEE,SUGAR=YES,MILK=NO)

Great!

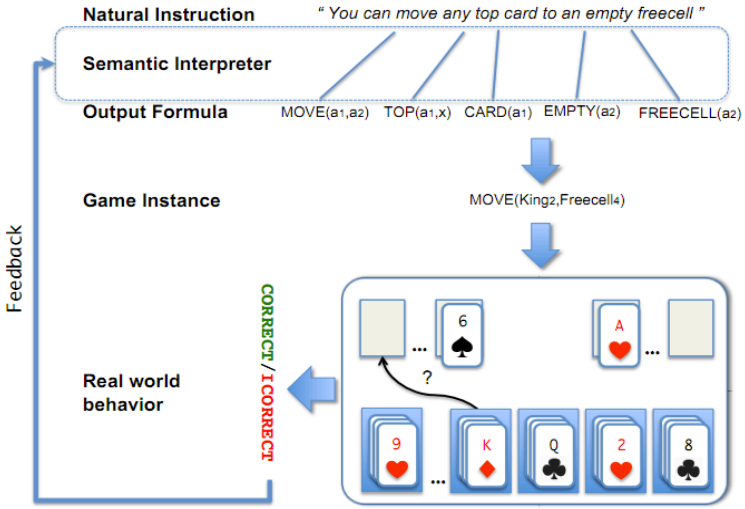
Arggg

- How to recover meaning from text?
- Annotate with meaning representation; use (standard) “example based” ML
  - **Teacher needs deep understanding of the learning agent**
  - **Annotation burden; not scalable.**
- Instructable computing
  - **Natural communication between teacher/agent**

Scenarios I: Understanding Instructions [IJCAI'11]

- Understanding Games' Instructions

A top card can be moved to the tableau if it has a different color than the color of the top tableau card, and the cards have successive values.



- Allow a teacher to interact with an automated learner using natural instructions
  - **Agonstic to agent's internal representations**
  - **Contrasts with traditional 'example-based' ML**

$move(a_1, a_2)$
$top(a_1, x_1)$ $card(a_1)$ $freecell(a_2)$ $empty(a_2)$

- "You can move any of the top cards to an empty free-cell"

$move(a_1, a_2)$
$top(a_1, x_1)$ $card(a_1)$ $tableau(a_2)$ $top(x_2, a_2)$
$color(a_1, x_3)$ $color(x_2, x_4)$ $not\text{-}equal(x_3, x_4)$
$value(a_1, x_5)$ $value(x_2, x_6)$ $successor(x_5, x_6)$

- "A top card can be moved to a tableau if it has a different color than the color of the top tableau card, and the cards have successive values"

What to Learn from Natural Instructions?

- Two conceptual ways to think about learning from instructions

- (i) Learn directly to play the game [EMNLP'09; Barziley et. al 10,11]
  - Consults the natural language instructions
  - Use them as a way to improve your feature based representation
- (ii) Learn to interpret a natural language lesson [IJCAI'11]
  - And (jointly) how to use this interpretation to do well on the final task.
  - Will this help generalizing to other games?

- Semantic Parsing into some logical representation is a necessary intermediate step
  - Learn how to semantically parse from task level feedback
  - Evaluate at the task level rather than the representation level

Scenario I': Semantic Parsing [CoNLL'10,ACL'11…]

X :"*What is the largest state that borders New York and Maryland ?*"

Y: largest(state(next_to(state(NY)) AND next_to (state(MD))))

- Successful interpretation involves multiple decisions
  - **What entities appear in the interpretation?**
  - **"New York" refers to a state or a city?**
  - **How to compose fragments together?**
    - **state(next_to()) >< next_to(state())**
- Question: How to learn to semantically parse from "task level" feedback.

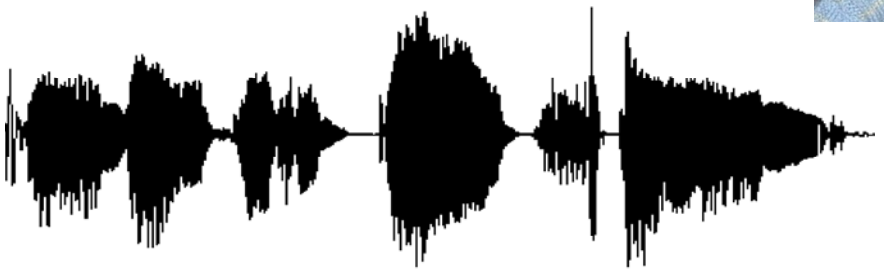
In all cases: There is an important intermediate representation between the input and the task level. We are interested in learning to do well on the task, without getting any feedback in the intermediate level.

g problem

“the world”

- How do we acquire language?

“the language”



[Topid rivvo den marplox.]

- Is it possible to learn the meaning of verbs from natural, behavior level, feedback? (no “intermediate representation” level feedback)

Outline

- Background: NL Structure with Integer Linear Programming
  - **Global Inference with expressive structural constraints in NLP**

- Constraints Driven Learning with Indirect Supervision
  - **Training Paradigms for latent structure**
  - **Indirect Supervision Training with latent structure (NAACL'10)**
  - **Training Structure Predictors by Inventing binary labels (ICML'10)**

- Response based Learning
  - **Driving supervision signal from World's Response (CoNLL'10,IJCAI'11)**
  - **Semantic Parsing ; Playing Freecell; Language Acquisition**
  - **Some work in progress**

Interpret Language Into An Executable Representation

X :"*What is the largest state that borders New York and Maryland ?*"

Y: largest(state(next_to(state(NY) AND next_to (state(MD))))

- Successful interpretation involves multiple decisions
  - **What entities appear in the interpretation?**
  - **"New York" refers to a state or a city?**
  - **How to compose fragments together?**
    - **state(next_to()) >< next_to(state())**

- Question: How to learn to semantically parse from "task level" feedback.

Learning and Inference in NLP

- Natural Language Decisions are Structured
  - Global decisions in which several local decisions play a role but there are mutual dependencies on their outcome.
- It is essential to make coherent decisions in a way that takes the interdependencies into account. Joint, Global Inference.
- But: Learning structured models requires annotating structures.
- Interdependencies among decision variables should be exploited in Decision Making (Inference) and in Learning.
  - Goal: learn from minimal, indirect supervision
  - Amplify it using interdependencies among variables

CCMs can be viewed as a general interface to easily combine declarative domain knowledge with data driven statistical models

Const

Penalty for violating the constraint.

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y) - \sum_{i=1}^{K} \rho_i d(y, 1_{C_i(x)})$$

(Soft) constraints component

Weight Vector for "local" models

Features, classifiers; log-linear models (HMM, CRF) or a combination

How far y is from a "legal" assignment

How to solve?

This is an Integer Linear Program

Solving using ILP packages gives an exact solution.

Cutting Planes, Dual Decomposition & other search techniques are possible

How to train?

Training is learning the objective function

Decouple? Decompose?

How to exploit the structure to minimize supervision?

Three Ideas

- **Idea 1:** Modeling

  Separate modeling and problem formulation from algorithms
  - Similar to the philosophy of probabilistic modeling

- **Idea 2:** Inference

  Keep model simple, make expressive decisions (via constraints)
  - Unlike probabilistic modeling, where models become more expressive

- **Idea 3:** Learning

  Expressive structured decisions can be supervised indirectly via related simple binary decisions
  - Global Inference can be used to amplify the minimal supervision.

Constrained Conditional Models

- Constrained Conditional Models – ILP formulations – have been shown useful in the context of many NLP problems

- [Roth&Yih, 04,07: Entities and Relations; Punyakanok et. al: SRL …]
  - Summarization; Co-reference; Information & Relation Extraction; Event Identifications; Transliteration; Textual Entailment; Knowledge Acquisition; Sentiments; Temporal Reasoning, Dependency Parsing,…

- Some theoretical work on training paradigms [Punyakanok et. al., 05 more; Constraints Driven Learning, PR, Constrained EM…]

- **NAACL'12 Tutorial: http://L2R.cs.uiuc.edu/**

Outline

- Background: NL Structure with Integer Linear Programming
  - **Global Inference with expressive structural constraints in NLP**

- Constraints Driven Learning with Indirect Supervision
  - **Training Paradigms for latent structure**
  - **Indirect Supervision Training with latent structure (NAACL'10)**
  - **Training Structure Predictors by Inventing binary labels (ICML'10)**

- Response based Learning
  - **Driving supervision signal from World's Response (CoNLL'10,IJCAI'11)**
  - **Semantic Parsing ; playing Freecell; Language Acquisition**

Semantic Parsing as Structured Prediction

X :"*What is the largest state that borders New York and Maryland ?*"

Y: largest(state(next_to(state(NY) AND next_to (state(MD))))

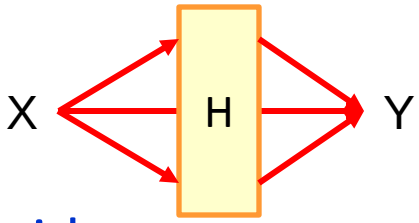
- Successful interpretation involves multiple decisions
  - What entities appear in the interpretation?
  - "New York" refers to a state or a city?
  - How to compose fragments together?
    - **state(next_to()) >< next_to(state())**

1. Learning latent structure from minimal supervision
2. Invent/solicit supervision to learn structure
3. Use inference to solicit task specific feedback
   - Learn structure

- Question: How to learn to semantically parse from "task level" feedback.

I. Paraphrase Identification

Given an input $x \in X$
Learn a model $f : X \rightarrow \{-1, 1\}$



- Consider the following sentences:
- S1: Druce will face murder charges, Conte said.
- S2: Conte said Druce will be charged with murder .

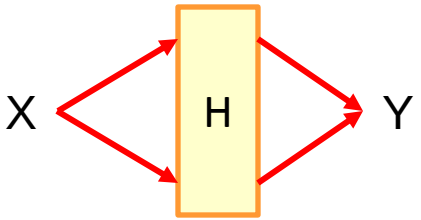
We need latent variables that explain why this is a positive example.

- Are S1 and S2 a paraphrase of each other?
- There is a need for an intermediate representation to justify this decision

Given an input $x \in X$
Learn a model $f : X \rightarrow H \rightarrow \{-1, 1\}$

Algorithms: Two Conceptual Approaches

- Two stage approach (a pipeline; typically used for TE, paraphrase id, others)
  - Learn hidden variables; fix it
    - **Need supervision for the hidden layer (or heuristics)**
  - For each example, extract features over x and (the fixed) h.
  - Learn a binary classier for the target task



- Proposed Approach: Joint Learning
  - Drive the learning of h from the binary labels
  - Find the best h(x)
  - **An intermediate structure representation is good to the extent it supports better final prediction.**
  - Algorithm? How to drive learning a good H?

Learning with Constrained Latent Representation (LCLR): Intuition

- If x is positive
  - There must exist a good explanation (intermediate representation)
  - $\exists$ h, $w^T \phi(x,h) \geq 0$
  - or, $\max_h w^T \phi(x,h) \geq 0$

This is an inference step that will gain from the CCM formulation CCM on the latent structure

- If x is negative
  - No explanation is good enough to support the answer
  - $\forall$ h, $w^T \phi(x,h) \leq 0$
  - or, $\max_h w^T \phi(x,h) \leq 0$

New feature vector for the final decision. Chosen h selects a representation.

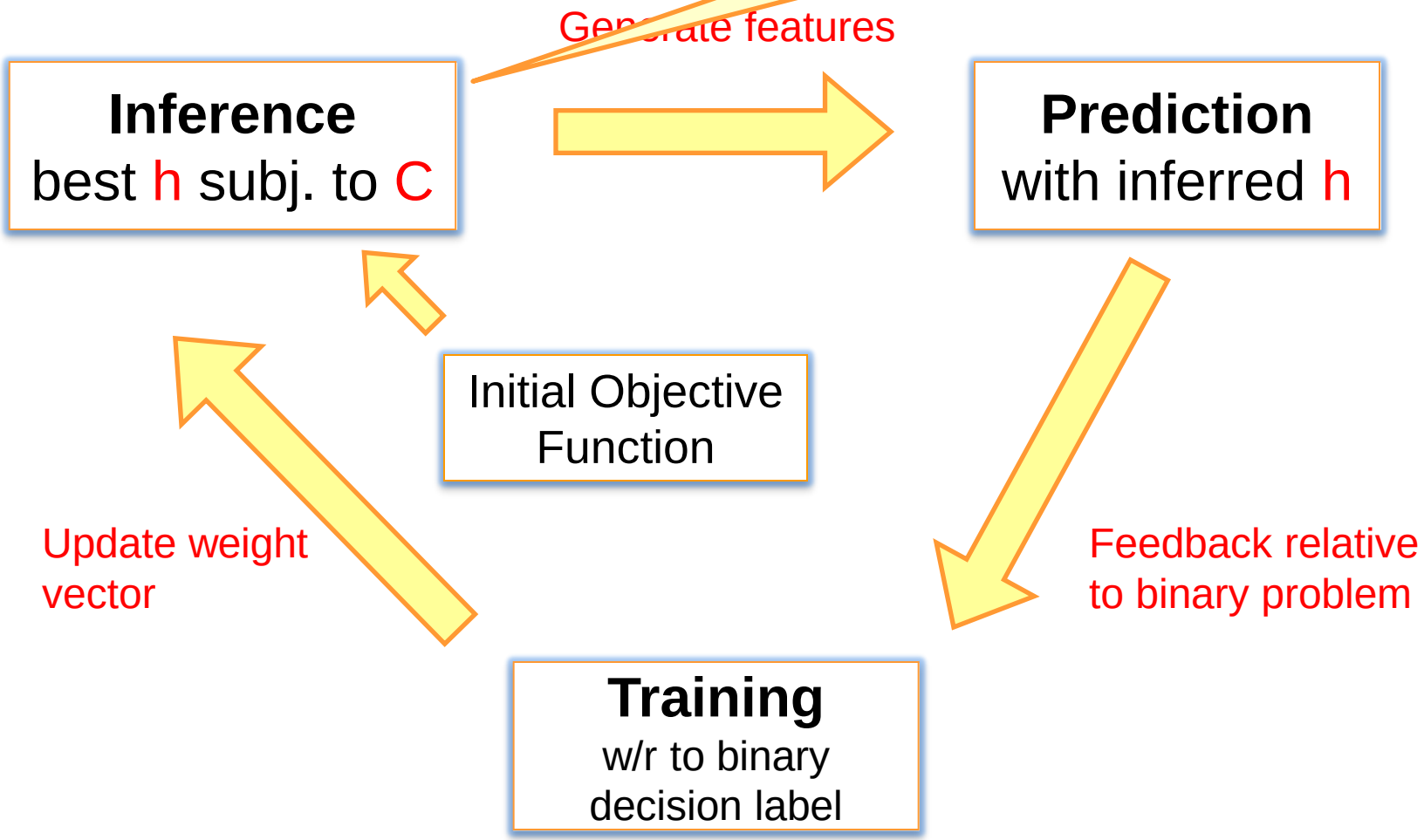
- Altogether, this can be combined into an objective function:

$$\text{Min}_w \; 1/2||w||^2 \; + \; C\sum_i L(1\text{-}z_i \max_{h \in \mathcal{C}} w^T \sum_{\{s\}} h_s \phi_s(x_i))$$

Inference: best h subject to constraints C

Iterative Objective Function Learning

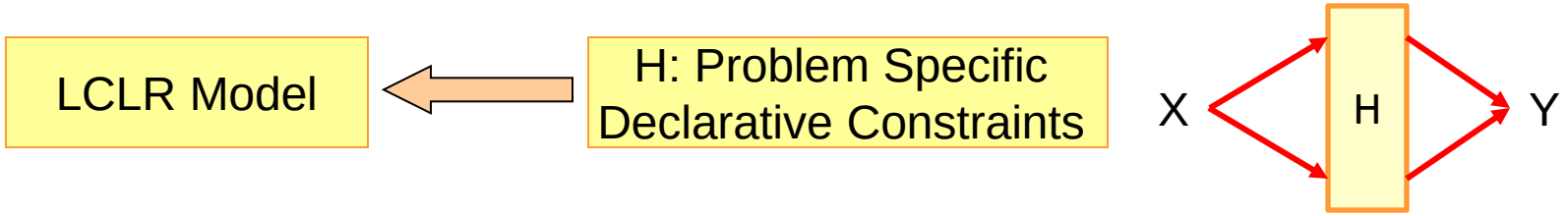
ILP inference discussed earlier; restrict possible hidden structures considered.



- Formalized as Structured SVM + Constrained Hidden Structure
- **LCRL: Learning Constrained Latent Representation**

Learning with Constrained Latent Representation (LCLR): Framework

- LCLR provides a general inference formulation that allows the use of expressive constraints to determine the hidden level
  - Flexibly adapted for many tasks that require latent representations.



- Paraphrasing: Model input as graphs, $V(G_{1,2})$, $E(G_{1,2})$
  - Four (types of) Hidden variables:
    - **$h_{v1,v2}$ – possible vertex mappings; $h_{e1,e2}$ – possible edge mappings**

$$\forall v_1 \in V(G_1), \sum_{v_2 \in V(G_2)} h_{v_1,v_2} + h_{v_1,*} = 1, \quad \forall v_2 \in V(G_2), \sum_{v_1 \in V(G_1)} h_{v_1,v_2} + h_{*,v_2} = 1$$

$$\forall e_1 \in E(G_1), \sum_{e_2 \in E(G_2)} h_{e_1,e_2} + h_{e_1,*} = 1, \quad \forall e_2 \in E(G_2), \sum_{e_1 \in E(G_1)} h_{e_1,e_2} + h_{*,e_2} = 1$$

$$h_{v_1,v_2} + h_{v_1',v_2'} - h_{e_1,e_2} \leq 1, \quad h_{v_1,v_2} \geq h_{e_1,e_2}, \quad h_{v_1',v_2'} \geq h_{e_1,e_2}$$

Experimental Results

Transliteration:

Transliteration System	Acc	MRR
(Goldwasser and Roth 2008)	N/A	89.4
Alignment + Learning	80.0	85.7
LCLR	**92.3**	**95.4**

Recognizing Textual Entailment:

Entailment System	Acc
Median of TAC 2009 systems	61.5
Alignment + Learning	65.0
LCLR	**66.8**

Paraphrase Identification:*

Alignment + Learning	72.00
LCLR	**72.75**

Outline

- Background: NL Structure with Integer Linear Programming
  - **Global Inference with expressive structural constraints in NLP**

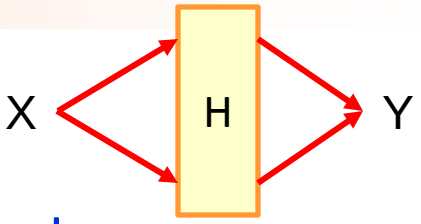
- Constraints Driven Learning with Indirect Supervision
  - **Training Paradigms for latent structure**
  - **Indirect Supervision Training with latent structure (NAACL'10)**
  - **Training Structure Predictors by Inventing binary labels (ICML'10)**

- Response based Learning
  - **Driving supervision signal from World's Response (CoNLL'10,IJCAI'11)**
  - **Semantic Parsing ; playing Freecell; Language Acquisition**

II: Structured Prediction

- Before, the structure was in the intermediate level
  - We cared about the structured representation only to the extent it helped the final binary decision
  - The binary decision variable was given as supervision
- What if we care about the structure?
  - Information & Relation Extraction; POS tagging, Semantic Parsing
- Invent a companion binary decision problem!

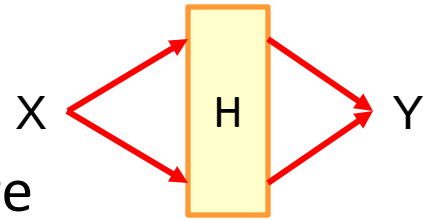
Structured Prediction



- Before, the structure was in the intermediate level
  - We cared about the structured representation only to the extent it helped the final binary decision
  - The binary decision variable was given as supervision
- What if we care about the structure?
  - Information Extraction; Relation Extraction; POS tagging, many others.
- Invent a companion binary decision problem!
  - **Parse Citations:** Lars Ole Andersen . Program analysis and specialization for the C Programming language. PhD thesis. DIKU , University of Copenhagen, May 1994 .
  - **Companion**: Given a citation; does it have a legitimate citation parse?
  - **POS Tagging**
  - **Companion:** Given a word sequence, does it have a legitimate POS tagging sequence?
- Binary Supervision is almost free

Companion Task Binary Label as Indirect Supervision

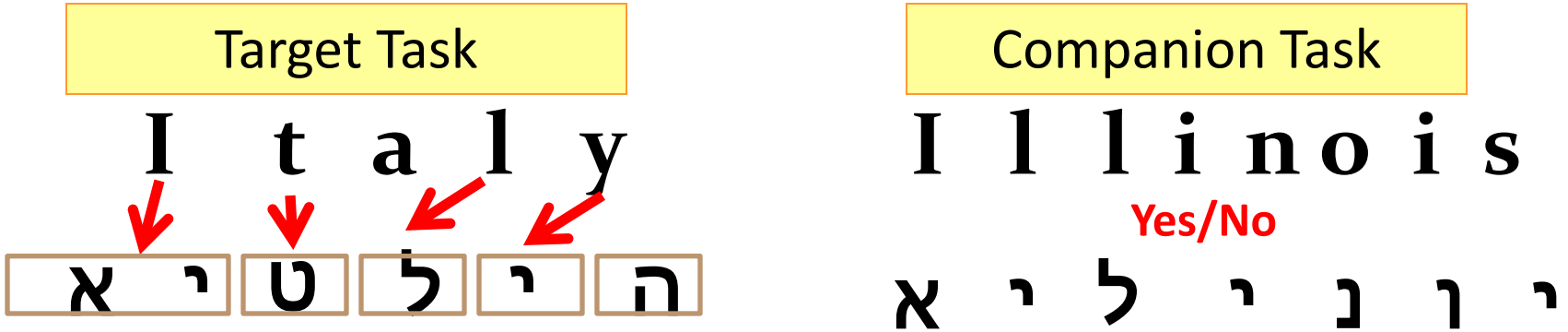
- The two tasks are related just like the binary and structured tasks discussed earlier



- All positive examples must have a good structure
- Negative examples cannot have a good structure
- We are in the same setting as before
  - Binary labeled examples are easier to obtain
  - We can take advantage of this to help learning a structured model
- Algorithm: (1) Use previous algorithm
  (2) Augment it with structured supervision

Joint Learning Framework

- Joint learning : If available, make use of both supervision types



Loss function – same as described earlier.

Key: the same parameter **w** for both components

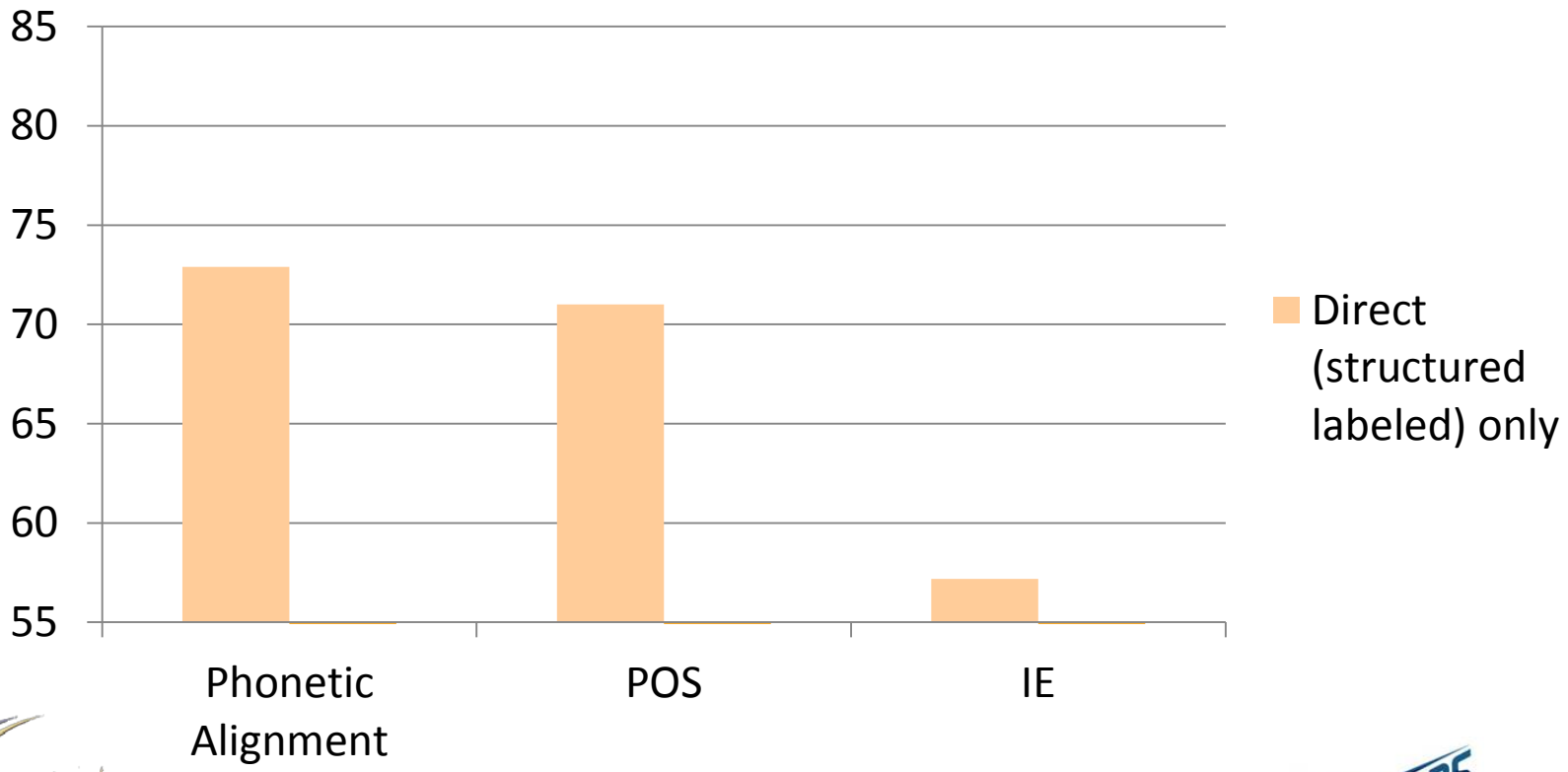
$$\min_{w} \frac{1}{2} w^T w + C_1 \sum_{i \in S} L_S(x_i, y_i; w)$$

Loss on Target Task

Loss on Companion Task

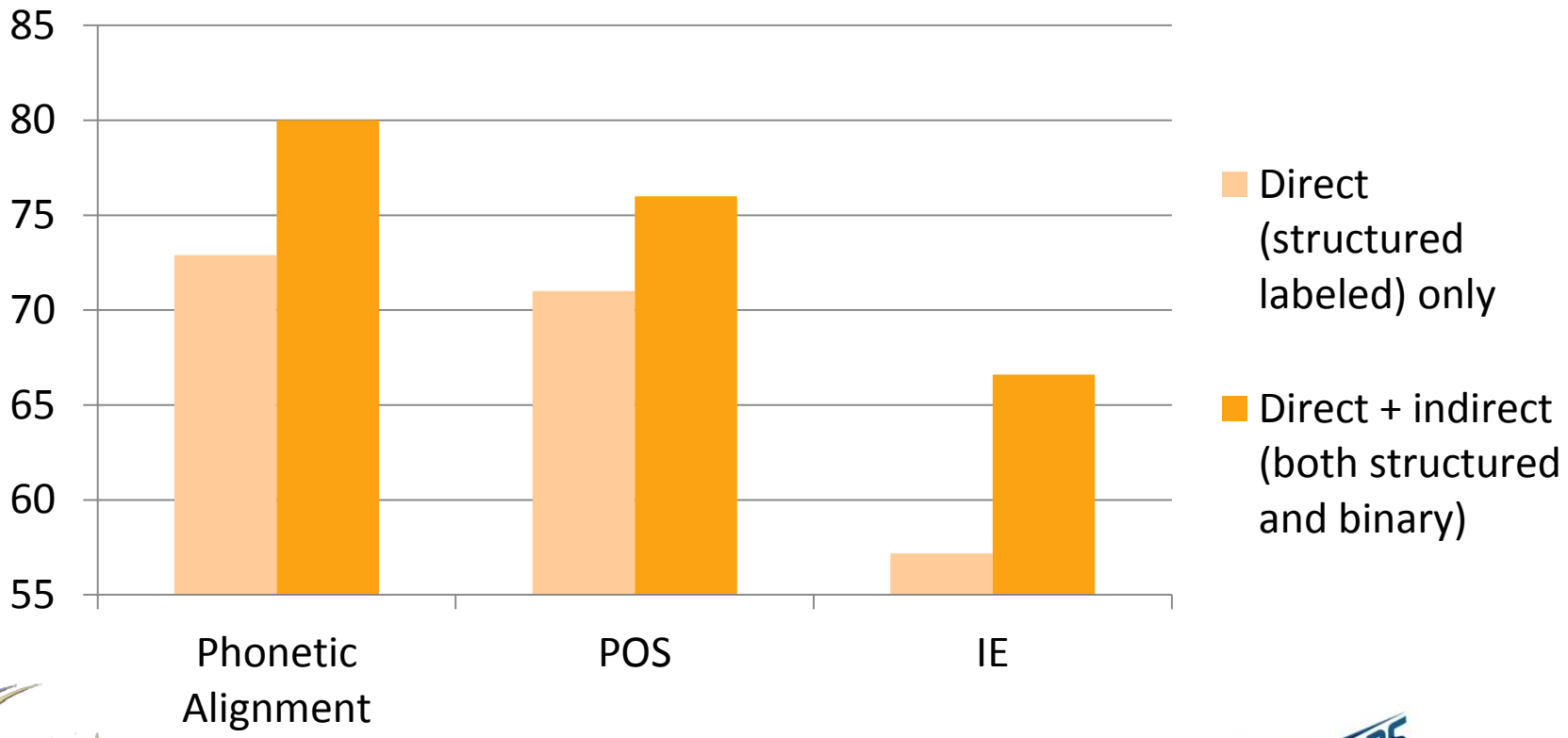
Experimental Result

- Very little direct (structured) supervision.



Experimental Result

- Very little direct (structured) supervision.
- (Almost free) Large amount binary indirect supervision



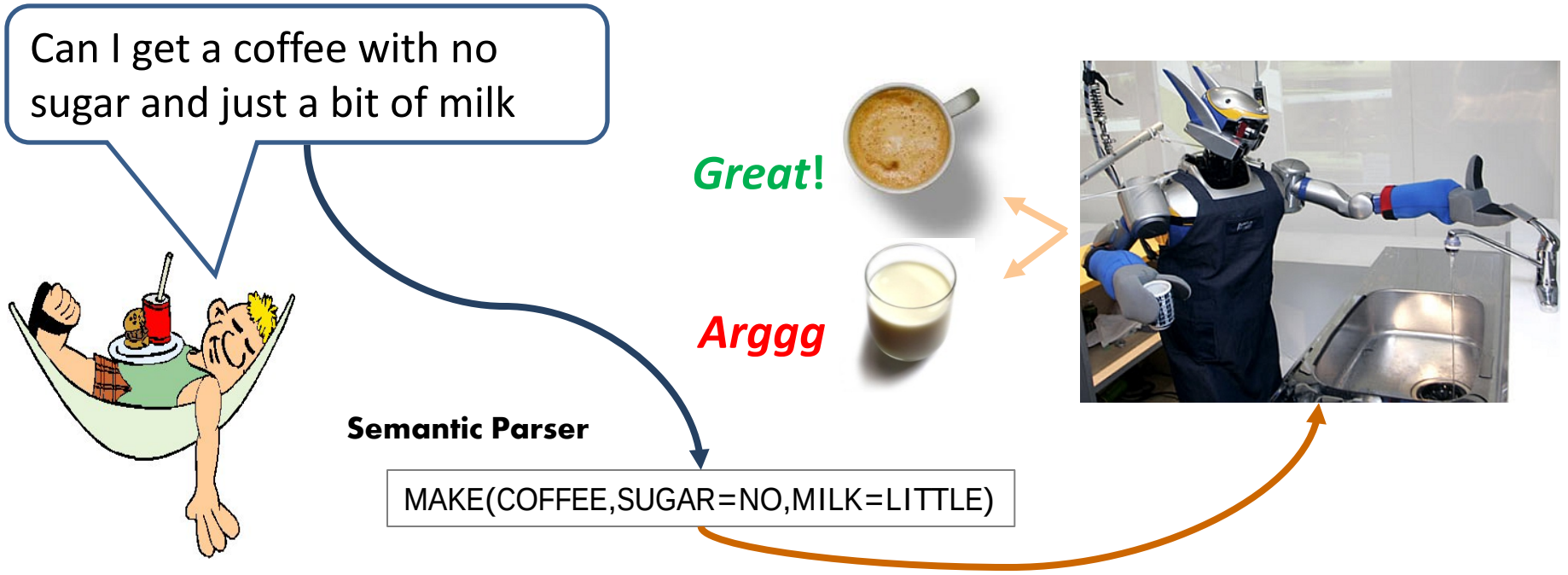
Outline

- Background: NL Structure with Integer Linear Programming
  - **Global Inference with expressive structural constraints in NLP**

- Constraints Driven Learning with Indirect Supervision
  - **Training Paradigms for latent structure**
  - **Indirect Supervision Training with latent structure (NAACL'10)**
  - **Training Structure Predictors by Inventing binary labels (ICML'10)**

- Response based Learning
  - **Driving supervision signal from World's Response (CoNLL'10,IJCAI'11)**
  - **Semantic Parsing ; playing Freecell; Language Acquisition**

Connecting Language to the World [CoNLL'10,ACL'11,IJCAI'11]



Can we rely on this interaction to provide supervision?

Real World Feedback

Supervision = Expected Response

Traditional approach: learn from logical forms and gold alignments

EXPENSIVE!

Our approach: use only the responses

x NL Query

y Logical Query

r Query Response:

"What is the largest state that borders NY?"

largest(state(next_to(const(NY))))

Pennsylvania

Pennsylvania

Interactive Computer System

Binary Supervision

Check if Predicted response == Expected response

Semantic parsing is a structured prediction problem: identify mappings from text to a meaning representation

Expected : Pennsylvania
Predicted : Pennsylvania
Positive Response

Expected : Pennsylvania
Predicted : NYC
Negative Response

Train a structured predictor with this binary supervision !

Response Based Learning

Input: Inputs $\{\mathbf{x}^l\}_{l=1}^N$,
$Feedback : \mathcal{X} \times \mathcal{Y} \rightarrow \{+1, 1\}$,
initial weight vector $\mathbf{w}$

1: **repeat**
2: **for** $l = 1, \ldots, N$ **do**
3: $\hat{\mathbf{h}}, \hat{\mathbf{y}} = \arg\max_{\mathbf{h},\mathbf{y}} \mathbf{w}^T \Phi(\mathbf{x}^l, \mathbf{h}, \mathbf{y})$
4: $f = Feedback(\mathbf{x}^l, \hat{\mathbf{y}})$
5: add $(\Phi(\mathbf{x}^l, \hat{\mathbf{h}}, \hat{\mathbf{y}}), f)$ to B
6: **end for**
7: $\mathbf{w} \leftarrow TRAIN(B)$
8: **until** Convergence
9: **return** $\mathbf{w}$

Difficulties:
- **Need to generate training examples**
- **Negative examples give no information**

Basic Algorithm:
- Try to generate good structures
  - Inference w Constraints
- Get world's response
- Update parameters
  - Previous algorithms

TRAIN: Try to get more positive examples (representations with positive feedback)
Direct (Binary) protocol: a binary classifier on **Positive/Negative ex's**
(**Problem: many good sub-structures are being demoted**)
Structured Protocol: Use only correct structures.
(Problem: ignores negative feedback)

Constraints Drive Inference

Repeat
  for all input sentences **do**
    Find best structured output
    Query *feedback* function
  end for
  Learn new W using feedback
Until Convergence

- X: What is the largest state that borders NY?
- Y: largest(state(next_to(const(NY))))

$$y^* = F_w(x) = \arg\max_{y \in Y} score(x, y) = \arg\max_{y \in Y, h \in H} w^T \Phi(x, y, h)$$

So Far

And now…

- Decompose into two types of decisions:
  - **First order:** Map lexical items to logical symbols
    - {*"largest"* → `largest()`, *"borders"* → `next_to()`,.., *"NY"* → `const(NY)`}
  - **Second order:** Compose meaning from logical fragments
    - `largest(state(next_to(const(NY))))`
- Domain's semantics is used to constrain interpretations
  - declarative constraints: Lexical resources (wordnet); type consistency: distance in sentence, in dependency tree,…

Empirical Evaluation [CoNLL'10,ACL'11]

- Key Question: **Can we learn from this type of supervision?**

Algorithm	# training structures	Test set accuracy
No Learning: Initial Objective Fn	0	22.2%
Binary signal: **Binary Protocol**	0	69.2 %
Binary signal: **Structured Protocol**	0	**73.2 %**
Improved Protocol:	0	**79.6%**
Improved Protocol + Loss Fn	0	**81.6%**
WM*2007 (fully supervised – uses gold structures)	**310**	**75 %**

*[WM] Y.-W. Wong and R. Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. ACL.

Current emphasis: Learning to understand natural language instructions for games via response based learning

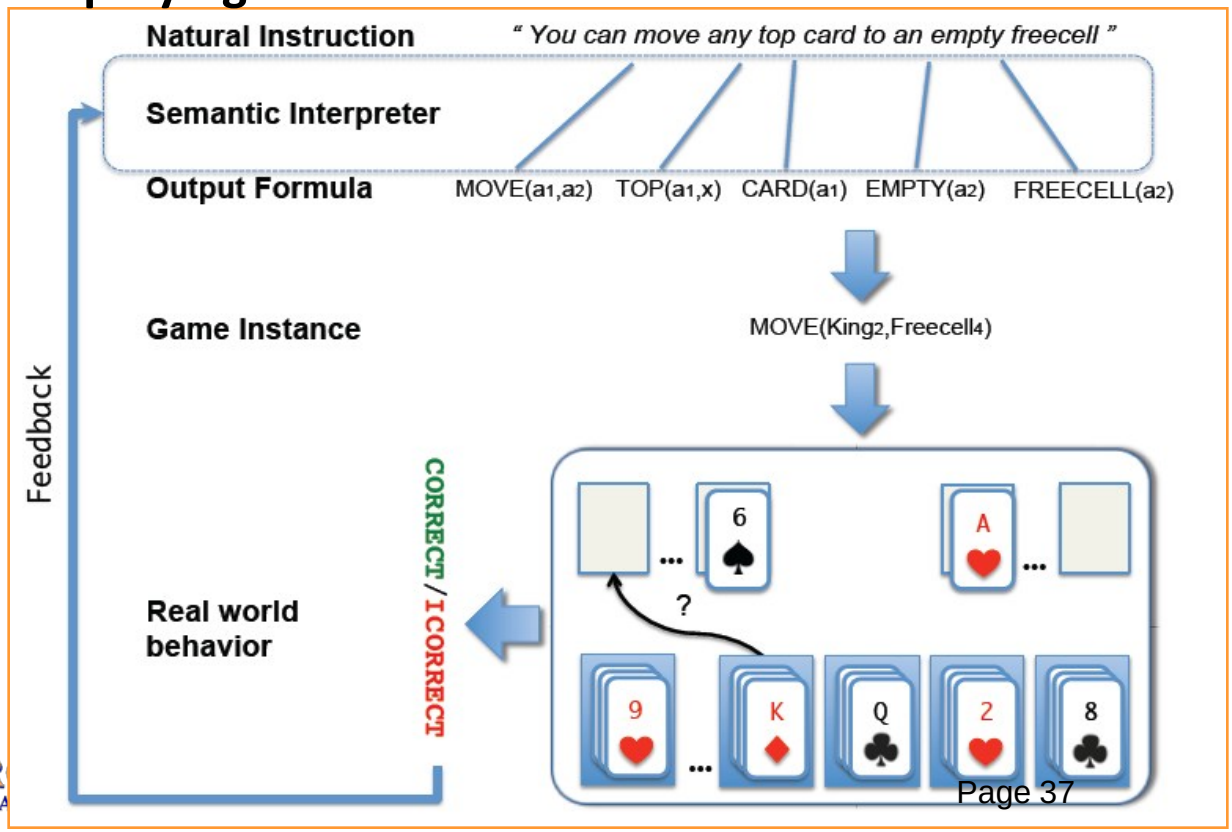
Learning from Natural Instructions

- A human teacher interacts with an automated learner using natural instructions
- Learner is given:
  - **A lesson describing the target concept directly**
  - **A few instances exemplifying it**

Challenges:

(1) how to interpret the lesson and

(2) **how to use this interpretation to do well on the final task**.



Lesson Interpretation as an inference problem

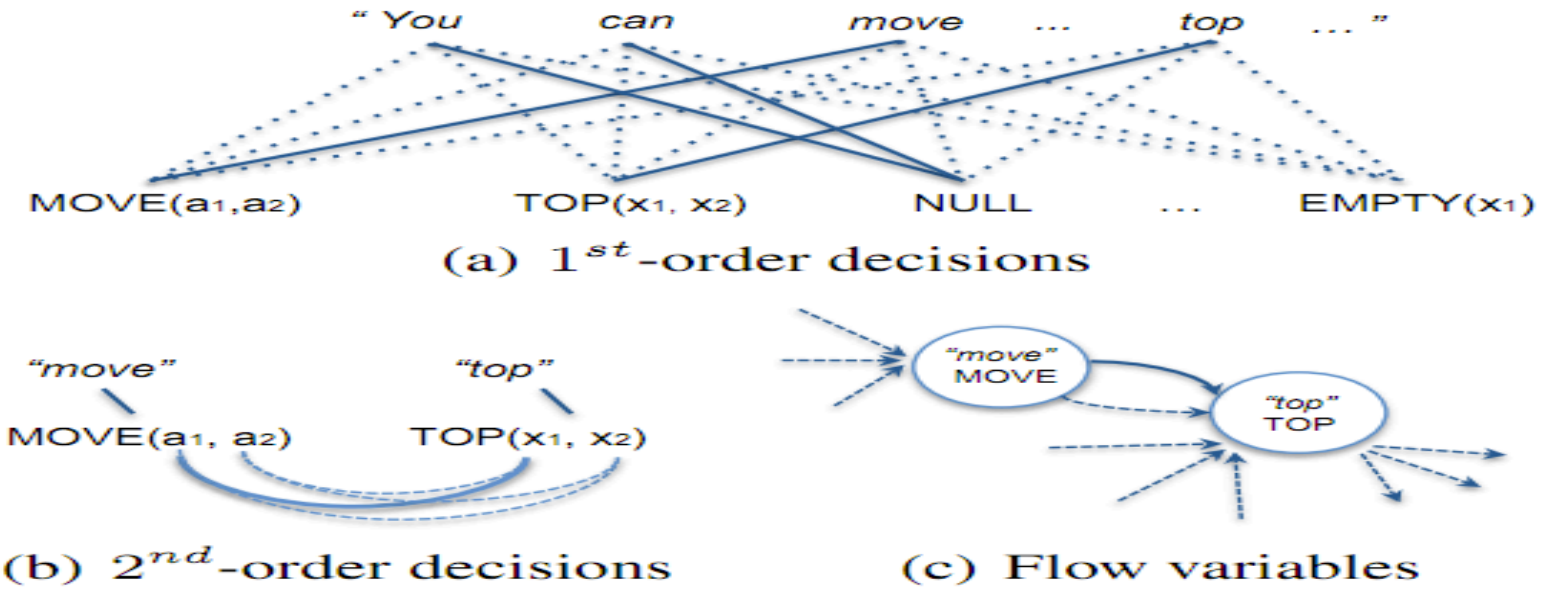
- X: You can move any top card to an empty freecell

- Y: Move(a1,a2) Top(a1, x) Card (a1) Empty(a2) Freecell(a2)

$$y^* = F_w(x) = \arg\max_{y \in Y} score(x, y) = \arg\max_{y \in Y, h \in H} w^T \Phi(x, y, h)$$

- Semantic interpretation is framed as an Integer Linear Program with three types of constraints:
  - Lexical Mappings: (1st order constraints)
    - **At most one predicate mapped to each word**
  - Argument Sharing Constraints (2nd order constraints)
    - **Type consistency; decision consistency**
  - Global Structure Constraints
    - **Connected structure enforced via flow constraints**

Lesson Interpretation as an inference problem



with three types of constraints:

- Lexical Mappings: (1st order constraints)
  - **At most one predicate mapped to each word**
- Argument Sharing Constraints (2nd order constraints)
  - **Type consistency; decision consistency**
- Global Structure Constraints
  - **Connected structure enforced via flow constraints**

Empirical Evaluation [IJCAI'11]

- Can the induced game-hypothesis generalize to new game instances?
  - Accuracy was evaluated over previously unseen game moves

Target Concept	Initial Model	Learned Model
FREECELL	0.78	0.956
HOMECELL	0.532	0.672
TABLEAU	0.536	0.628

- Can the learned reader generalize to new inputs?
  - Accuracy was evaluated over previously unseen game moves using classification rules generated from previously unseen instructions.

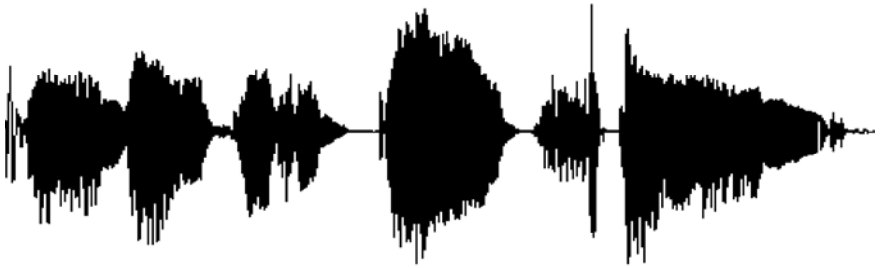
Target Concept	Initial Model	Learned Model
FREECELL	0.78	0.967
HOMECELL	0.532	0.668
TABLEAU	0.536	0.608

The language-world mapping problem

Skip

- How do we acquire language?

“the language”



[Topid rivvo den marplox.]

“the world”

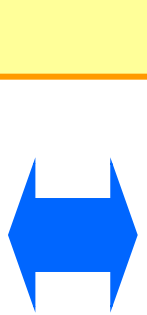
BabySRL: Learning Semantic Roles From Scratch

- A joint line of research with Cindy Fisher's group.
- Driven by Structure-mapping: a starting point for syntactic bootstrapping
- Children can learn the meanings of some nouns via cross-situational observations alone [Fisher 1996, Gillette, Gleitman, Gleitman, & Lederer, 1999;more]

- But how do they learn the meaning of verbs?
  - Sentences comprehension is grounded by the acquisition of an initial set of concrete nouns
  - These nouns yields a skeletal sentence structure — candidate arguments; cues to its semantic predicate—argument structure.
  - Represent sentence in an abstract form that permits generalization to new verbs

[Johanna rivvo den sheep.]

Nouns identified



BabySRL [Connor et. al, CoNLL'08, '09,ACL'10, IJCAI'11]

- Realistic Computational model developed to experiment with theories of early language acquisition
  - SRL as minimal level language understanding: who does what to whom.
  - Verbs meanings are learned via their syntactic argument-taking roles
  - Semantic feedback to improve syntactic & meaning representation

- Inputs and knowledge sources
  - Only those we can defend children have access to

- Key Components:
  - Representation: Theoretically motivated representation of the input
  - Learning: Guided by knowledge kids have

Exciting results – generalization to new verbs, reproducing and recovering from mistakes made by young children.

Minimally Supervised BabySRL [IJCAI'11]

- Goal: Unsupervised "parsing" – identifying arguments & their roles
- Provide little prior knowledge & only high level semantic feedback
  - Defensible from psycholinguistic evidence

- Unsupervised parsing
  - Identifying part-of-speech states
- Argument Identification
  - Identify Argument States
  - Identify Predicate States
- Argument Role Classification
  - Labeled Training using predicted arguments

Learning with Indirect Supervision

Input + Distributional Similarity

Structured Intermediate Representation (no supervision)

Binary weak supervision for the final decision

- Learning is done from CHILDES corpora
- IJCAI'11: indirect supervision driven from scene feedback

Conclusion

Thank You!

- Study of machine learning protocols that are based on natural language interpretation and world feedback
- Motivation:
  - Reduce annotation cost
  - Learning from Natural Instructions
  - Language Acquisition
- Technical approach is based on
  - (1) Learning structure with indirect supervision
  - (2) Constraining intermediate structure representation declaratively
- These were introduced via Constrained Conditional Models: Computational Framework for global inference and a vehicle for incorporating knowledge in structured tasks – Integer Linear Programming Formulations
- Applications: Game playing domain, Psycholinguistics, Semantic Parsing, ESL,…other structured tasks.