Learning and Inference in Natural Language

Dan Roth
University of Illinois, Urbana-Champaign
danr@cs.uiuc.edu
http://L2R.cs.uiuc.edu/~danr

Wen-tau Yih, Vasin Punyakanok; Chad Cumby

Comprehension

(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. Who is Christopher Robin? 2. When was Winnie the Pooh written?
3. What did Mr. Robin do when Chris was three years old?
4. Where did young Chris live? 5. Why did Chris write two books of his own?

Understanding Questions

Q: What is the fastest automobile in the world?

A1: …will stretch Volkswagen's lead in the [world's fastest growing vehicle market.] Demand for cars is expected to soar

A2: …the Jaguar XJ220 is the dearest (415,000 pounds), fastest (217mph) and most sought after car in the world.

Selecting an answer may require identifying some constraints on the answer, specified in the question, and selecting an answer that best satisfies them.

Ambiguity Resolution

Illinois' bored of education board

...Nissan Car and truck plant is ...
...divide life into plant and animal kingdom

(This Art) (can N) (will MD) (rust V) V,N,N

The dog bit the kid. He was taken to a veterinarian
a hospital

More NLP Tasks

- Prepositional Phrase Attachment

  buy shirt with sleeves, buy shirt with a credit card

- Word Prediction

  She ___ the ball on the floor (wrote, dropped;...)

- Name Entity/ Categorization

  Tiger was in Washington for the GPA Tour

- Information Extraction Tasks

  afternoon, Dr. Ab C will talk in Ms. De. F class..

Inference with Classifiers

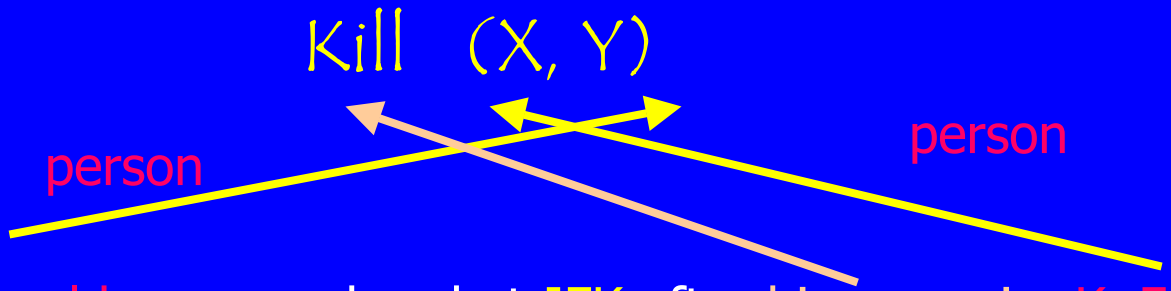
[$_{NP}$ He] [$_{VP}$ reckons] [$_{NP}$ the current account deficit] [$_{VP}$ will narrow] [$_{PP}$ to] [$_{NP}$ only # 1.8 billion] [$_{PP}$ in] [$_{NP}$ September]

- Classifiers
  1. Recognizing "The beginning of NP"
  2. Recognizing "The end of NP"
  3. Also for other kinds of phrases…
- Some Constraints
  1. Phrases do not overlap
  2. Order of phrases
  3. Length of phrases
- Use classifiers to infer a coherent set of phrases

Inference with Classifiers

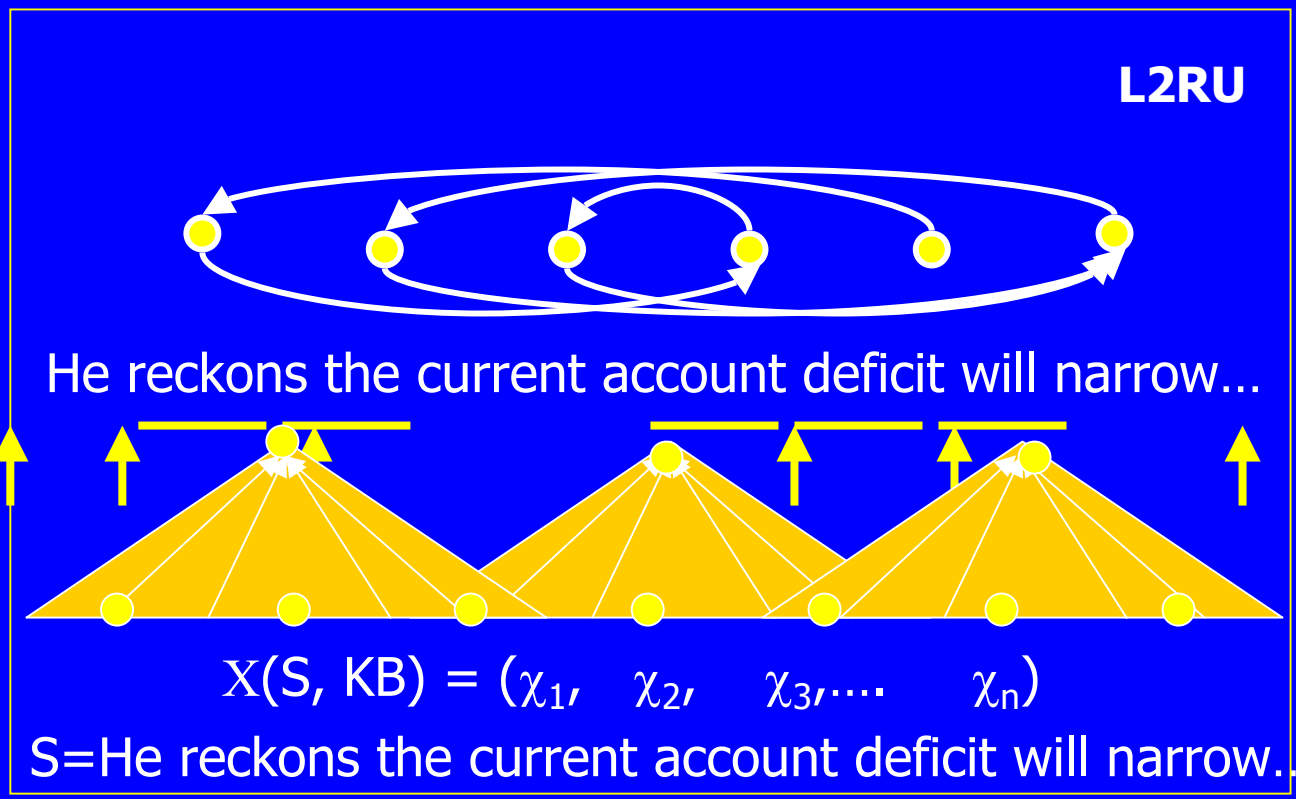
J.V. Oswald was murdered at JFK after his assassin, K. F. Johns…

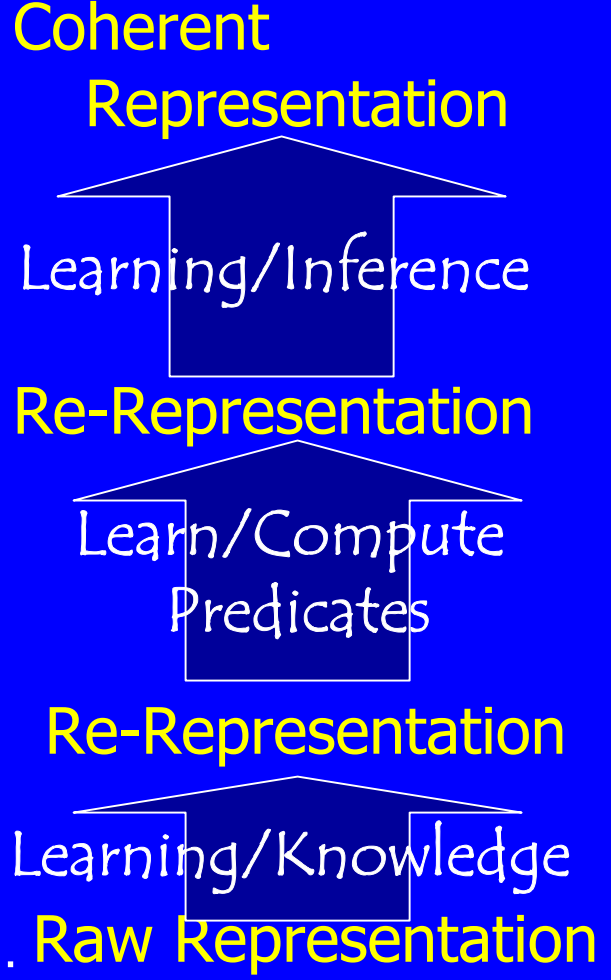
Identify:



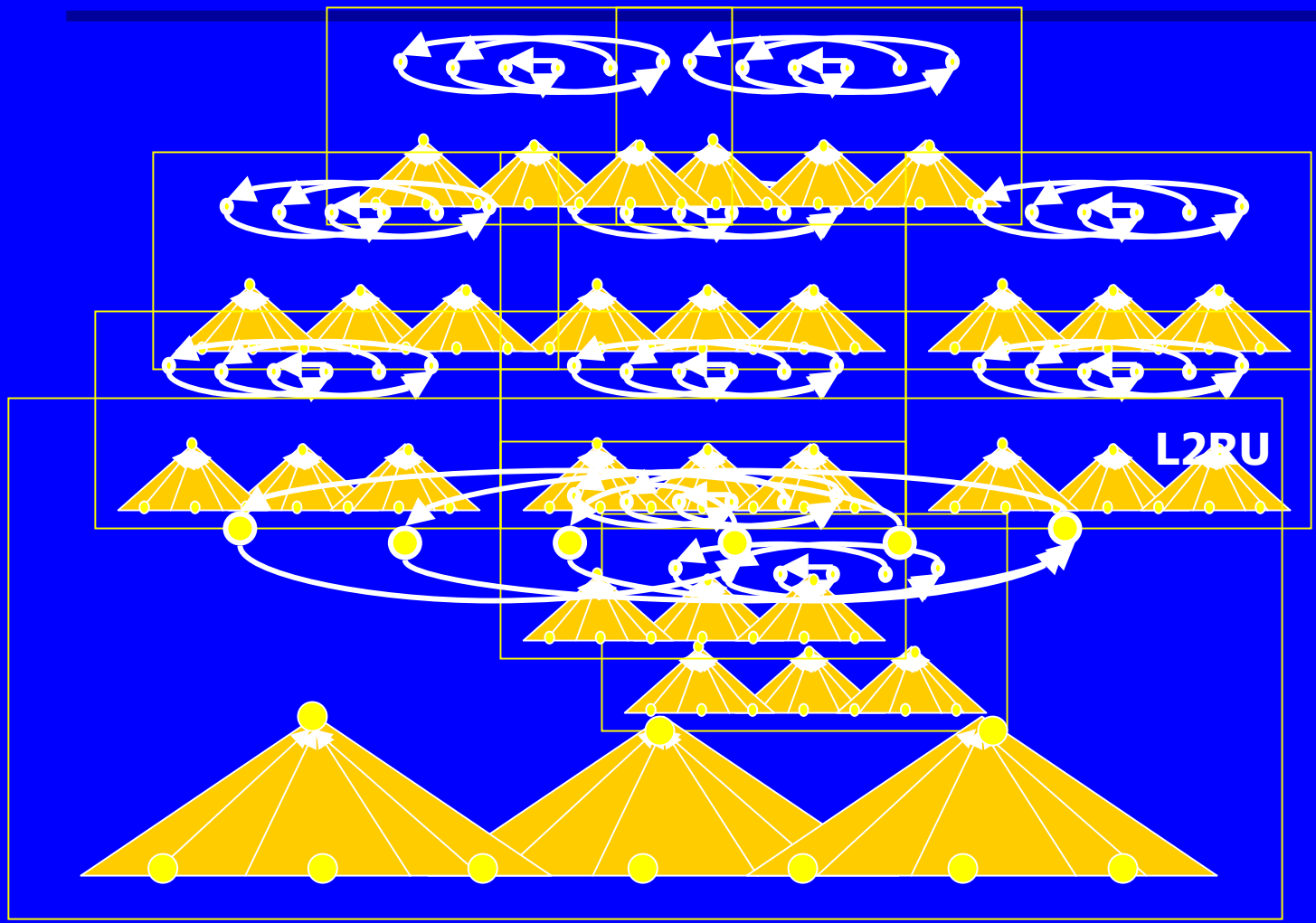
The Big Picture

[$_{NP}$ He] [$_{VP}$ reckons] [$_{NP}$ the current account deficit] [$_{VP}$ will narrow]....

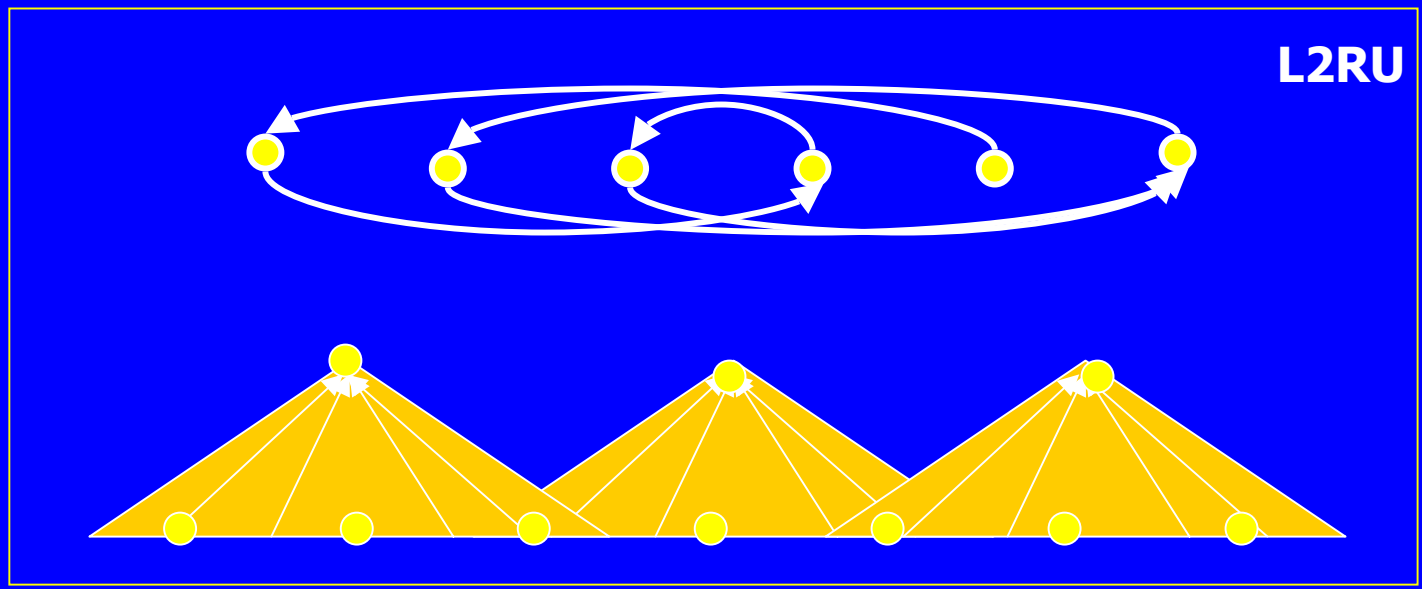




The Big Picture

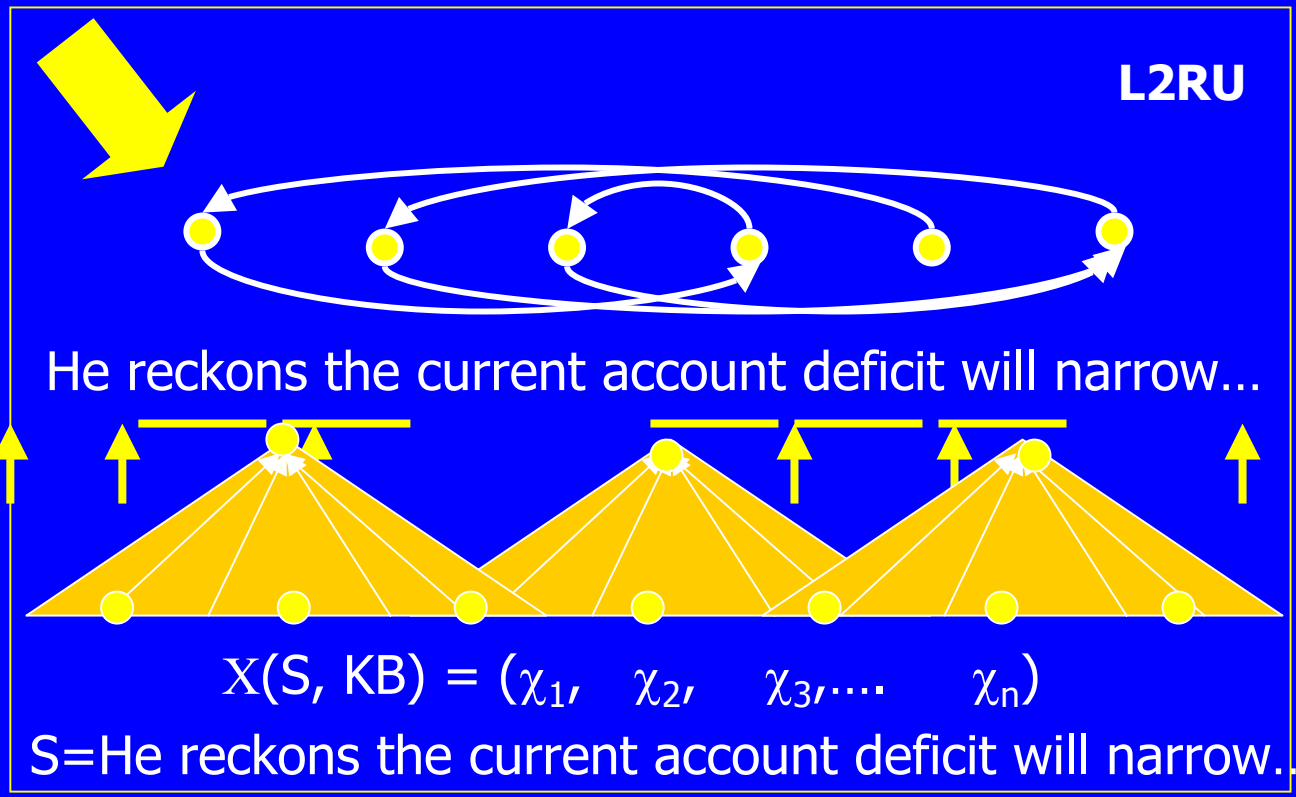


The Big Picture



The Big Picture

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow]....



Coherent Representation

Learning/Inference

Re-Representation

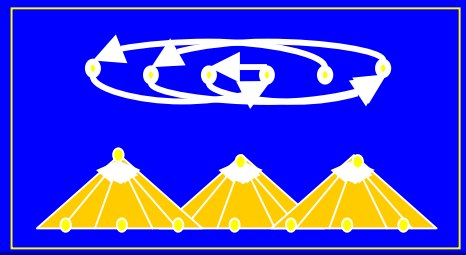
Learn/Compute Predicates

Re-Representation

Learning/Knowledge

Raw Representation

Plan of the Talk



Inference with classifiers

The use of different classifiers to yield a coherent inference.

- Inference with Sequential Constraints
  Phrase Identification Problem
- Classification
  Intermediate Representation; Conditional Probability
- Inference with General Constraint Structure
  Recognizing Entities and Relation

Identifying Phrase Structure

[$_{NP}$ He] [$_{VP}$ reckons] [$_{NP}$ the current account deficit] [$_{VP}$ will narrow] [$_{PP}$ to] [$_{NP}$ only # 1.8 billion] [$_{PP}$ in] [$_{NP}$ September]

- Classifiers
  1. Recognizing "The beginning of NP"
  2. Recognizing "The end of NP"
  3. Also for other kinds of phrases…
- Some Constraints
  1. Phrases do not overlap
  2. Order of phrases
  3. Length of phrases
- Use classifiers to infer a coherent set of phrases

Identifying Phrase Structure

- General Paradigm: Inference with Classifiers

Applications:

- Shallow parsing: chunking [Punyakanok, Roth NIPS'00]

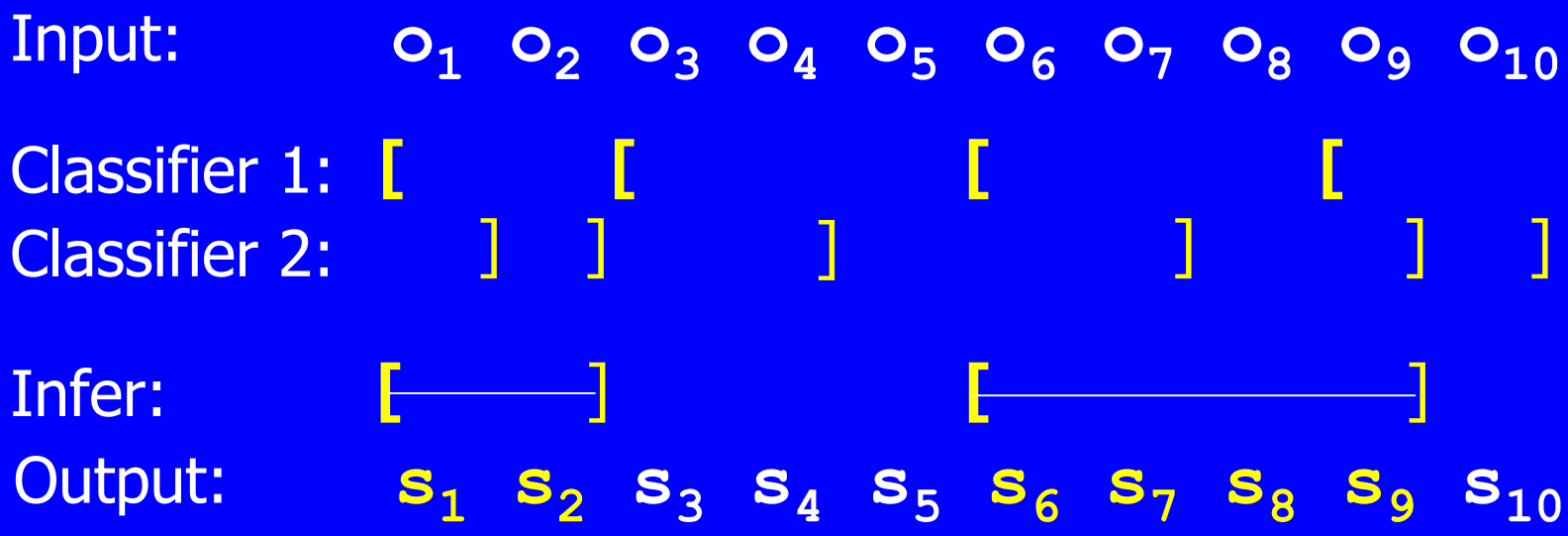
Other Application

- Names Entity Recognition
- Identifying document structure
- Shallow parsing: Clausing

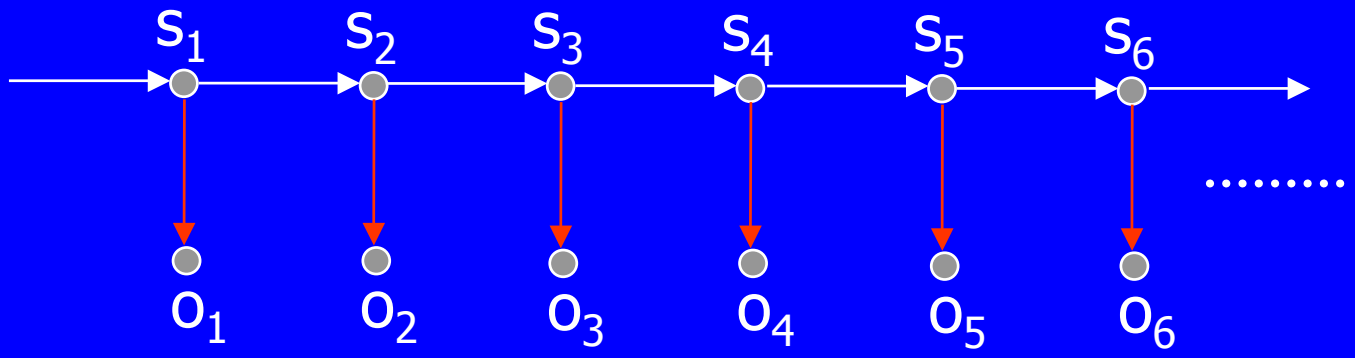
- Computational Biology: Detecting Splice Sites

Phrase Identification Problem

- Use classifiers' outcomes to identify phrases
- Phrase structure needs to satisfy some constraints



Hidden Markov Model

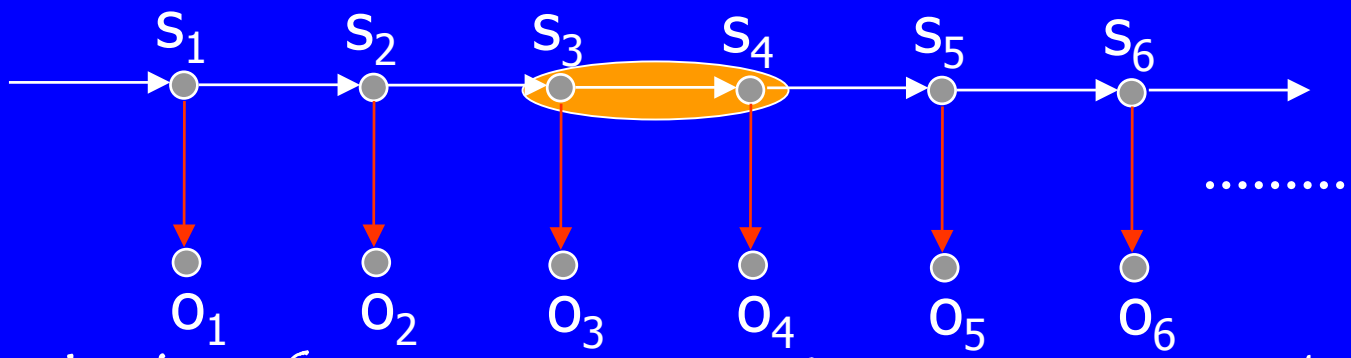


- Estimate
  - Initial state probability $P_1(s)$
  - Transition probability $P(s|s')$
  - Observation probability $P(o|s)$
- Goal
  - $\text{argmax}_{<S>} P(S|O)$
  - Can use dynamic programming (Viterbi)

Only local information is taken into account

Not exactly what we want

HMM with Classifiers



- Each classifier's output can be viewed as P(s|o)

$$P_t(o \mid s) = \frac{P_t(s \mid o) P_t(o)}{P_t(s)}$$

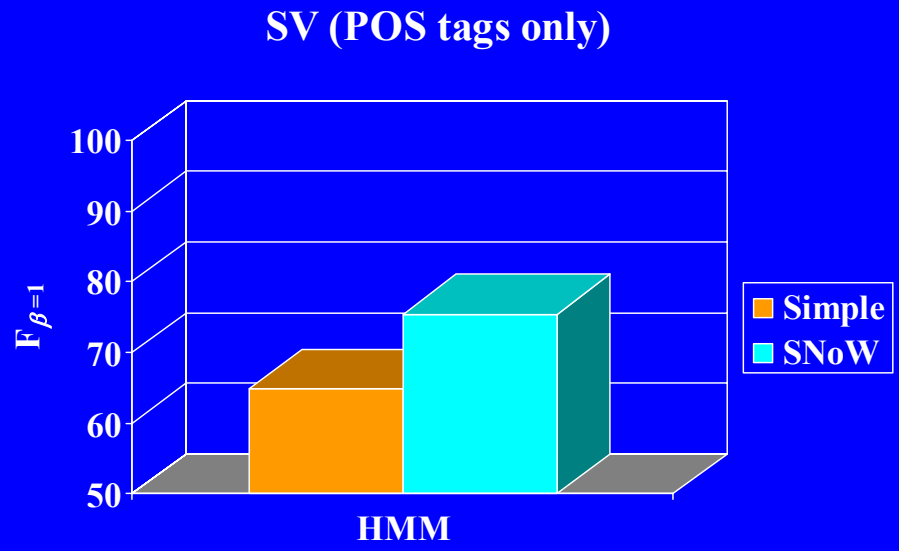
Global information can be taken into account

Constant at time t

$$P_t(s) = \sum_{s'} P_t(s \mid s') P_{t-1}(s')$$

➔ Constraints are incorporated via the transition probability

HMM with Classifiers



Standard (WSJ) Data Set
SV: 25k (3k) patterns

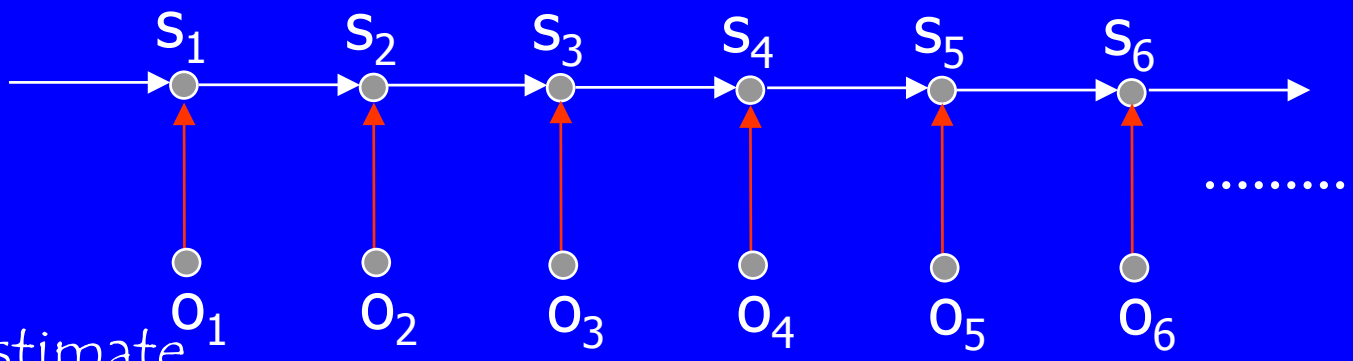
- Significant differences in performance
- Simple HMM not good enough for non trivial problems

- Adding Classifiers to the HMM scheme allows for modeling global correlations via classifiers' features
- Lost probabilistic interpretation of scoring function

Conditional Models

- Model States directly
- Directly incorporate the previous states in term of features
- Train many classifiers, each of which is projected on a previous state
  - More classifiers, but *simpler*

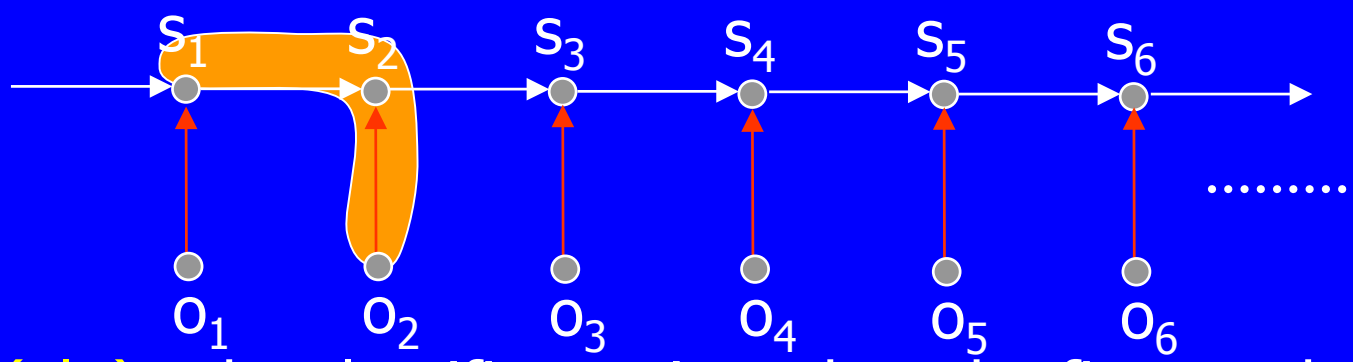
Projection-based Markov Model



- Estimate
  - Initial state probability $P_1(s|o)$
  - Transition probability $P(s|s',o)$
- Goal
  - $\text{argmax}_{<S>} P(S|O)$
  - $\text{argmax}_{<S>} \Pi_{t=2..n} [P(s_t|s_{t-1}, o_t)] P_1(s_1|o_1)$
  - Can use dynamic programming (Viterbi)

Unlike HMM, here the independence assumption allows $P(s|s',O)$

PMM with Projected Classifiers



- $P_1(s|o)$ - the classifier projected on the first symbol of sequences

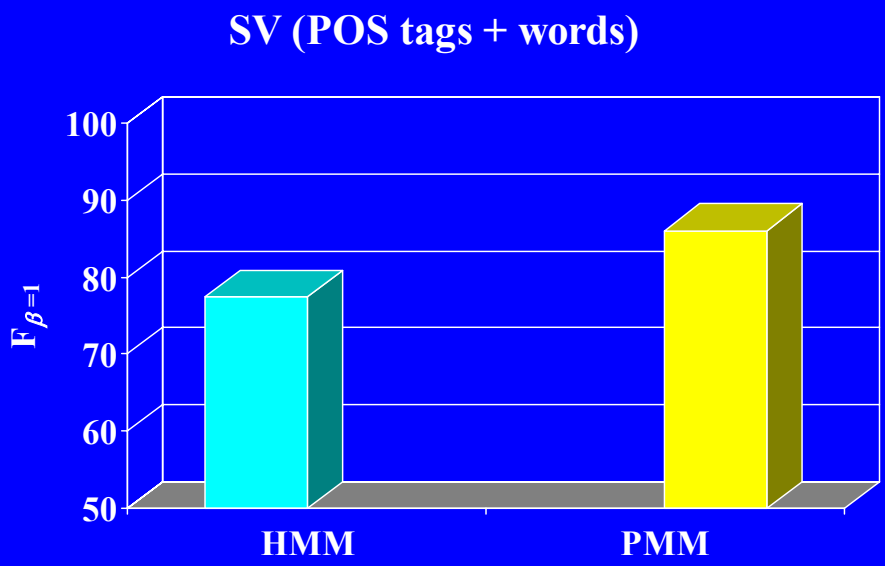
- $P(s|s',o) = P_{s'}(s|o)$ – the classifiers projected on each previous state [more classifiers, but same inference complexity]

➔ Constraints are incorporated via the transition probability

Can be used with more general distributional models [Lafferty et al.]

Projection-based Markov Model

- PMM significantly improves over HMM (with classifiers)
- State representation is better



Standard (WSJ) Data Set

- SV: 25k (3k) patterns

The Cost Function

- Markovian Method
  - Maximize the probability over the sequence
- The True Cost Function
  - Maximize the number of correct phrases
  - Minimize the number of wrong phrases

Constraint Satisfaction (CSCL)

- We extend the Boolean Constraint Satisfaction formalism to handle variables that are outcomes of classifiers
  - V - set of variables; Clauses: model constraints
  - f - A CNF, the CSP problem.
  - Satisfying assignment: τ: $V \rightarrow \{0,1\}$
  - Cost: c: $V \rightarrow \Re$
  - Find the solution τ that *minimizes the cost*

$$c(\tau) = \sum_{i=1..n} \tau(v_i)c(v_i)$$

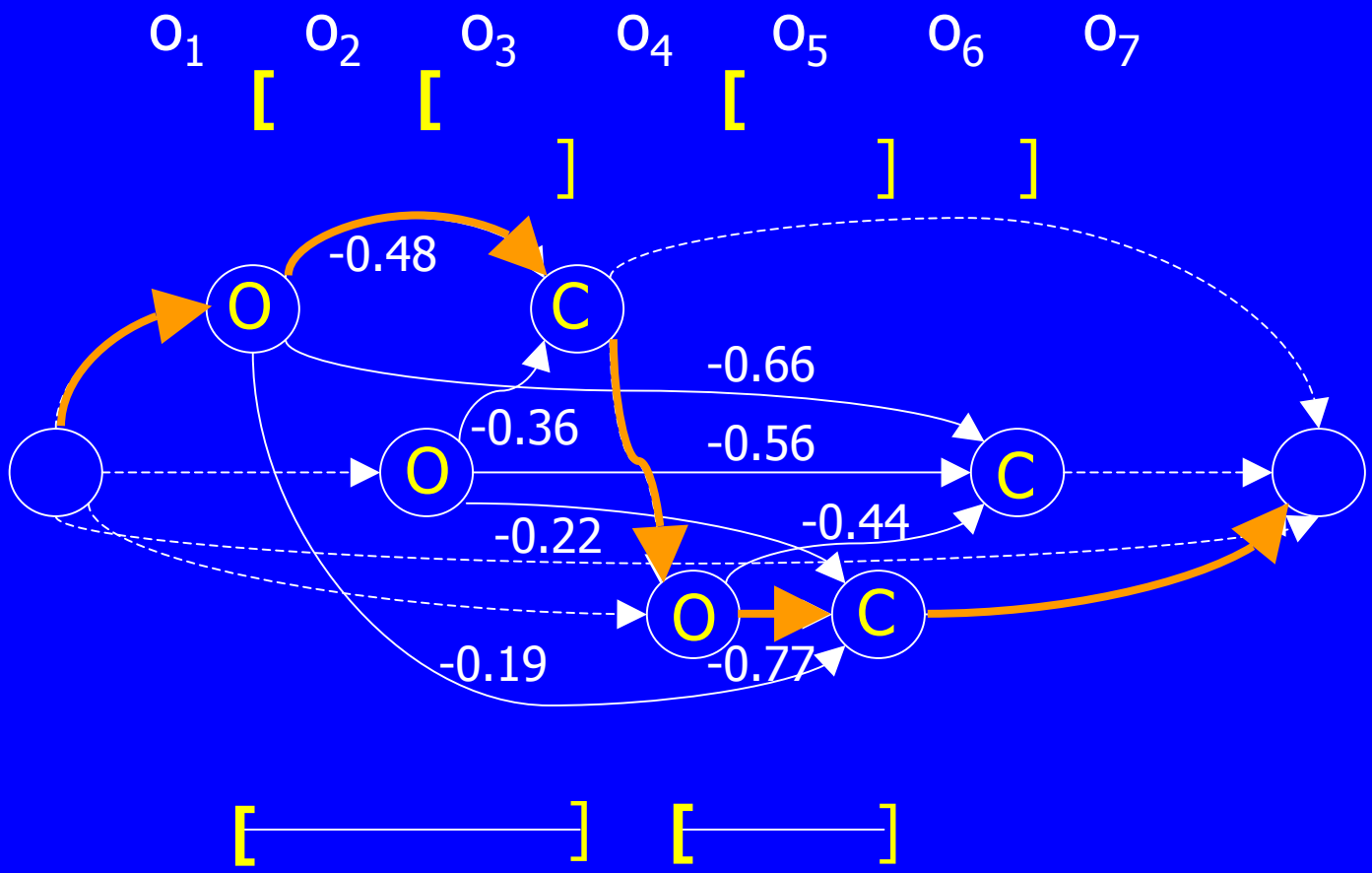
Modeling Constraints

- Let V be the set of all possible phrases

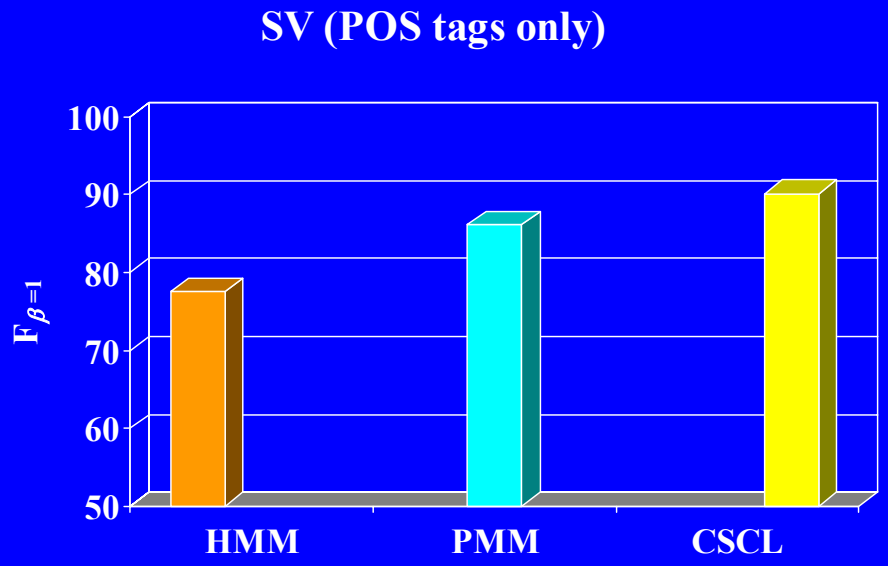
- $f = \bigwedge_{v_i \text{ overlaps } v_j} (\neg v_i \vee \neg v_j)$
- c = 1-P(O)P(C) → -P(O)P(C)
  - P(O) and P(C) are supplied by classifiers
  - Maximizes the expected number of correct phrases.

- CSP in general is hard
- Structure of the constraints yields a problem that can be solved by shortest path algorithm

Constraints Solution



CSCL

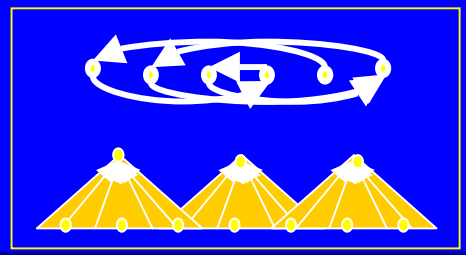


Standard (WSJ) Data Set

- SV: 25k (3k) patterns

- CSCL performs better
  - Handles better Longer patterns
  - Better cost function
  - Competitive with other approaches tried on this task.

Plan of the Talk



Inference with classifiers

The use of different classifiers to yield a coherent inference.

- Inference with Sequential Constraints
  Phrase Identification Problem

- Classification
  Intermediate Representation; Conditional Probability

- Inference with General Constraint Structure
  Recognizing Entities and Relation


IF I WAS IN CHARGE, WE'D NEVER SEE GRASS BETWEEN OCTOBER AND MAY.
ON "THREE," READY? ONE... TWO... THREE!
SNOW!
I SAID SNOW! C'MON! SNOW!
SNOW!
OK THEN, DON'T SNOW! SEE WHAT I CARE! I LIKE THIS WEATHER! LET'S HAVE IT FOREVER!
PLEEASE SNOW! PLEASE?? JUST A FOOT! OK, EIGHT INCHES! THAT'S ALL! C'MON! SIX INCHES, EVEN! HOW ABOUT JUST SIX??

SNoW http://L2R.cs.uiuc.edu/~danr/snow.html

- A successful learning approach tried on several NLP problems
- A learning architecture tailored for high dimensional problems
- Multi Class Learner; Robust confidence in prediction
- A network of linear representations
- Several update algorithms are available
- Most successful – a multiplicative update algorithm, a variation of Winnow (Littlstone'88)
- Feature space: Infinite Attribute Space $\{0,1\}^\infty$
  - examples of variable size: only active features
  - determined in a data driven way

Classifiers

- Output:

  What classifiers can we use?

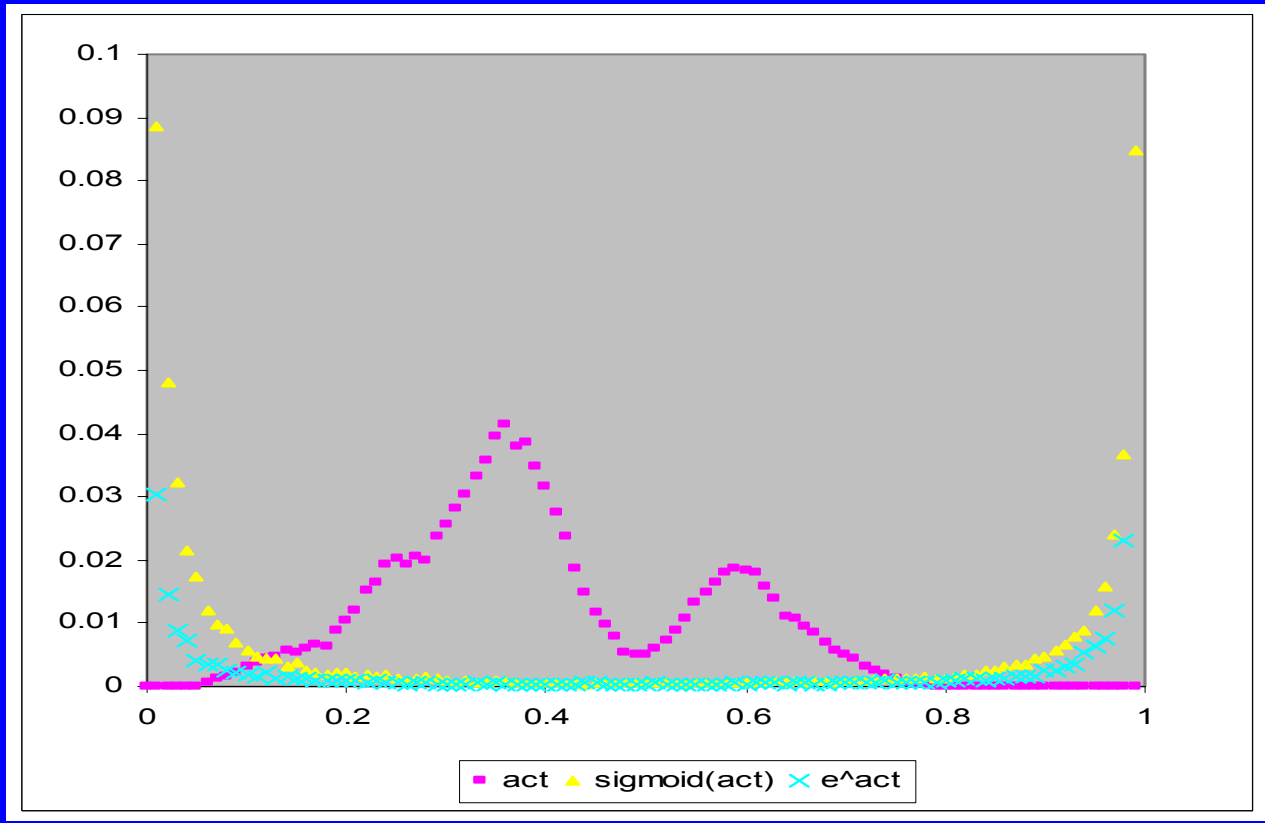
  Do we get what we want?

- Input:

  What features?

Conditional Probabilities

$y = \#\{z \mid f(z) = x\}$



- Data: Two class (Open/NotOpen Classifier)

Conditional Probabilities

For example z:

Y=Prob(label=1 | f(z)=x)

If Prob(1 | f(z)=x) = x

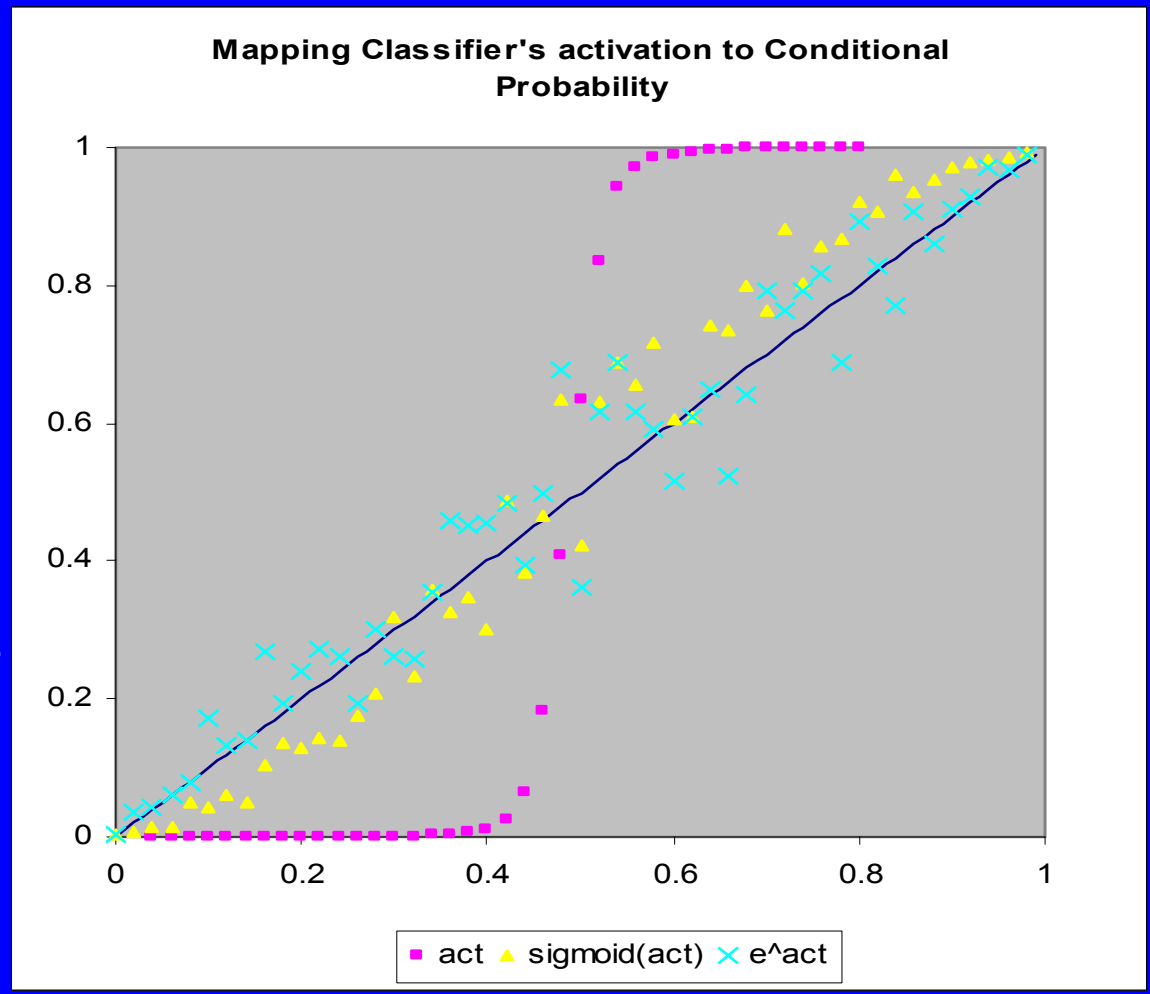
Then f(z) = Prob(1 | z)

Plotted for SNoW (Winnow)

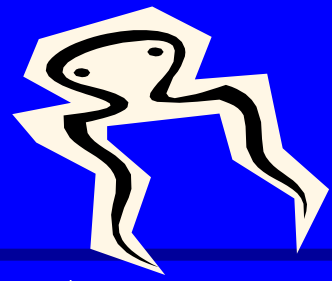
Holds for many classifiers

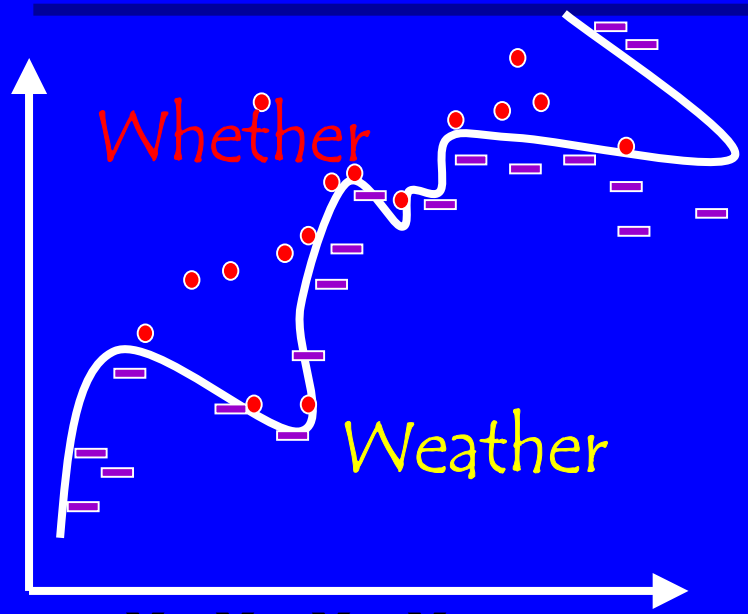
See Tong Zhang's ICML'02

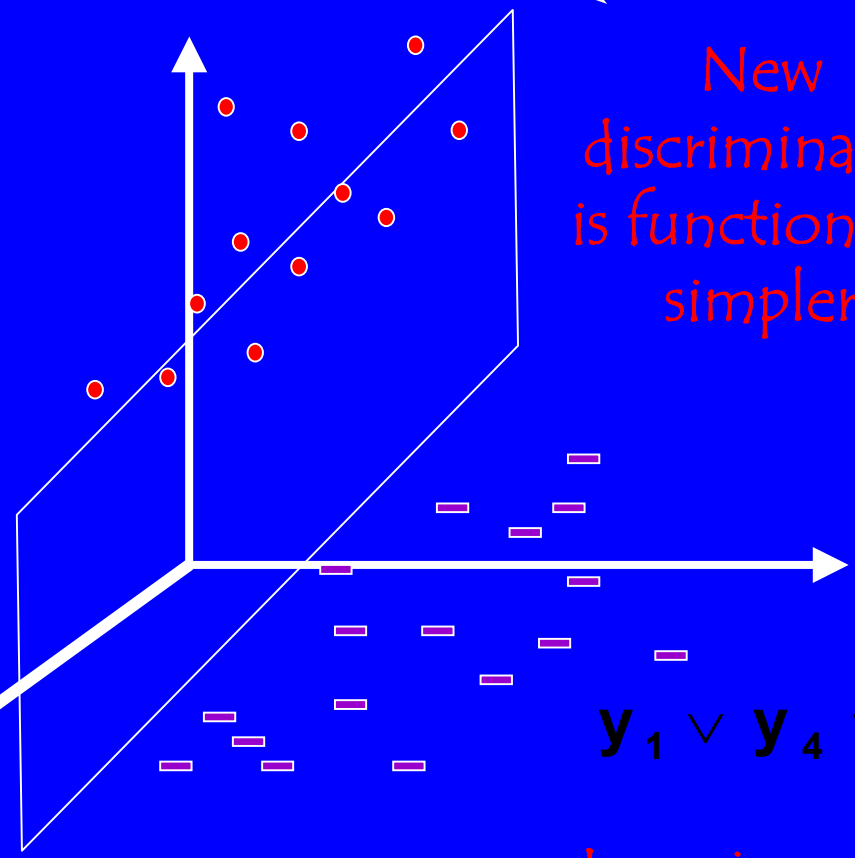
for theoretical justification



Scenario







New discriminator is functionally simpler

$x_1, x_2, x_3, x_4, \ldots$

$x_1 x_2 \bar{x}_3 \vee \bar{x}_1 x_4 \bar{x}_3 \vee x_3 \bar{x}_2 x_5$

$y_1 \vee y_4 \vee y_5$

Input Transformation

Learning

$\{x_1 x_2 \bar{x}_3, x_1 x_2 x_3, x_1 \bar{x}_2 \bar{x}_3, x_1 x_2 \bar{x}_4, \bar{x}_1 x_4 \bar{x}_3, x_3 \bar{x}_2 x_5, \ldots\}$

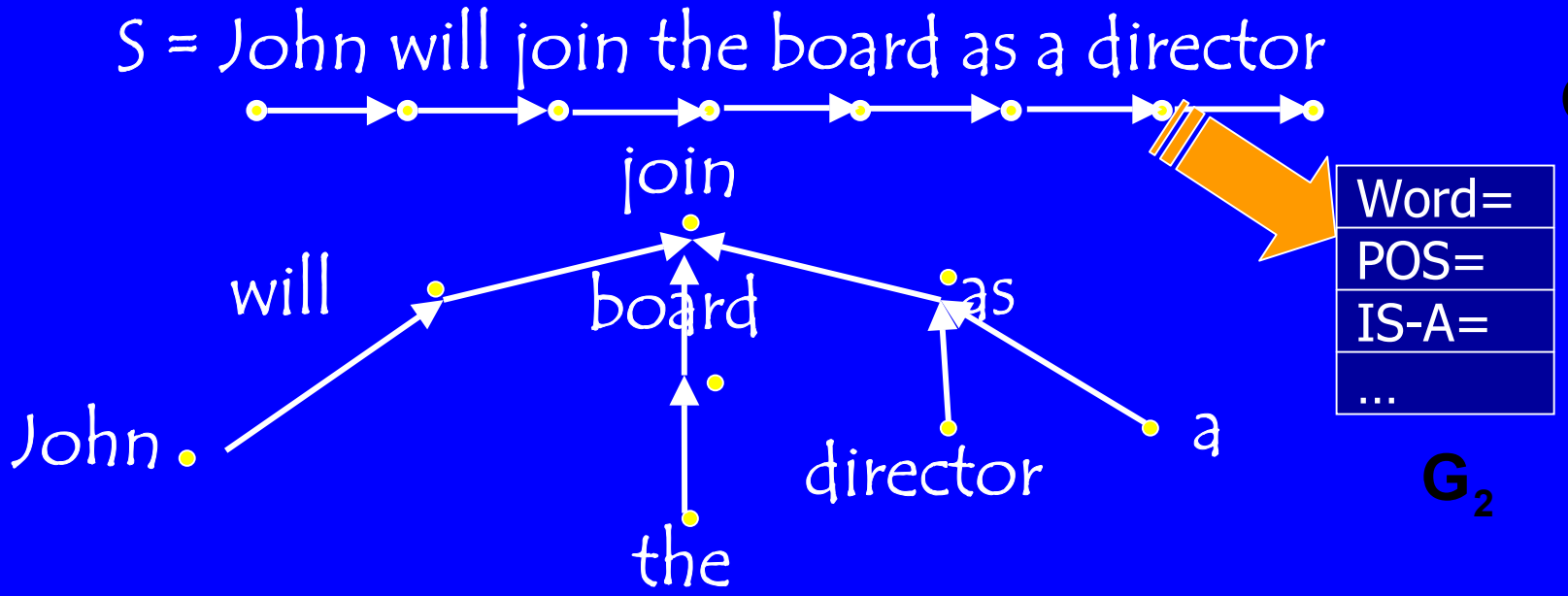
A Better Feature Space

- Feature efficient algorithms allow us to the extend the types of intermediate representations used.
- More potential features is not a problem
- Representing interesting concepts often requires
  - The use of relational expressions.
  - Better exploitation of the structure
- Generate complex features that represent (also) relational (FOL) constructs
- Structure: Extend the generation of features beyond the linear structure of the sentence.

Structured Domain

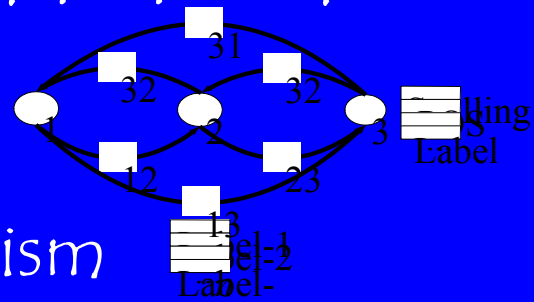
afternoon, → Dr. → Ab → C → ...in → Ms. → De. F class..

[$_{NP}$ Which type] [$_{PP}$ of] [$_{NP}$ submarine] [$_{VP}$ was bought]
[$_{ADVP}$recently] [$_{PP}$ by] [$_{NP}$ South Korea] (. ?)



Structured Domain

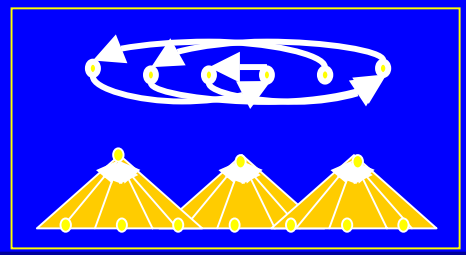
- Domain Elements are represented as labeled graphs



- Feature Description Logic formalism
  Re-representation of a domain element as a feature vector done via subsumption

- Feature are generated in a way that allows abstraction over different instantiations (relational) [Roth;Yih IJCAI'01; Cumby;Roth ILP'02].

Plan of the Talk



Inference with classifiers

The use of different classifiers to yield a coherent inference.

- Inference with Sequential Constraints
  Phrase Identification Problem

- Classification
  Intermediate Representation; Conditional Probability

- Inference with General Constraint Structure
  Recognizing Entities and Relation

Extensions

- Dealing with hierarchical structure

  [Carreras, Marquez, Punyakanok, Roth, ECML'02]

- Dealing with more general structure of constraints on the classifiers outcome

  [Roth, Yih COLING'02]

Clause Identification (I)

- A clause is a sequence of words in a sentence that contains a subject and a predicate:

([$_{NP}$ Balcor], **(** [$_{NP}$ which] **(** [$_{VP}$ has] [$_{NP}$ interests] [$_{PP}$ in] [$_{NP}$ real estate] **))** , [$_{VP}$ said] **(** [$_{NP}$ the position] [$_{VP}$ is newly created]**) .)**

- Chunks, annotated with their types are part of the input.

Clause Identification (II)

Classifiers:

- Start of a clause
- End of a clause
- Score of a clause (s,e)

Algorithm:

- Recursively, score splits of sentences into clauses.

$$S = \text{argmax} \sum_{(s,e)} \text{score}(s,e)$$

- use dynamic programming

Clause Identification (III)

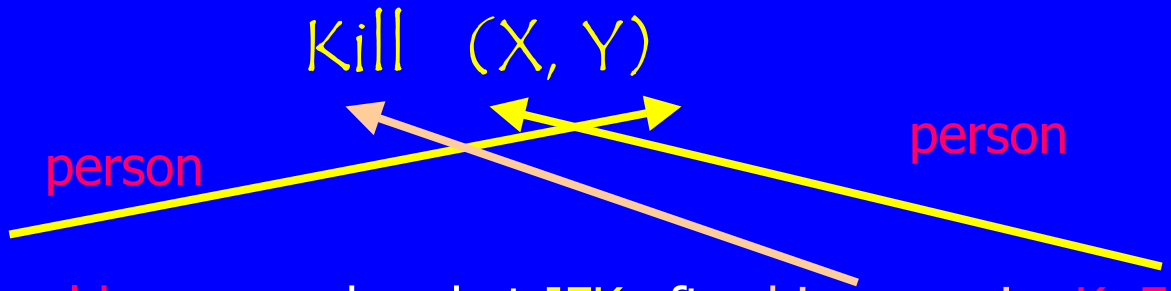
- Several Scoring functions are possible
- Other schemes, generalizing previous schemes are possible.

- Results are significantly better than local classifiers based approaches (CoNLL'01)

Inference with Classifiers

J.V. Oswald was murdered at JFK after his assassin, K. F. Johns…

Identify:



Identifying Entities and Relations

- Recognizing and classifying entities and relations in a key task in many NLP problems
  - Information Extraction
    - Extracting meaningful entities like *title* and *salary*
    - Knowing if these entities are associated with the same position
  - Question Answering
    - "Where was Poe born?"
    - Finding a *person* (who is Poe), a *place*
    - Knowing that the person and the place has relation *born_in*

Inference with Classifiers

1. Learn classifiers for each entity and relation.

2. Classifiers represent a conditional probability for each variable, given the observed data.

3. Incorporate this information, along with constraints, in making global inference for the most probable assignment to all variables of interest (entities and relations).

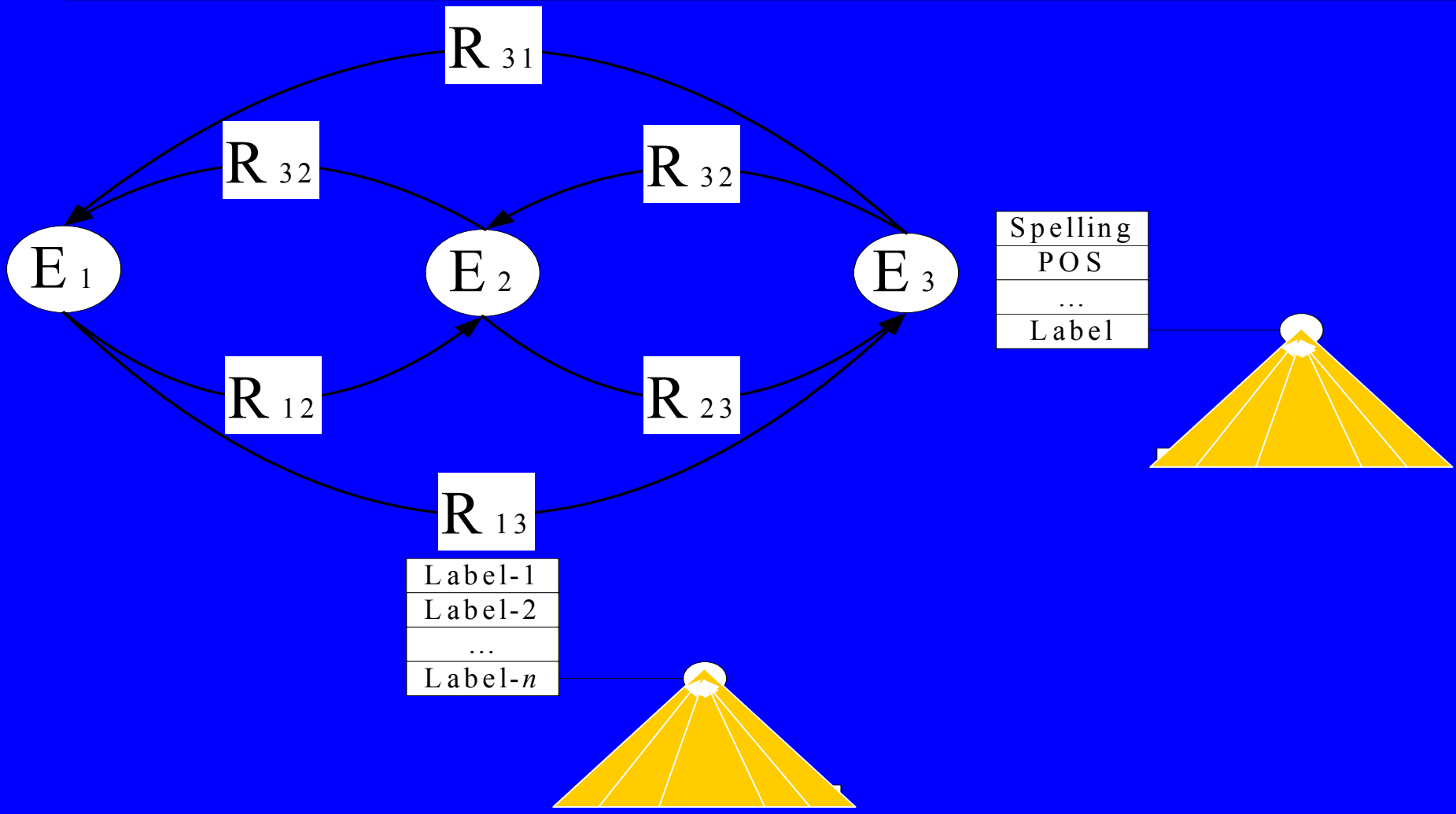
Basic Terms

- Dole's wife, Elizabeth, is a native of Salisbury, N.C.

  E_1 E_2 E_3

- Entity
  - A single word or a set of consecutive words with a predefined boundary.
  - Segmentation (phrase detection) assumed solved.
- (Binary) Relation
  - Any pair of entities (R_{12}, R_{21}, R_{13}, R_{31}, R_{23}, R_{32})

Conceptual View



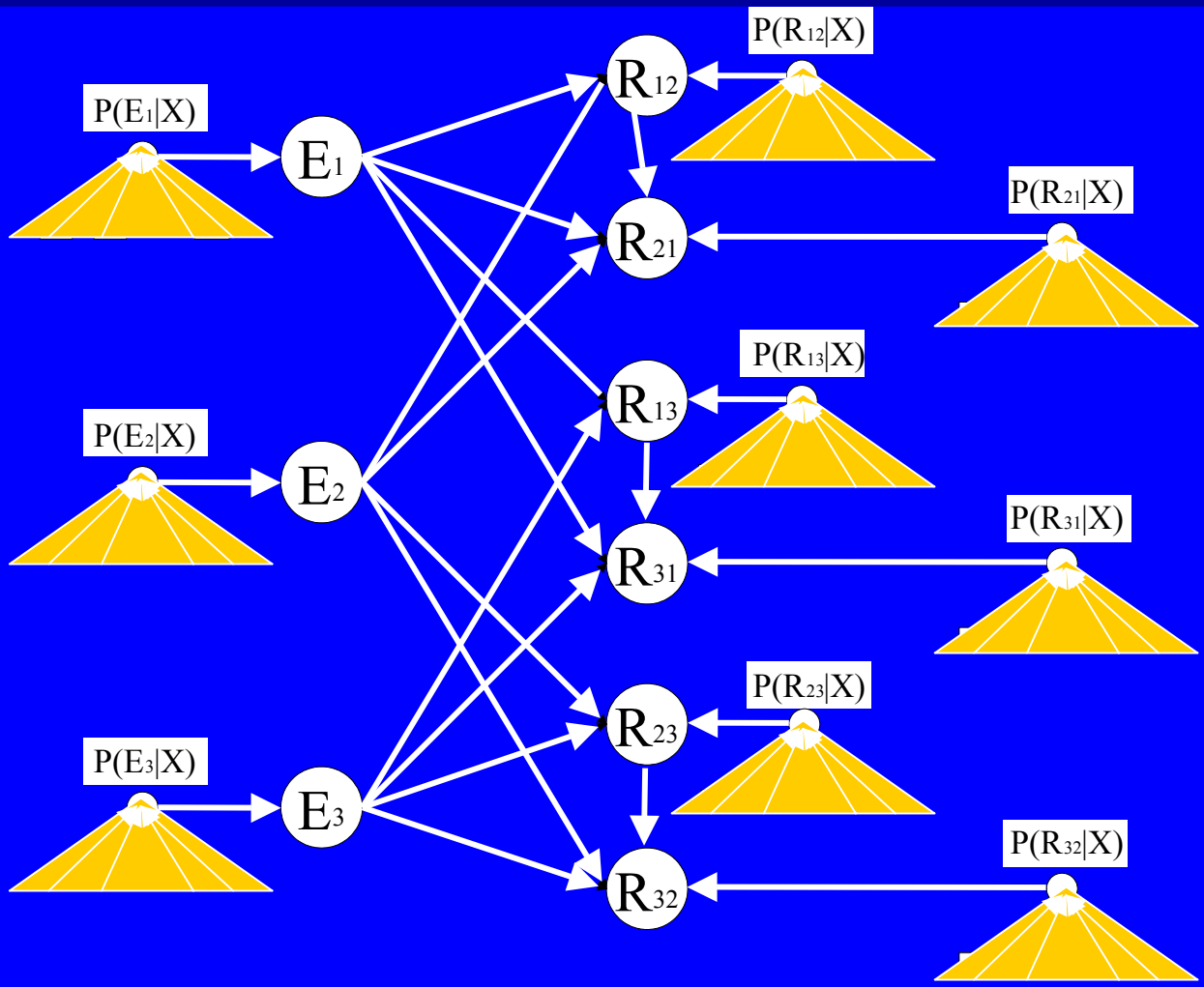
Identifying Entities and Relations

- Goal – coherently label entities & relations
- Exploit mutual dependency
  - The value of an entity or relation depends not only on its local properties, but also on properties of other entities and relations.
  - The outcomes of entity and relation predictors are mutually dependent.
  - E.g. **E1** depends on $\mathbf{R_{12}}$; $\mathbf{R_{12}}$ depends on **E1** and **E2**

Constraints

- A constraint C is a 3-tuple (R, E^1, E^2)
  - If the relation is R, then the legitimate class labels of its two entity arguments are E^1 and E^2.
- Examples
  - (born_in, person, location)
  - (spouse_of, person, person)
  - (murder, person, person)
- Constraints are modeled as conditional probabilities in a Bayesian network. $P(R \mid E^1, E^2)$

Belief Network



Experiments

- Basic : Local classifiers
  - Tests baseline
  - May produce predictions that are not coherent
- BN : belief network inference model
  - Can do exact inference
  - Most variables are abstracted aways and used only in learning

$$(e_1, ..., e_n, r_{12}, ..., r_{n(n-1)}) = \arg\max_{e_i, r_{jk}} Prob(E_1, ..., E_n, R_{12}, ..., R_{n(n-1)})$$

Results

Approach	person			location			kill		
	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1
Basic	95.8	92.2	94.4	66.9	92.5	76.9	76.1	39.3	51.7
BN	89.0	96.6	92.6	70.9	89.3	78.5	59.3	79.7	63.3
Improvement	-7.1%	4.8%	-1.4%	6.1%	-3.4%	2.1%	-29.1%	102.6%	22.6%

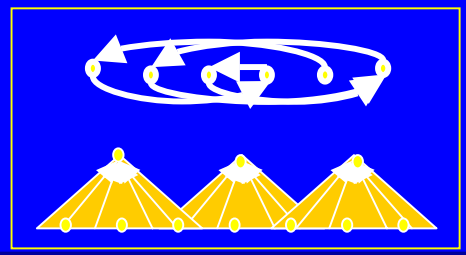
Approach	person			location			born_in		
	Rec	Prec	F_1	Rec	Prec	F_1	Rec	Prec	F_1
Basic	90.0	89.1	89.4	89.7	94.0	91.6	87.3	50.3	63.6
BN	84.5	89.7	86.8	87.7	94.2	90.6	86.5	70.8	76.32
Improvement	-6.2%	0.6%	-2.9%	-2.1%	0.3%	-1.1%	-0.9%	41.0%	22.0%

Discussion

- Weaknesses of preliminary approach
  - Modeling: directed model
  - Data

- Current/Future work :
  - Markov Random Fields
  - Bootstrapping:
    Use partial labeling to exploit indirect constraints-based correlation to replace direct supervision

Final Thoughts



- Research on a unified view of
  Learning, Knowledge Representation, Inference
  aiming at making progress in natural language

- Supported by theoretical work on learning in high dimensions, knowledge representation, inference algorithms,…

- In addition to theoretical and algorithmic research there is a need for a programming paradigm that allows one to reason at the right level.

Comprehension

(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. Who is Christopher Robin? 2. When was Winnie the Pooh written?
3. What did Mr. Robin do when Chris was three years old?
4. Where did young Chris live? 5. Why did Chris write two books of his own?

Comprehension

(NEW YORK: May 1, 1931)-The world's tallest building opened today in New York City. It is called the Empire State Building.

At noon, two small children cut a ribbon. It was in front of the main door. The ribbon was made from paper. After it was cut, people walked through the door for the first time. Hundreds of people were there. All day long, they took part in a big party on a floor 86 stories high.

This building holds as many people as there are in some cities. Each day, 25,000 workers will ride one of the 63 elevators. Another 15,000 people will visit. They might shop or get their hair cut.

The Empire State Building is a skyscraper. It is so tall that it seems to scrape the skies. At the very top is a tall, pointed tower. People can go to the top and look at the views. They can see at least 50 miles away.

1. Who cut the ribbon?
2. What is the name of the building?
3. When was the ribbon cut?
4. Where is the building?
5. Why do you think people cannot see the top of the building on some days?

Comprehension

(SALEM, MASSACHUSETTS, 1899) - The merry-go-round is 100 years old this year! No other park ride has lasted so long. The first merry-go-round in the United States was built in 1799. It was built in a park in Salem.

A merry-go-round has wooden animals on it. The most favorite are the horses. They are attached to poles. They can move up and down. The animals are on a platform. It turns in a circle. The merry-go-round spins to the sound of music. In time, the weather damages the animals. They lose their bright colors. Then, workers must fix the animals. They sand away all the old paint. Then they patch the broken parts. The next step is to paint the animals white. After this, bright colors of paint are added. Then the animals are carefully put back in place. Another name for a merry-go-round is "carousel" (CAR-uh-sel). Call it what you like. By any name, it's great fun!

1. Who fixes the merry-go-round? 2. Why do merry-go-rounds need to be fixed?
3. What is another name for a merry-go-round?
4. When was the first one built in the United States?
5. Where was the first one built in the United States?