Global Learning of Entailment Graphs

Jonathan Berant

Joint work with Ido Dagan and Jacob Goldberger

January, 2011

Task: Entailment (inference) Rules between Predicates

- Binary predicates: specify relations between a pair of arguments:
  - *X cause an increase in Y*
  - *X treat Y*
- *Entailment rules*:
  - *Y is a symptom of X → X cause Y*
  - *X cause an increase in Y → X affect Y*
  - *X's treatment of Y → X treat Y*

Motivation

- Textual entailment systems are useful for applications

QA:
Question: What affects blood pressure?
"Salt causes an increase in blood pressure"

IE: X purchase Y

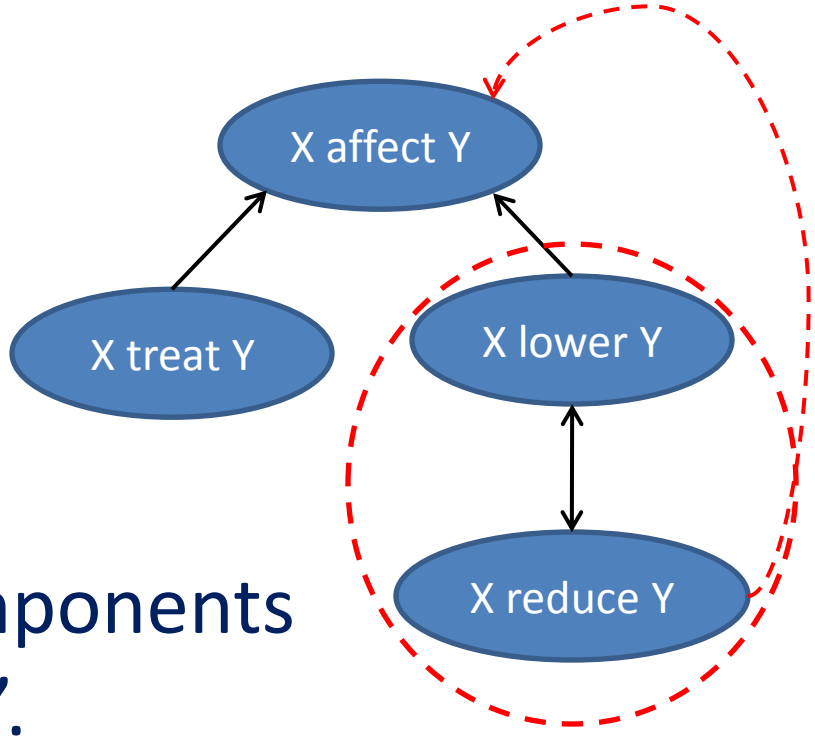
IBM	Coremetrics
Google	reMail
Yahoo	Overture

IR:
Query: symptoms of IBS

"IBS is characterized by vomiting"

Entailment Graphs

- Nodes: predicates
- Edges: entailment rules



- Entailment is transitive
- Strong connectivity components represent "equivalence".
- The DIRT rule-base (Lin and Pantel, 2001) uses only pairwise information.

Contributions

- Global graph learning: utilizing interactions between entailment rules for learning
- Integer Linear Programming (ILP) formulation
- Scaling: applying ILP on larger graphs using *Decomposition* and *Incremental ILP*

Outline

1. Background
2. Learning algorithm (high-level)
3. Focused entailment graphs
4. Typed entailment graphs
5. Conclusions

Background

Local Learning

Input: pair of predicates (p_1,p_2)
Question: $p_1 \rightarrow p_2$? (or $p_1 \leftrightarrow p_2$)

- Sources of information (monolingual corpus):
  1. Lexicographic: WordNet (Szpektor and Dagan, 2009), FrameNet (Ben-aharon et al, 2010)
  2. Pattern-based (Chklovsky and Pantel, 2004)
  3. Distributional similarity (Lin and Pantel, 2001; Sekine, 2005; Bhagat et al, 2007;Yates and Etzioni, 2009; Poon and Domingos 2010; Schoenmackers et al., 2010)

Properties of Information Sources

1. Lexicographic:
   - WordNet relations: hyponym, derivation, entailment
   - Limited coverage: *"X cause a reduction in Y"*
2. Pattern-based:
   - *"he <u>scared</u> and even <u>startled</u> me"*
   - Requires very large corpus (web-scale)
   - Distinguishes different semantic relations:
     - "to X and even Y" vs. "Either X or Y"
3. **Distributional similarity**

Distributional Similarity

X affect Y		
X	Y	#
insulin	metabolism	7
Zantac	BP	4
…	…	…

X treat Y		
X	Y	#
Zantac	BP	9
diet	diabetes	2
…	…	…

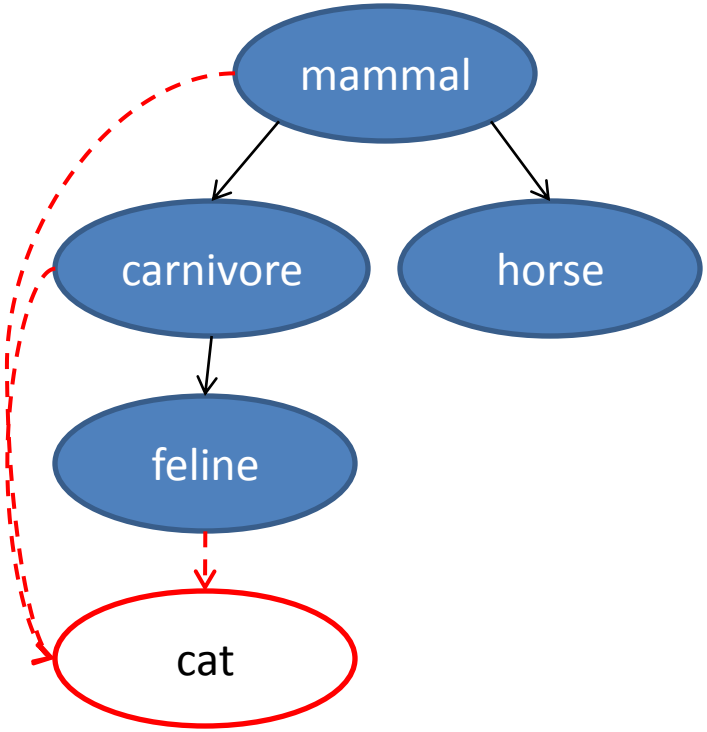
- Vary w.r.t to representation and similarity computation
- Good coverage
- Discerning exact semantic relation is difficult

Global Learning

Input: Set of predicates P

Question: Find $E = \{ (p_1,p_2) \mid p_1 \rightarrow p_2 \}$

- Snow et al. (2006) presented an algorithm for taxonomy induction.
- At each step add the concept that maximizes the likelihood under a transitivity constraint.



Global Learning

- Resolver (Yates and Etzioni, 2009)
  - Clustering of concepts and relations
- OntoUSP (Poon and Domingos, 2010)
  - Unsupervised semantic parsing
  - Ontology Induction (is-a hierarchy)
- Co-reference Resolution (Finkel and Manning, 2008)
- Temporal Information Extraction (Ling and Weld, 2010)

High-level Description of Algorithm

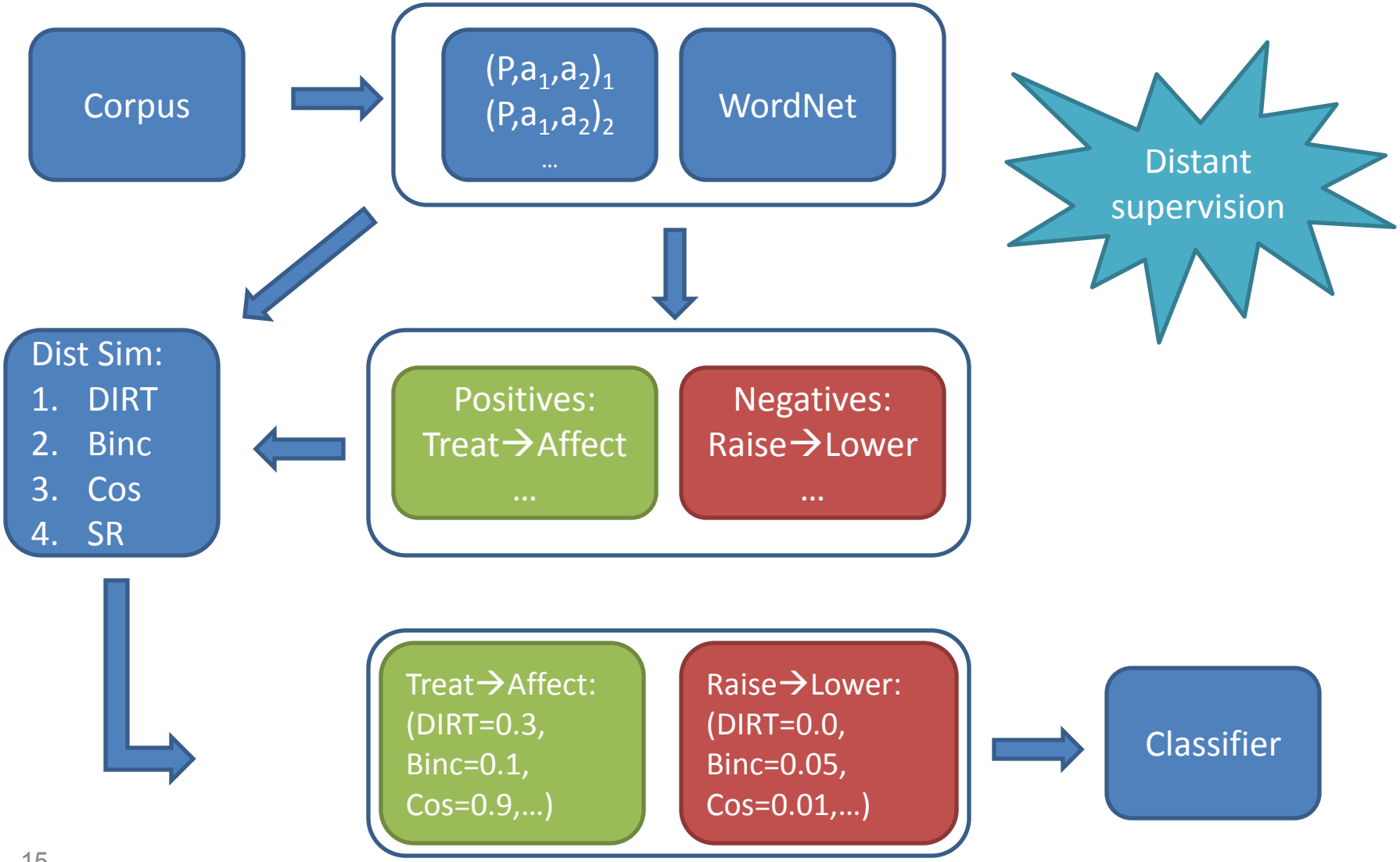
Learning Entailment Graph Edges

- Two step algorithm:

Input: set of predicates *P*

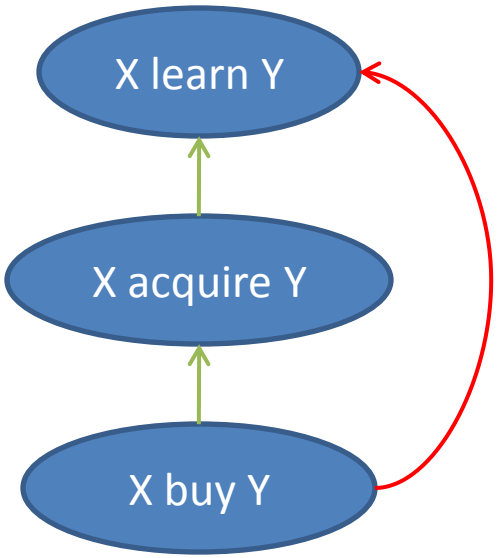
1. Train a **local** entailment classifier: given the predicates (p_1,p_2), estimate a score for $p_1 \rightarrow p_2$
2. Learn the edges of the graph using the local entailment classifier and a transitivity constraint

Local Entailment Classifier – Step 1



Learning Edges - Challenges

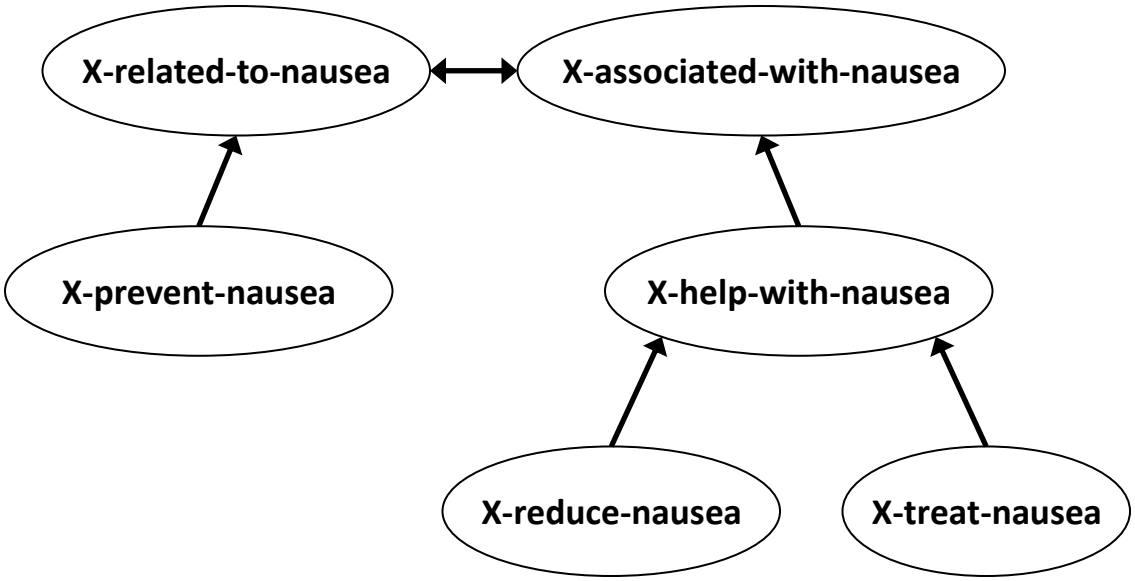
- Ambiguity:
  - Focused entailment graphs
  - Typed entailment graphs



- Scalability – problem is NP-hard:
  - ILP formulation
  - Scaling: *decomposition and incremental ILP*

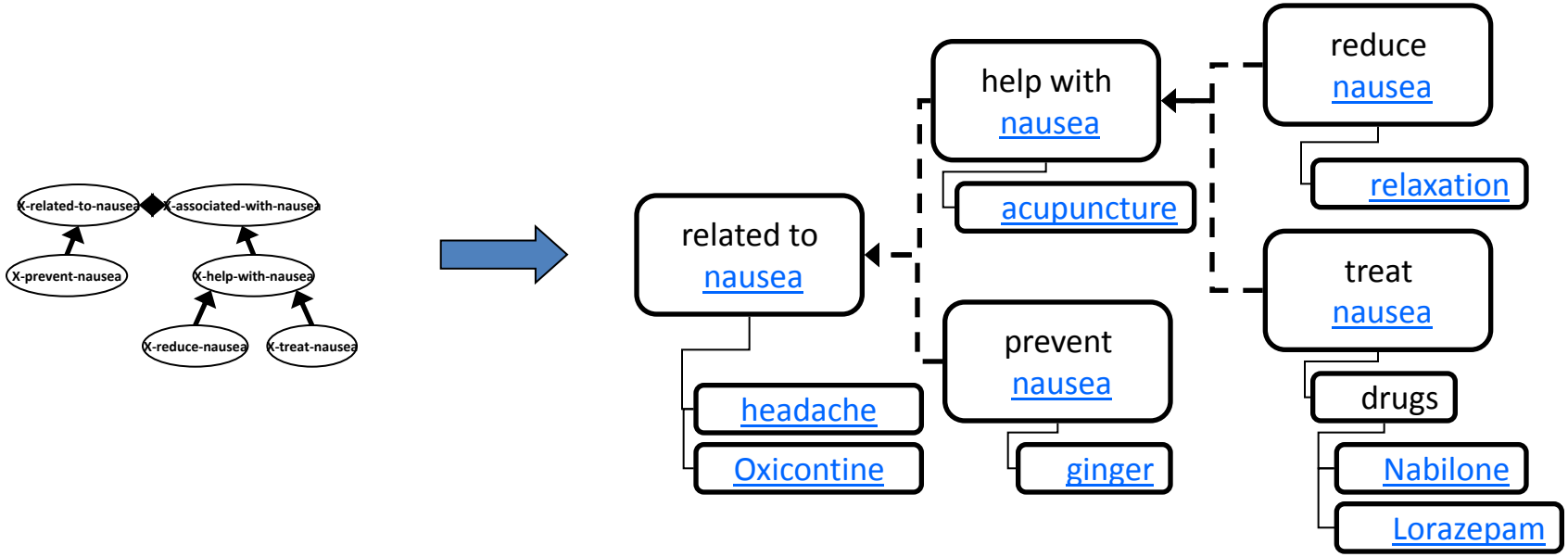
Focused Entailment Graphs

Focused Entailment Graphs



- Argument is instantiated by a target concept *(nausea)*
- Instantiating an argument reduces ambiguity and scalability problems

Motivation - Hierarchical Summarization



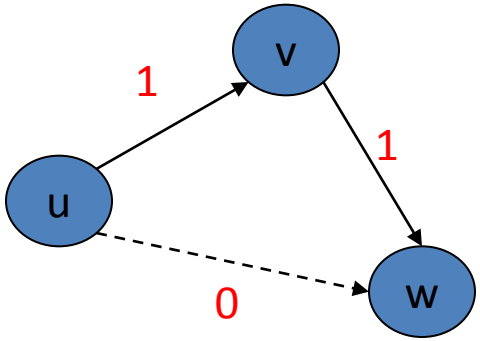
- Scenario: user queries about a concept *(nausea)*
- Summarize facts using a predicate entailment hierarchy interleaved with a concept taxonomies.

Global Learning of Edges

Input: Set of nodes V, weighting function $f: V \times V \rightarrow R$
Output: Set of directed edges E respecting transitivity that maximizes sum of weights

- Problem is NP-hard:
  - Reduction from "Transitive Subgraph" (Yannakakis, 1978)
- Integer Linear Programming (ILP) formulation:
  - Optimal solution (not approximation)
  - Often an LP relaxation will provide an integer solution

Integer Linear Program



$$\hat{G} = \arg\max \sum_{u \neq v} f(u,v) \cdot X_{uv}$$

$$\forall u,v,w \in V . X_{uv} + X_{vw} - X_{uw} \leq 1$$

$$\forall (u,v) \in NEG . X_{uv} = 0$$

$$\forall (u,v) \in POS . X_{uv} = 1$$

$$X_{uv} \in \{0,1\}$$

- Indicator variable X_{uv} for every pair of nodes
- Objective function maximizes sum of edge scores
- Transitivity and background-knowledge provide constraints

Objective Function

- Given a probabilistic classifier that estimates $P_{uv} = P(X_{uv}=1 \mid F_{uv})$ and some simplifying independence assumptions:

$$\hat{G} = \arg\max_{G} P(G \mid F)$$

$$= \arg\max_{G} \sum_{u \neq v} \log \frac{P_{uv} \cdot P(X_{uv}=1)}{(1-P_{uv}) \cdot P(X_{uv}=0)} \cdot X_{uv}$$

$$= \arg\max_{G} \sum_{u \neq v} \log \frac{P_{uv}}{(1-P_{uv})} \cdot X_{uv} + \lambda \cdot |E|$$

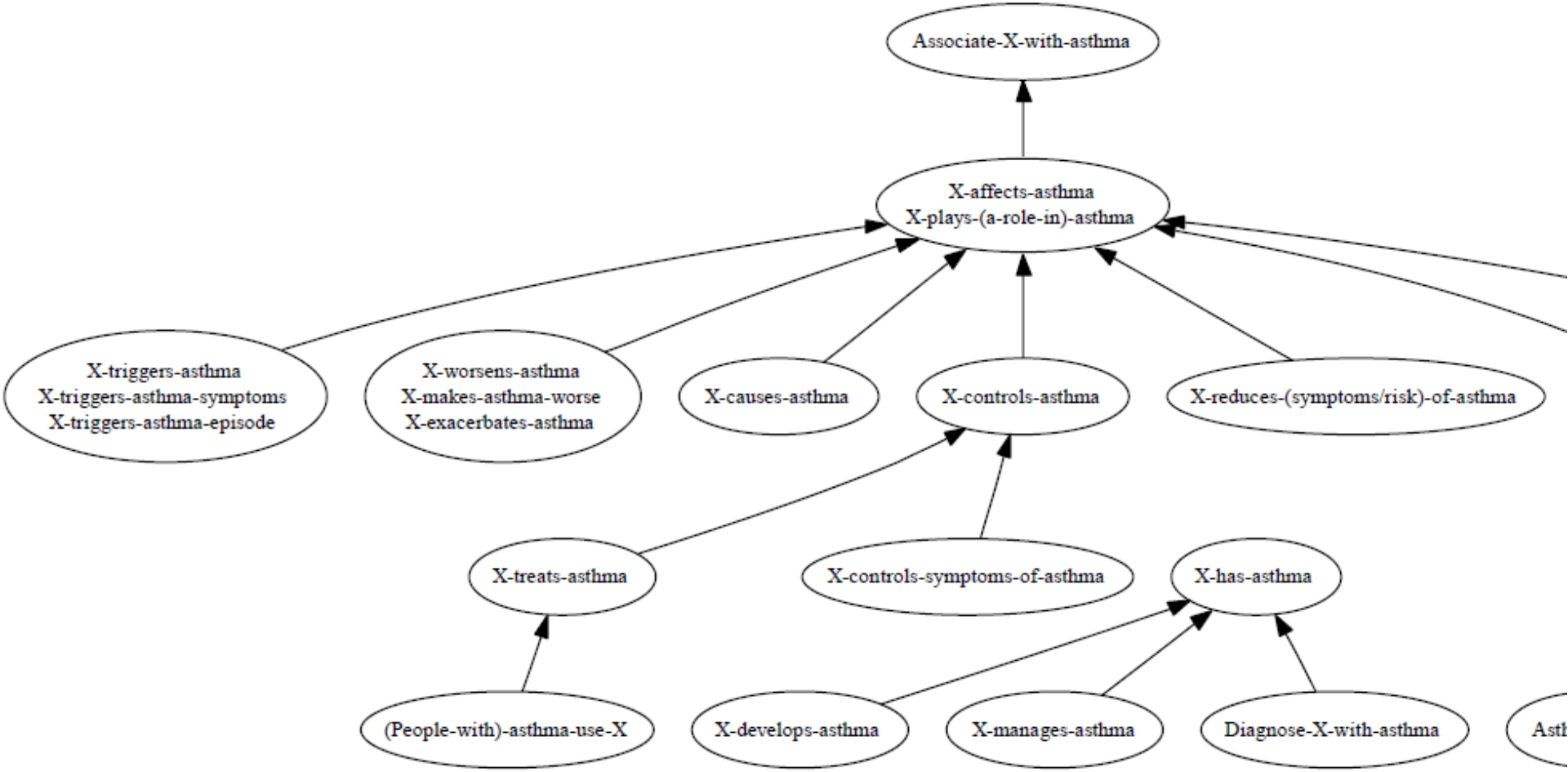
f(u,v)

prior

Experimental Evaluation

- 50 million word tokens **healthcare** corpus
- Ten medical students prepared gold standard graphs for 23 medical concepts:
  - Smoking, seizure, headache, lungs, diarrhea, chemotherapy, HPV, Salmonella, Asthma, etc.
- Evaluation:
  - F_1 on set of learned edges vs. gold standard
  - F_1 on set of learned facts vs. gold standard

Gold Standard Graph



Evaluated algorithms

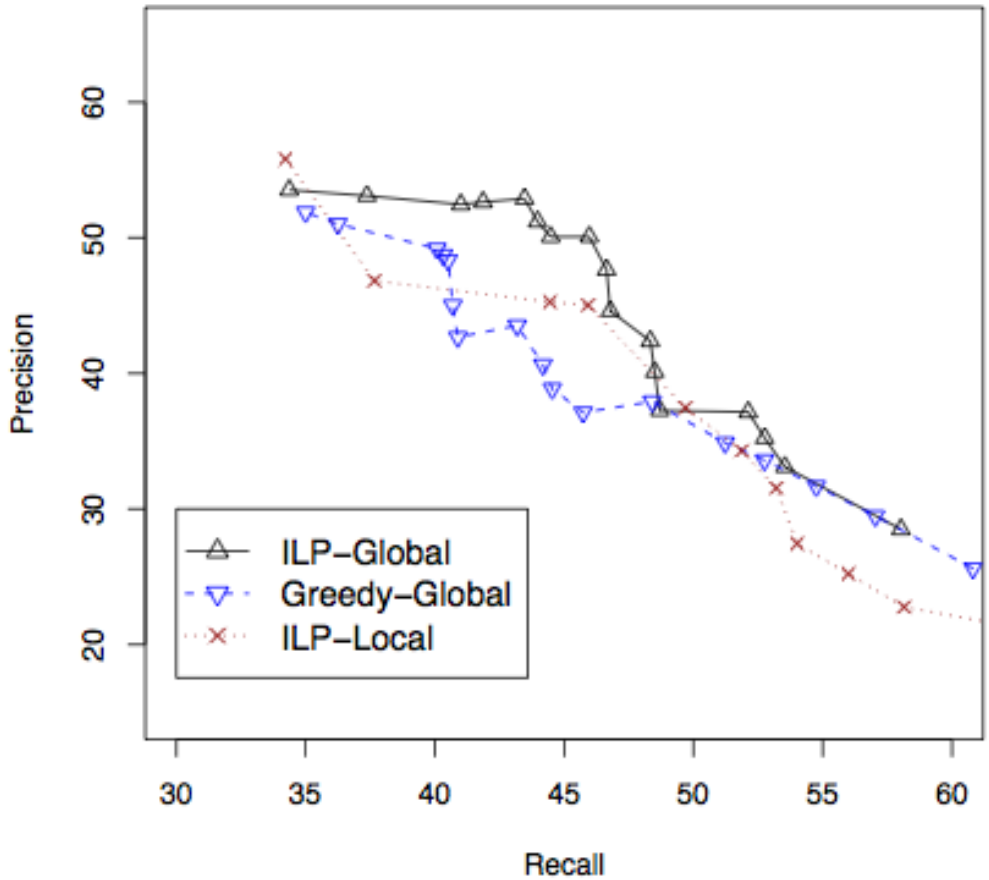
- Local algorithms
  - Distributional similarity (DIRT, BInc, etc.)
  - WordNet
  - ILP with No transitivity constraints
- Global algorithms
  - ILP/Snow et al. (greedy optimization)

Results

	Edges			Propositions		
	recall	Precision	F_1	recall	Precision	F_1
ILP-global	46.0	50.1	**43.8***	67.3	69.6	**66.2***
Greedy	45.7	37.1	36.6	64.2	57.2	56.3
ILP-local	44.5	45.3	38.1	65.2	61.0	58.6
$Local_1$	53.5	34.9	37.5	73.5	50.6	56.1
$Local_2$	52.5	31.6	37.7	69.8	50.0	57.1
$Local^*_1$	53.5	38.0	39.8	73.5	54.6	59.1
$Local^*_2$	52.5	32.1	38.1	69.8	50.6	57.4
WordNet	10.8	44.1	13.2	39.9	72.4	47.3

- The algorithm significantly outperforms all other baselines.

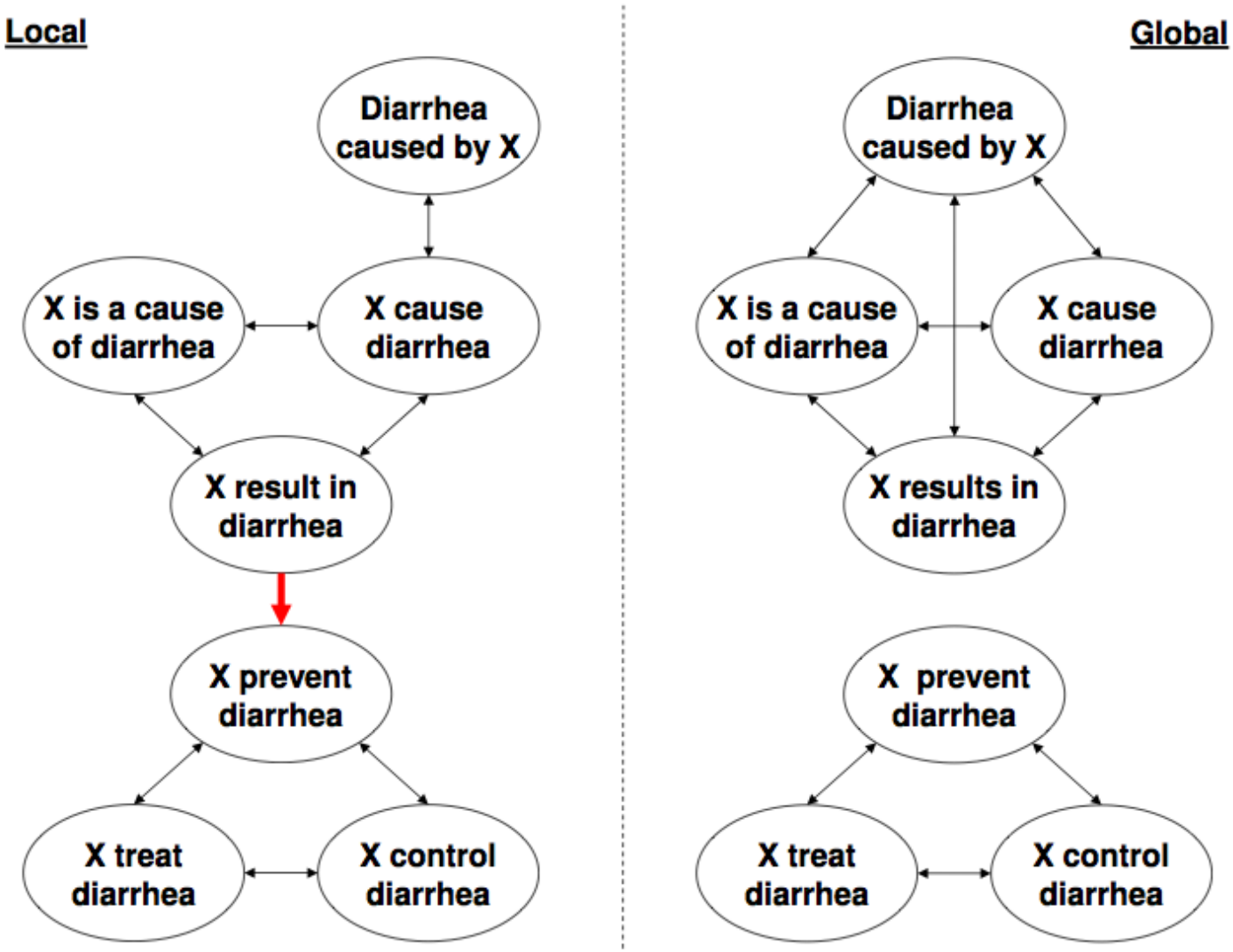
Precision-recall Curve



Results

	Global=true/Local=false	Global=false/Local=true
Gold standard = true	48	42
Gold standard = false	78	494

- Global algorithm avoids false positives.


Local
Diarrhea caused by X
X is a cause of diarrhea
X cause diarrhea
X result in diarrhea
X prevent diarrhea
X treat diarrhea
X control diarrhea
Global
Diarrhea caused by X
X is a cause of diarrhea
X cause diarrhea
X results in diarrhea
X prevent diarrhea
X treat diarrhea
X control diarrhea

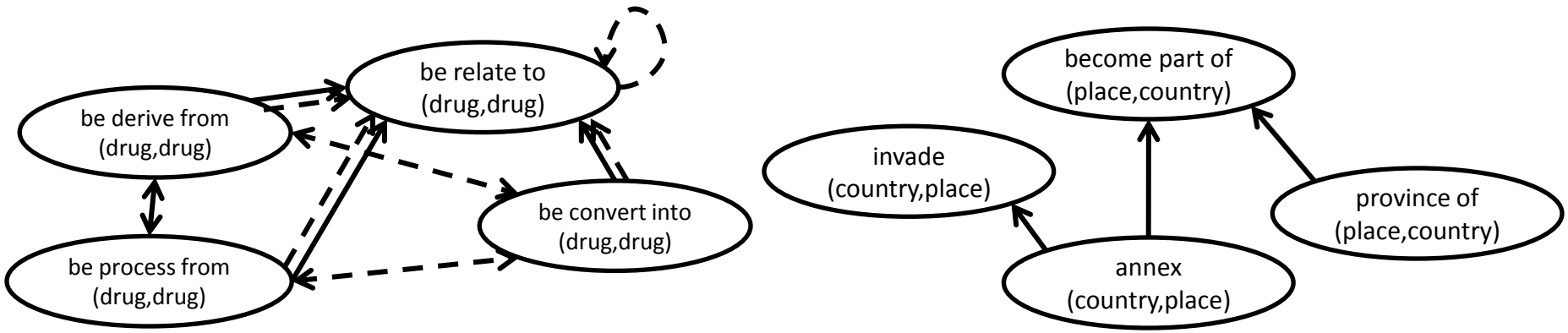
Typed Entailment Graphs

Typed Predicates

- Focused graphs learn low-applicability rules

- Predicate variables are typed:
  - *$X_{company}$ acquire $Y_{company}$ → $Y_{company}$ is sold to $X_{company}$*
- Rules of wide-applicability but less ambiguous
- Schoenmackers et al. (EMNLP 2010) used a <u>local</u> algorithm to learn 30,000 entailment rules
- We want to use a <u>global</u> algorithm

Typed Entailment Graphs



- A graph is defined for every pair of types
- "single-type" graphs contain "direct-mapping" edges and "transposed-mapping" edges
- Problems:
  - How to represent "single-type" graphs
  - Hard to solve graphs with >50 nodes

ILP for "single-type graphs"

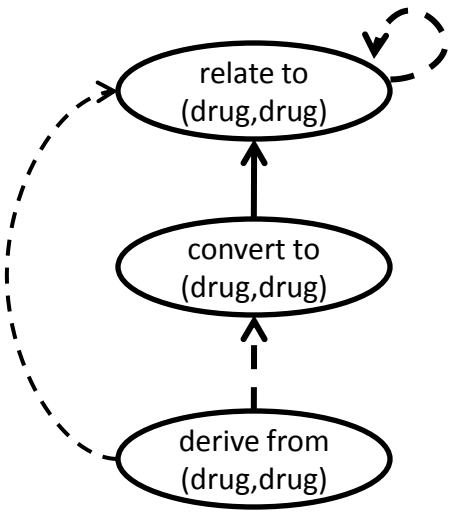
$$\hat{G} = \arg\max \sum_{u \neq v} f_x(u,v) \cdot X_{uv} + \sum_{u,v} f_y(u,v) \cdot Y_{uv}$$

$$\forall u,v,w \in V . X_{uv} + X_{vw} - X_{uw} \leq 1$$

$$\forall u,v,w \in V . X_{uv} + Y_{vw} - Y_{uw} \leq 1$$

$$\forall u,v,w \in V . Y_{uv} + X_{vw} - Y_{uw} \leq 1$$

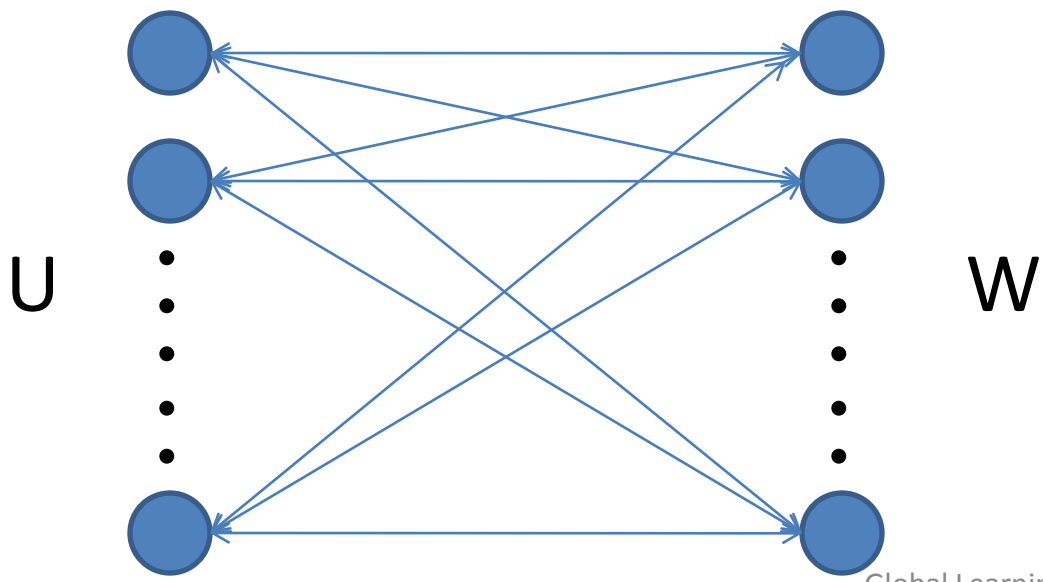
$$\forall u,v,w \in V . Y_{uv} + Y_{vw} - X_{uw} \leq 1$$



- The functions f_x and f_y provide local scores for direct and reversed mappings
- Cut the size of ILP in half comparing to naïve solution

Decomposition

- Sparsity: Most predicates do not entail one another
- Proposition: If we can partition the nodes to *U,W* such that *f(u,w)<0* for every *u,w* then any *(u,w)* is not an edge in the optimal solution



Decomposition Algorithm

<u>Input</u>: Nodes V and function $f{:}V\text{x}V \rightarrow R$

1. Insert undirected edges for any (u,v) such that $f(u,v)>0$
2. Find connected components $V_1,\ldots V_k$
3. For $i = 1\ldots k$

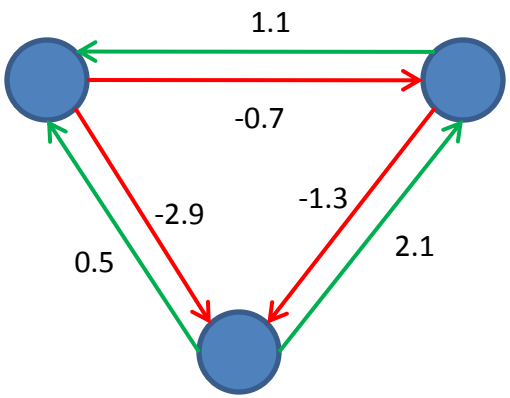
   $E_i = ILP\text{-}solve(V_i,f)$

<u>Output</u>: $E_1,\ldots,E_k$ guaranteed to be optimal

- Step 1 and 2 are efficient

Incremental ILP

- Given a good classifier most transitivity constraints are not violated



- Add constraints only if they are violated

Incremental ILP Algorithm

Input: Nodes V and function $f: V \times V \rightarrow R$

1. $ACT, VIO = \phi$
2. repeat
   a) $E = ILP\text{-}solve(V,f,ACT)$
   b) $VIO = violated(V,E)$ ← Needs to be efficient
   c) $ACT = ACT \cup VIO$
3. Until $|VIO| = 0$

Output: E guaranteed to be optimal

- Empirically converges in 6 iterations and reduces number of constraints from 10^6 to 10^3-10^4

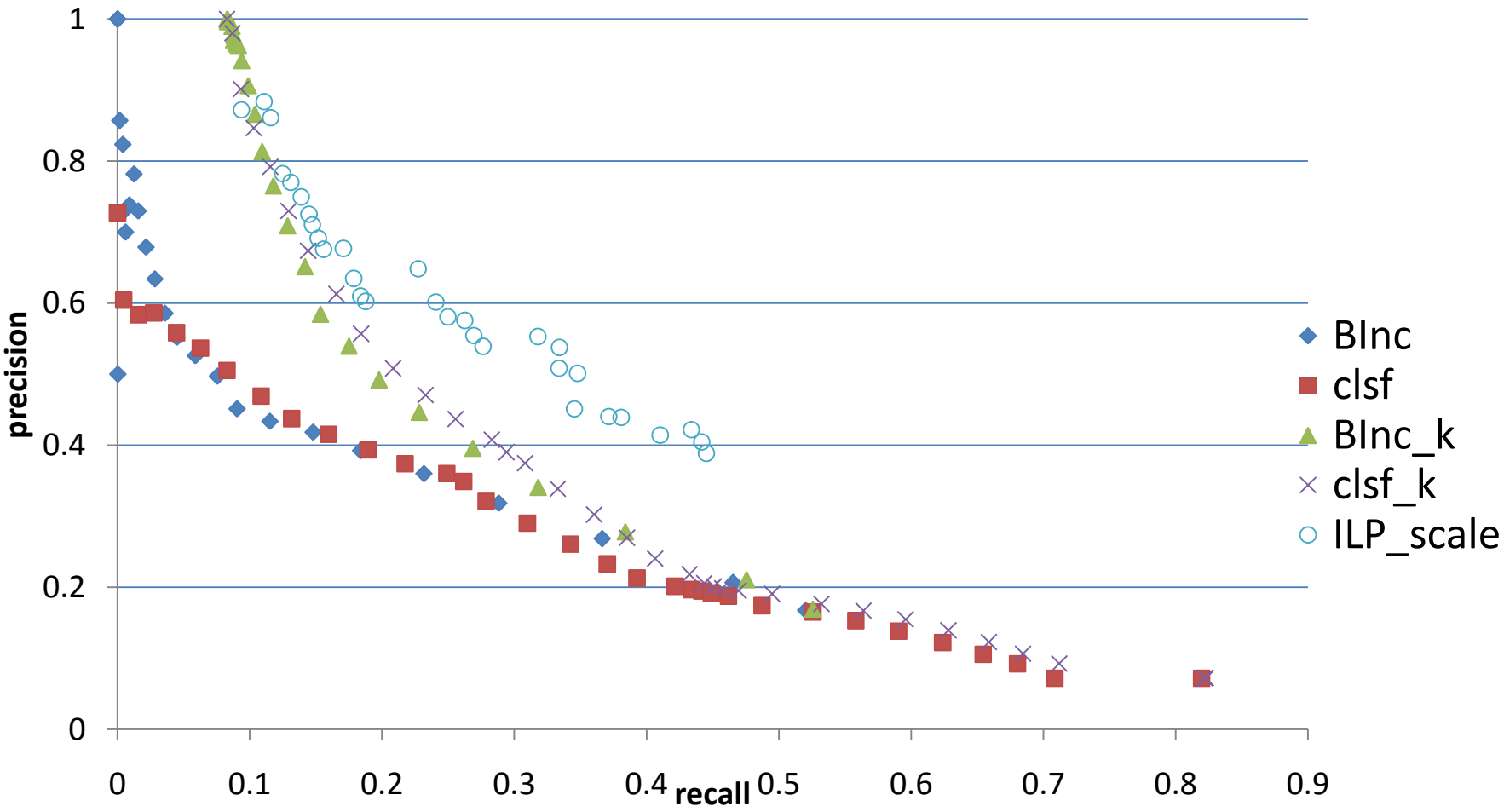
Experiment 1 - Transitivity

- 1 million TextRunner tuples over 10,672 typed predicates and 156 types
- Consist ~2,000 typed entailment graphs
- 10 gold standard graphs of sizes: 7, 14, 22, 30, 38, 53, 59, 62, 86 and 118
- Evaluation:
  - F_1 on set of edges vs. gold standard
  - Area Under the Curve (AUC)

Evaluated algorithms

- Local algorithms
  - Distributional similarity (DIRT, BInc, etc.)
  - Distributional similarity + background knowledge
  - ILP with No transitivity constraints
  - Sherlock rule resource
- Global algorithm: ILP

Precision-Recall Curve



Results

	Micro-average			AUC
	R	P	F_1	
ILP	**43.4**	42.2	**42.8**	**0.22**
Clsf	30.8	37.5	33.8	0.17
Sherlock	20.6	**43.3**	27.9	N/A
SR	38.4	23.2	28.9	0.14
DIRT	25.7	31.0	28.1	0.13
BINC	31.8	34.1	32.9	0.17

- R/P/F_1 at point of maximal micro-F_1
- Transitivity improves rule learning over typed predicates

Experiment 2 - Scalability

- Run ILP with and without *Decompose Incremental-ILP* over ~2,000 graphs
- Compare for various sparsity parameters:
  - Number of unlearned graphs
  - Number of learned rules

Results

Sparsity	# unlearned graphs	# learned rules	Δ(%)	Reduction (%)
-1.75	9/0	6,242/7,466	+20	75
-1	9/1	16,790/19,396	+16	29
-0.6	9/3	26,330/29,732	+13	14

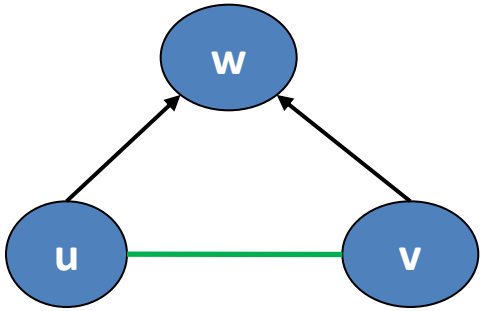
- Scaling techniques add 3,500 new rules to best configuration
- Corresponds to 13% increase in relative recall

Conclusions

- Algorithm for learning entailment rules given both local information and a global transitivity constraint
- ILP formulation for learning entailment rules
- Algorithms that scale ILP to larger graphs
- Application for hierarchical summarization of information for query concepts
- Resource of 30,000 domain-independent typed entailment rules

Future Work

- Untyped graphs
  - Ambiguity
  - Scalability
- Add types of edges.



- Improve entailment classifier

Thank you!