Coreference Resolution with ILP-based Weighted Abduction

Naoya Inoue[†], Ekaterina Ovchinnikova[‡], Kentaro Inui[†], Jerry Hobbs[‡]

[†]Tohoku University, Japan
[‡]ISI/USC, USA

Motivation

- **Long-term goal:** unified framework for discourse processing
- **Solution:** logical inference-based approach
  - World knowledge: set of logical formulae
  - Discourse processing: logical inference to logical forms (LFs) of target discourse
  - Interpretation as Abduction [Hobbs+ 93]

[Hobbs+ 93]

Interpretation as Abduction

- **Abduction:** inference to the best explanation to observation
- Interpreting sentences: finding best explanation to LFs of sentences
- Best explanation gives solution for broad range of NLP tasks
  – By-product of abductive inference

Abductive interpretation: example

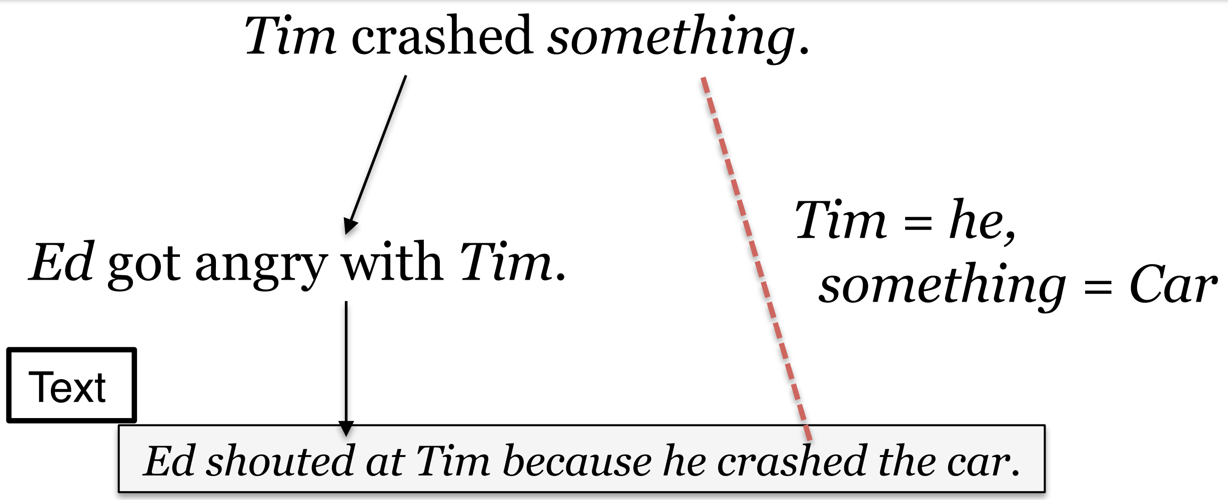
X crash Y → Z angry-with X
X angry-with Y → X shout-at Y

kn

Candidate explanations

	Ed got angry with Tim	Tim crashed something	something = car	Ed = he	Tim = he
1					
2	✓				
3	✓	✓			
					✓
					✓

Promising as unified framework for NLP
- ✓ Jointly solve several NLP tasks
- ✓ Make implicit information explicit in text



Attractive but...

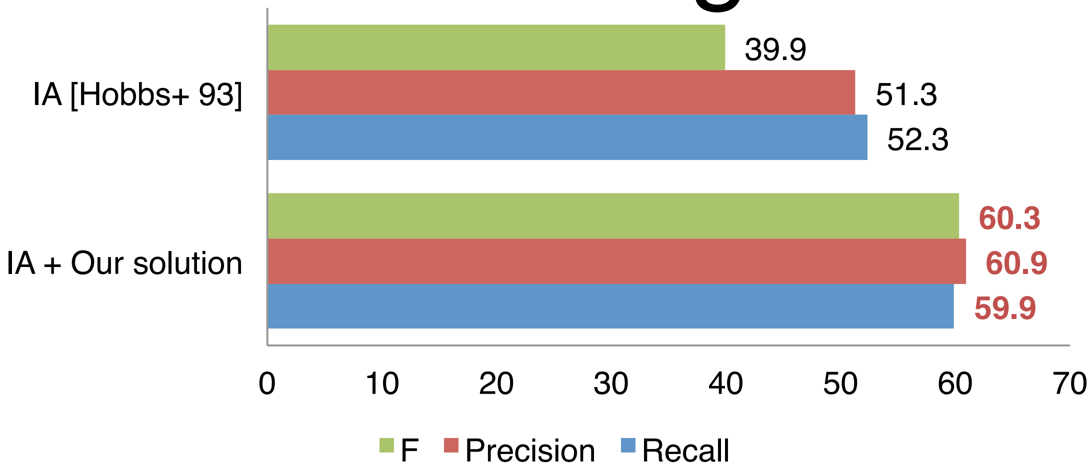
- Abductive discourse processing: attractive but still has many issues
  - How to perform efficient inference?
    - Best explanation finding: NP-hard
  - How to measure goodness of explanation?
    - Heuristic tuning: itractable on large BK
  - ... etc.

Our work

- **This talk:** address overmerging issue in abductive discourse processing
  - Finding least-cost explanation often produces wrong eq assumptions
    - Equality = Coreference
    - Critical issue in abductive discourse processing
  - Explore through coreference evaluation

Sneak preview (1/2)

- Successfully prohibit wrong merges
  – 28,233 wrong merges/33,775 merges (83.6%) → 7,489/11,001 (68.0%)
- Improve overmerging problem by 20 % BLANC-F over original IA



Sneak preview (2/2)

- Coreference study perspective: novel coreference model
  - Document-wise
  - Logical inference-based
  - Integrate statistical machine learning of logical inference with traditional clustering-based approach

Talk outline

- ✓ Introduction
- ☐ Key Idea
- ☐ Our system
- ☐ Evaluation
- ☐ Conclusion

[Hobbs+ 93]

Weighted Abduction (WA)

- **Input:** background knowledge (BK) B, observation O
  - B: set of first-order logical formulas (LFs)
  - O: set of first-order literals
- **Output:** least-cost explanation H of O w.r.t. B
  - H: set of first-order literals, such that:

$$\underset{\text{(BK)}}{B} \cup \underset{\text{(Explanation)}}{H} \models \underset{\text{(Observation)}}{O}$$

$$B \cup H \not\models \bot$$

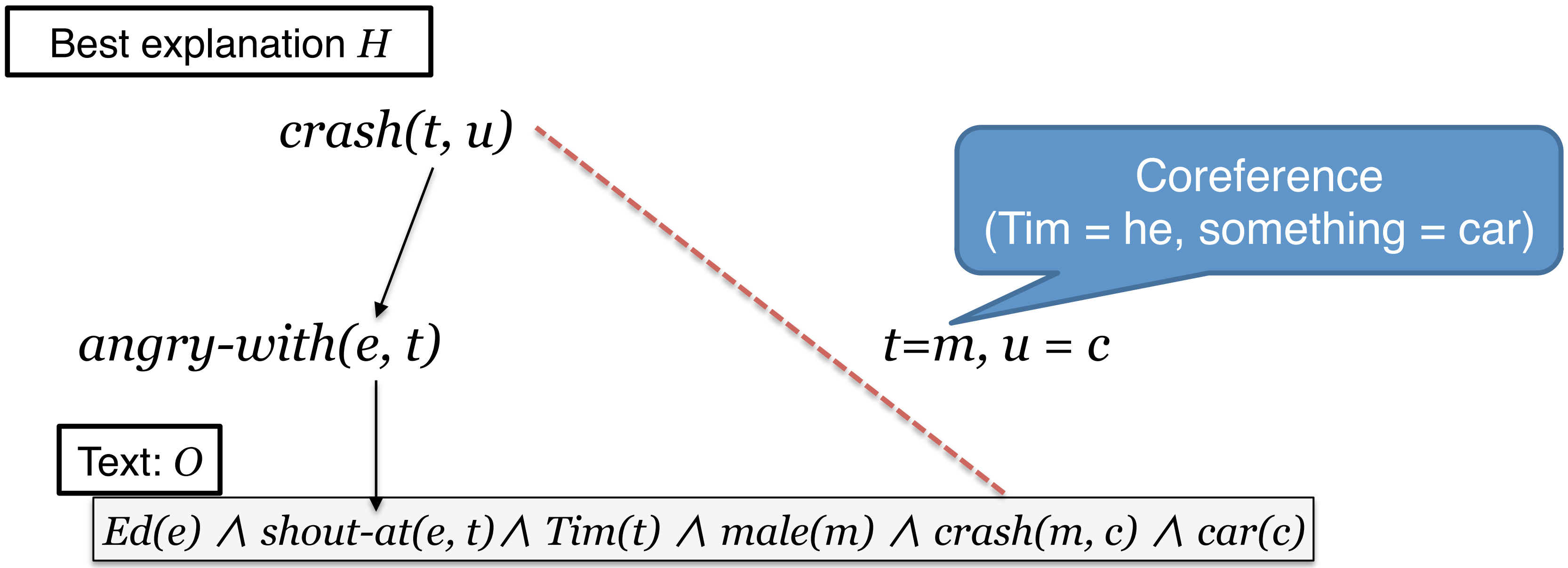
Abductive interpretation: example

crash(X, Y)
→ angry-with(Z, X)
angry-with(X, Y)
→ shout(X, Y)

World knowledge: *B*

Candidate explanations

	angry-with(e, t)	crashed(t, u)	u = c	e = m	t = m	cost(H)
1						30.0
2	✓					10.0
3	✓	✓				31.0
4	✓	✓	✓			53.0
5	✓	✓	✓	✓		12.0
6	✓	✓		✓	✓	33.0
7	✓	✓	✓		✓	5.0
:			:			:



Problem: overmerging

- Abduction: find least-cost explanation
  - Finding least-cost explanation ⇒ making equality assumptions as much as possible
  - Unification of two literals leads to minimal explanation
    - $H = \{p(x), p(y), p(z)\} \rightarrow H' = \{p(x), x{=}y{=}z\}$
- Frequently produces inconsistent explanation

Overmerging example

$run(y)$

$run(x)$ … …

$(\exists x, y) \ldots \wedge cat(x) \wedge dog(y) \wedge \ldots$

" … There are cat and dog. …"

Knowledge about disjointness:
$\forall x\ cat(x) \Rightarrow \neg dog(x)$

$(\exists x, y, e_1, e_2)\ cat(x) \wedge dog(x) \wedge run(x) \wedge x{=}y$

"A cat and dog run. Cat and dog refers to the same entity."

Problem: overmerging

- **Key problem:** knowledge about disjointness is not sufficient
  - Few knowledge acquisition study focus on disjointness knowledge
  - Assuming complete disjointness knowledge is not reasonable
    - Could be low coverage and/or noisy

Key idea: weighted unification

- **Solution:** cost for unification
  - Weighted abduction [Hobbs+ 93]: cost is not needed for unification
    - Unification always reduces cost
  - Modeled by weighted feature function
    - Features: disjointness knowledge base + linguistically-motivated features
    - Discriminative training of cost function from coreference-annotated dataset

Trainable cost function for weighted unification

Hypothesis: $run(x) \wedge run(y) \wedge x{=}y$

Hypothesis (or observation):	***cost(x=y)***	
	WA [Hobbs+ 93]	Our work
$cat(x) \wedge dog(y)$	0.0	
$cat(x) \wedge animal(y)$	0.0	

$$cost(x, y; \mathbf{w}) = \mathbf{w} \cdot \mathbf{\Phi}(x, y, H)$$

w: weight vector (trained)
Φ: feature vector (describing incompatibility, or compatibility)

Novelty

- Abduction perspective
  - First work to exploit learning-based approach for overmerging problem
    - [Ovchinnikova+ 11]: rule based
- Coreference resolution perspective
  - Latent clustering-based coreference resolution model
    - Latent variables: explanation of text
  - Exploit logical inference for coref resolution
    - [Poon & Domingos 08, Song+ 12]: Markov Logic-based, but not for reasoning

System overview

- **Preparation:**
  - Encode world knowledge as a set of logical formulae (= B)
- **Input:** text (one document)
  1) Generate LFs of text
  2) Perform weighted abduction, where:
     - **Observation:** LFs of text
     - **Background knowledge:** world knowledge (= B)
     - **Cost function:** [Hobbs+ 93] + weighted unification
  3) Build up coreference clusters from explanation
- **Output:** set of coreference clusters

1) Generate LFs

- Exploit off-the-shelf semantic parser, Boxer [Bos 09]

$(\exists\, e, t, m, c)\ Ed(e) \wedge shout\text{-}at(e, t) \wedge Tim(t) \wedge male(m) \wedge crash(m, c) \wedge car(c)$

Ed shouted at Tim because he crashed the car.

2) Abductive interpretation

Background knowledge:

$(\forall x, y)\ crash(x, y) \rightarrow (\exists z)\ angry\text{-}with(z, x)$

$(\forall x, y)\ angry\text{-}with(x, y) \rightarrow shout\text{-}at(x, y)$

$crash(t, u)$

$t{=}m,\ u{=}c$

$angry\text{-}with(e, t)$

$(\exists e, t, m, c)\ Ed(e) \wedge shout\text{-}at(e, t) \wedge Tim(t) \wedge male(m) \wedge crash(m, c) \wedge car(c)$

Ed shouted at Tim because he crashed the car.

Cost function (1/2)

$$cost(H;\mathbf{w}) = \underbrace{\sum_{h \in L(H)} cost(h)}_{\text{(a) [Hobbs+ 93]}}$$

- Two parts:
  - a) Costs of assumed literals [Hobbs+ 93]
    - Assumed literals: literals not explained
  - b) Costs of equality assumptions (our extension)
    - Cost: calculated by weighted linear feature function

Cost function (2/2)

$cost(H; \mathbf{w}) =$

$crash(t, u)^{\$14}$

pay pay

$t{=}m, u{=}c$

$angry\text{-}with(e, t)^{\$12}$

pay

$(\exists e, t, m, c)\ Ed(e)^{\$10} \wedge shout\text{-}at(e, t)^{\$10} \wedge Tim(t)^{\$10} \wedge male(m)^{\$10} \wedge crash(m, c)^{\$10}$

pay pay pay

Feature vector: $\boldsymbol{\Phi(x, y, H)}$

- WordNet-based features
  - Are x and y antonym? Are x and y siblings?
  - Are x and y proper names not belonging to the same synset?
- Lexico-syntactic patterns
  - Do x and y appear in explicit non-identity expressions?
    - e.g. x is different from y
  - Do x and y appear in functional predicates?
    - e.g. x is father of Ed. y is father of John.
  - Are x and y owned by same literal?
    - e.g. eat(x, y)

Weight vector $\mathbf{w}$: how to tune?

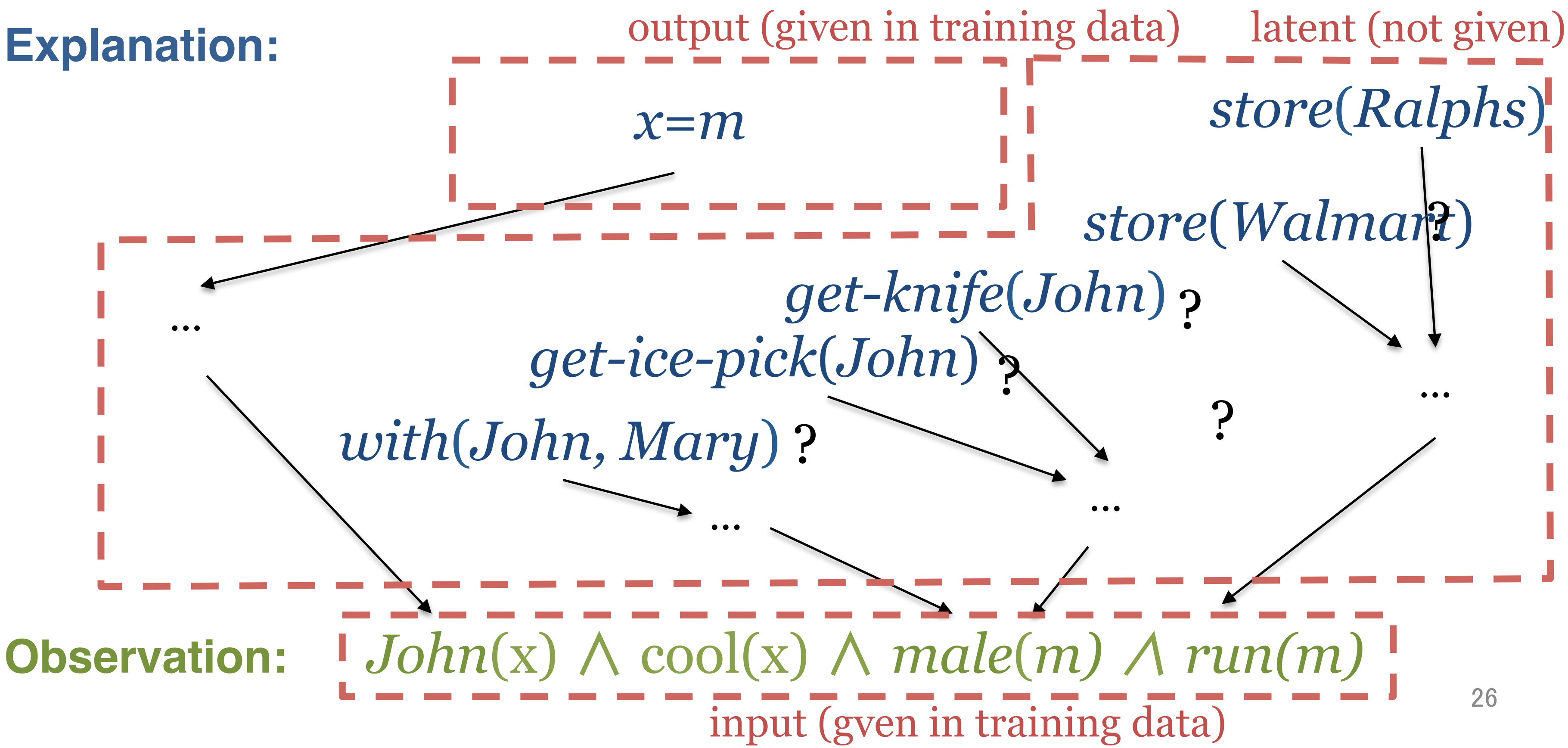
- Interpret the cost function as a latent coreference resolution model, where:
  - Output variables: coreference relations
  - Latent variables: explanations
- Apply document-wise supervised learning
  - Online large-margin training: Passive Aggressive (PA) algorithm [Crammer+ 06] modified for learning with latent variables
  - Training data:
    - (Input: LFs of text, Output: equality assumptions describing coreference relations)
      - e.g. $(John(x) \wedge cool(x) \wedge male(m) \wedge run(m), x=m)$

Modified PA

- At high level: EM-like training
  - Repeat the following steps:
  - 1. Given observed states, estimate most probable states of unobserved (latent) variables with current weights
    - Observed: equality assumptions
    - Unobserved (latent): explanation
  - 2. Update weight vector as if all the states are fully observed
    - Large-margin update [Crammer+ 06]
    - All the states = best explanation

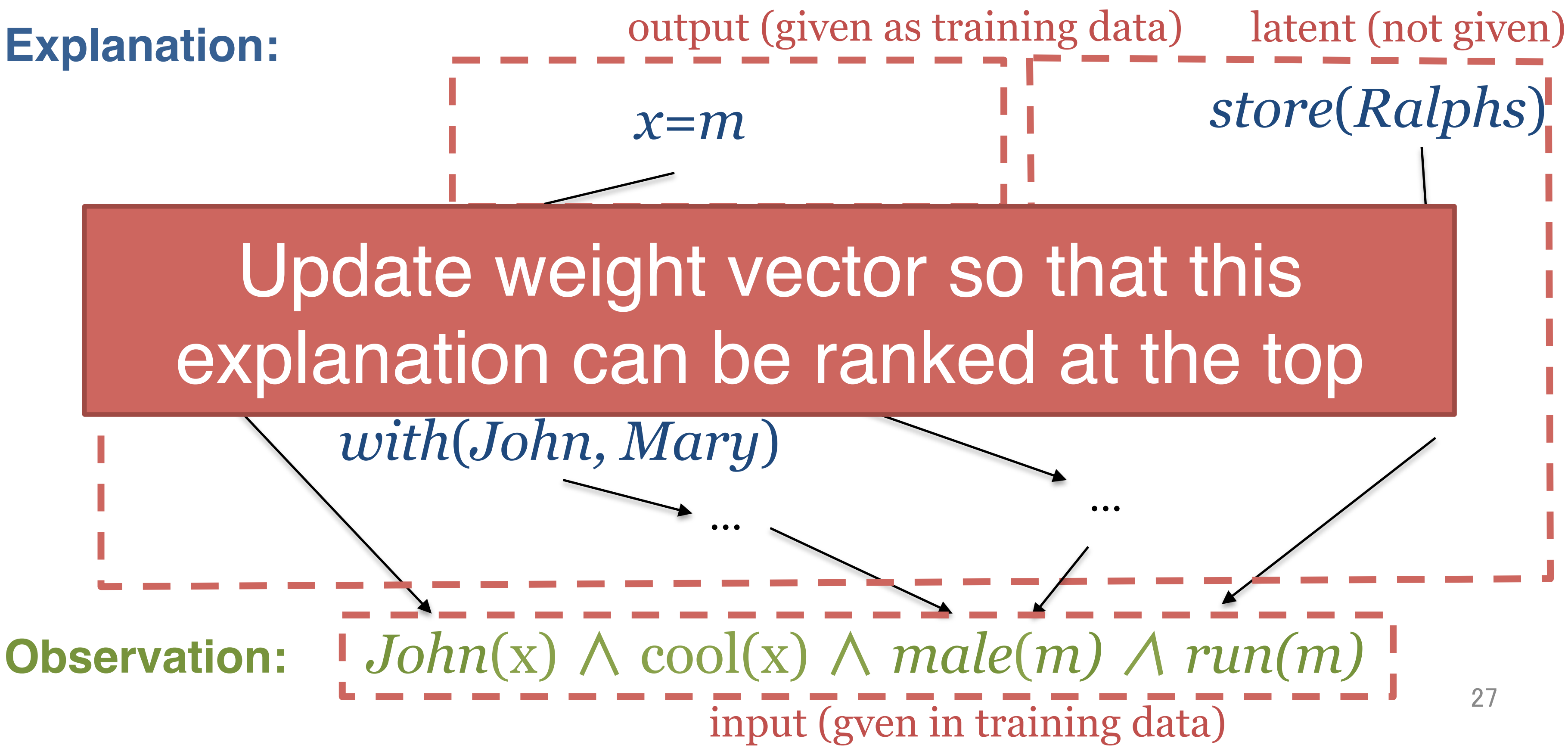
Example update

- Estimate most probable explanations consistent with gold equality assumptions



Example update

- Estimate most probable explanations consistent with gold equality assumptions



Inference

- Least-cost finding problem: NP hard
- Extend state-of-the-art ILP-based abductive reasoner [Inoue & Inui 12]
  – Lifted inference: directly perform abduction on first-order level
  – Use Integer Linear Programming technique for efficient search

3) Identify coreference clusters

Cluster: {***Tim***, ***he***}

$crash(t, u)$

$angry\text{-}with(e, t)$

$\boxed{t=m,\ u=c}$

$(\exists\, e, t, m, c)\ Ed(e) \wedge shout\text{-}at(e, t) \wedge \boxed{Tim(t)} \wedge \boxed{male(m)} \wedge crash(m, c) \wedge car(c)$

Ed shouted at Tim because he crashed the car.

Talk outline

✓ Introduction
✓ Key Idea
✓ Our system
☐ Evaluation
☐ Conclusion

Evaluation

- Dataset
  - CoNLL 2011 SharedTask [Pradhan+ 11]
    - Test: 101 documents from dev set
    - Training: 100 documents from training set
  - Background knowledge:
    - WordNet, FrameNet, Narrative Chains
- Evaluation criteria
  - Overmerging Rate, BLANC metrics [Recasens & Hovy 10]
    - Other criteron: not suitable for exploring overmerging issues

Background knowledge (1/2)

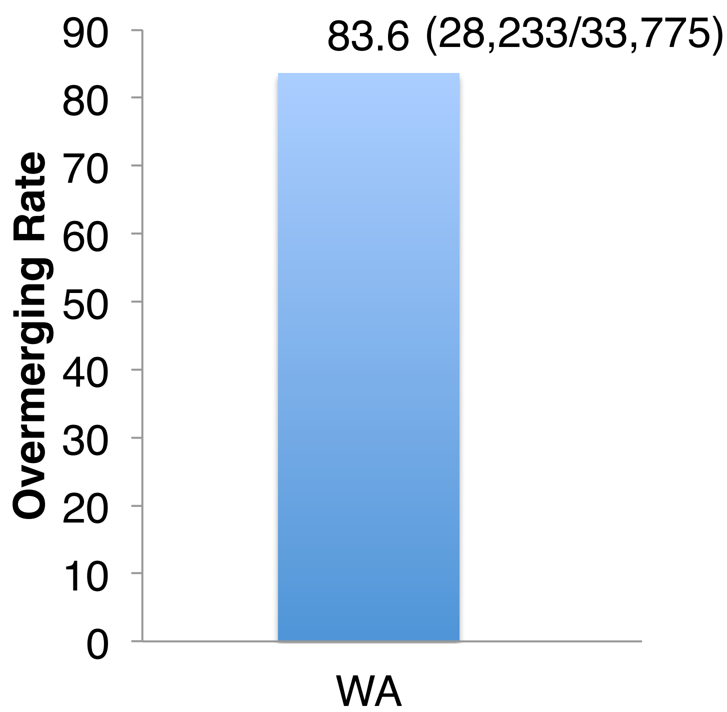
- WordNet [Fellbaum 98]: 22,815 axioms
  - Hyperonymy, Causation, Entailment, Meronymy, Membership
    - $(\forall x)\ synset1(x) \rightarrow synset2(x)$
- FrameNet [Ruppenhofer+ 10]: 12,060 axioms
  - Frame-lexeme mappings
    - e.g. $(\forall e_1, e_2, x_1, x_2, x_3)$ GIVING$(e_1) \wedge$ DONOR$(e_1, x_1) \wedge$ RECIPIENT$(e_1, x_2) \wedge$ THEME$(e_1, x_3) \rightarrow$ give$(e_1, x_1, x_3) \wedge$ to(e_2, e_1, x_2)
  - Frame-frame relations
    - e.g. GIVING causes GETTING

Background knowledge (2/2)

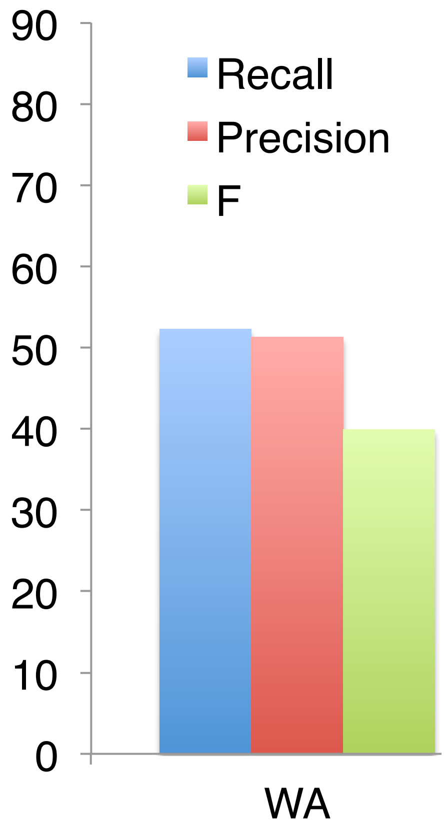
- Narrative chains [Chambers and Jurafsky 09]: 1,391,540 axioms
  - Partially ordered set of events in temporal order, with slot realizations
  - Verb-script mappings
    - e.g. $(\forall s, e_1, x_1, x_2, x_3)\ \mathrm{Script\#1}(s, e_1, x_1, x_2, x_3) \rightarrow \mathrm{arrest}(e_1, x_1, x_2, x_3) \land \mathrm{police}(e_2, x_1)$
- AIDA tool [Yosef+ 2011]
  - Normalization of proper names
    - e.g. "A. Einstein", "Einstein, Albert" → "Albert_Einstein"

Impact of our extension: Overmerging Rate

$$\text{Overmerging Rate (\%)} = \frac{\text{\# of wrong merges}}{\text{\# of merges}}$$



Impact of our extension: BLANC metrics



Why is it not comparable?

- Cannot capture deeper contradiction: more features are needed
  - Example deeper contradiction:
    - goods made in Japan, German goods
      $goods(x) \wedge make(e, u, x) \wedge in(e, Japan)$
      $goods(y) \wedge german(y)$
  - Solution: exploit syntactic clues, discourse saliency, distributional similarity etc.
- Low recall: more world knowledge is needed
  - e.g. YAGO, freebase, ConceptNet 5.0
- But has many interesting theoretical aspects, and highly extensible

Summary

- Address overmerging problem in abduction-based discourse processing
  - Extend Hobbs+ [93]'s cost function: add cost function for equality assumptions
    - Cost function is weighted feature function
  - Propose automatic tuning method of weights on coreference-annotated corpus
- Improvement by 20% BLANC-F over original weighed abduction

Future work

- Apply learning procedure to costs of assumed literals
  - Generalize cost function as weighted linear model, apply large-margin training
- Scale up reasoning process
  - Cutting plane-based MLNs [Riedel 08]
- Incorporate more features, and world knowledge for increasing both precision & recall

THANK YOU FOR YOUR ATTENTION!