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]
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
Ed got angry with Tim.
Tim crashed something.

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

Tim = he, something = Car

Ed shouted at Tim because he crashed the car.
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: intractable 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
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:

\[
B \cup H \models O
\]
Abductive interpretation: example

World knowledge: $B$

- $\text{crash}(X, Y)$
- $\Rightarrow \text{angry-with}(Z, X)$
- $\text{angry-with}(X, Y)$
- $\Rightarrow \text{shout}(X, Y)$

Text: $O$

Candidate explanations

<table>
<thead>
<tr>
<th></th>
<th>angry-with(e, t)</th>
<th>crashed(t, u)</th>
<th>$u = c$</th>
<th>$e = m$</th>
<th>$t = m$</th>
<th>cost(H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30.0</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10.0</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>31.0</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>53.0</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>12.0</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>33.0</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Best explanation $H$

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

Coreference: $(\text{Tim} = \text{he}, \text{something} = \text{car})$

$t = m, u = c$

Ed(e) $\land$ shout-at(e, t) $\land$ Tim(t) $\land$ male(m) $\land$ crash(m, c) $\land$ car(c)
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

(∃x, y) ... ∧ cat(x) ∧ dog(y) ∧ ...

“... There are cat and dog. ...”

Knowledge about disjointness:
∀x cat(x) ⇒ ¬dog(x)

(∃x, y, e₁, e₂) cat(x) ∧ dog(x) ∧ run(x) ∧ 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: \( \text{run}(x) \land \text{run}(y) \land x=y \)

<table>
<thead>
<tr>
<th>Hypothesis (or observation):</th>
<th>cost( (x=y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{cat}(x) \land \text{dog}(y) )</td>
<td>0.0</td>
</tr>
<tr>
<td>( \text{cat}(x) \land \text{animal}(y) )</td>
<td>0.0</td>
</tr>
</tbody>
</table>

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

- \( \mathbf{w} \): weight vector (trained)
- \( \Phi \): 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]

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

*Ed shouted at Tim because he crashed the car.*
2) Abductive interpretation

Background knowledge:
(\(\forall x, y\) crash(x, y) -> (\(\exists z\) angry-with(z, x))
(\(\forall x, y\) angry-with(x, y) -> shout-at(x, y))

(\(\exists e, t, m, c\) Ed(e) ^ shout-at(e, t) ^ Tim(t) ^ male(m) ^ crash(m, c) ^ car(c)

Ed shouted at Tim because he crashed the car.
Cost function (1/2)

\[ \text{cost}(H; w) = \sum_{h \in L(H)} \text{cost}(h) \]

(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)

\[
cost(H; w) =
\]

\[
\sum e, t, m, c \left( Ed(e)^{10} \wedge shout-at(e, t)^{10} \wedge Tim(t)^{10} \wedge male(m)^{10} \wedge crash(m, c)^{10} \right)
\]

\[
+ crash(t, u)^{14}
\]

\[
+ angry-with(e, t)^{12}
\]

\[
+ w_\varrho \Phi(t, m, H)
\]

\[
+ w_\varrho \Phi(u, c, H)
\]

\[
t=m, u=c
\]

\[
pay \quad pay
\]

Our extension
Feature vector: $\Phi(x, y, H)$

- **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. $\text{eat}(x, y)$
Weight vector \( w \): 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. \((\text{John}(x) \land \text{cool}(x) \land \text{male}(m) \land \text{run}(m), x=m)\)
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

Observation: $\text{John}(x) \land \text{cool}(x) \land \text{male}(m) \land \text{run}(m)$

input (given in training data)

Explanation:

$x=m$

output (given in training data)

latent (not given)

$\text{get-knife}(\text{John})$?

$\text{get-ice-pick}(\text{John})$?

$\text{with}(\text{John}, \text{Mary})$?
Example update

• Estimate most probable explanations consistent with gold equality assumptions

Observation: $John(x) \land cool(x) \land male(m) \land run(m)$

Explanation: $x=m$

output (given as training data)

latent (not given)

Update weight vector so that this explanation can be ranked at the top

store(Ralphs)

with(John, Mary)
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: \(\{\text{Tim, he}\}\)

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

\textit{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 criterion: not suitable for exploring overmerging issues
Background knowledge (1/2)

- **WordNet [Fellbaum 98]:** 22,815 axioms
  - Hyperonymy, Causation, Entailment, Meronymy, Membership
    - $(\forall x) \text{synset}_1(x) \rightarrow \text{synset}_2(x)$

- **FrameNet [Ruppenhofer+ 10]:** 12,060 axioms
  - Frame-lexeme mappings
    - e.g. $(\forall e_1, e_2, x_1, x_2, x_3) \text{GIVING}(e_1) \land \text{DONOR}(e_1, x_1) \land \text{RECIPIENT}(e_1, x_2) \land \text{THEME}(e_1, x_3) \rightarrow \text{give}(e_1, x_1, x_3) \land \text{to}(e_2, e_1, x_2)$
  - 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. \((\forall s, e_1, x_1, x_2, x_3) \text{Script#1}(s, e_1, x_1, x_2, x_3) \rightarrow \text{arrest}(e_1, x_1, x_2, x_3) \land \text{police}(e_2, x_1)\)

• AIDA tool [Yosef+ 2011]
  – Normalization of proper names
    • e.g. “A. Einstein”, “Einstein, Albert”
      \rightarrow “Albert_Einstein”
Impact of our extension: Overmerging Rate

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

Impact of our extension: Overmerging Rate

Overmerging Rate (%) = \( \frac{28,233}{33,775} = 83.6 \) \( \frac{7,489}{11,001} \)
Impact of our extension: BLANC metrics

- Recall
- Precision
- F

Impact by 20% BLANC-F

Still, not comparable to state-of-the-art resolver
Why is it not comparable?

- Cannot capture deeper contradiction: more features are needed
  - Example deeper contradiction:
    - \( \text{goods} \) made in Japan, German \( \text{goods} \)
    - \( \text{goods}(x) \land \text{make}(e, u, x) \land \text{in}(e, \text{Japan}) \land \text{goods}(y) \land \text{german}(y) \)
    - 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!