Annotating Anaphoric Shell Nouns with their Antecedents

Varada Kolhatkar and Graeme Hirst
University of Toronto

Heike Zinsmeister
Universität Stuttgart
Anaphoric Shell Nouns (ASNs)

New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
Why ASN annotation?
Ubiquity of Shell Nouns

- Occur frequently in all kinds of texts
- *fact, idea, problem*: among 100 most frequently occurring nouns (Schmid 2000)
- ~25 million occurrences in the NYT corpus (~1.3 billion tokens)
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...

• **Characterize** and **label** chunks of information
• **Cohesive devices**
• **Topic boundary markers**
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...

- **Characterize** and **label** chunks of information
- **Cohesive devices**
- **Topic boundary markers**
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...

- **Characterize** and **label** chunks of information
- **Cohesive devices**
- **Topic boundary markers**
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...

- **Characterize** and **label** chunks of information
- **Cohesive devices**
- **Topic boundary markers**
## Current Research: Gap

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ASN instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poesio and Artstein, 2008</td>
<td>455 abstract anaphor instances, very few ASNs</td>
</tr>
<tr>
<td>ARRAU</td>
<td></td>
</tr>
<tr>
<td>Kolhatkar and Hirst, 2012</td>
<td>188 <em>this issue</em> instances from Medline abstracts</td>
</tr>
<tr>
<td>Botley, 2006</td>
<td>462 ASN instances <em>(not available)</em></td>
</tr>
</tbody>
</table>
Current Research: Gap

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ASN instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poesio and Artstein, 2008 ARRAU</td>
<td>455 abstract anaphor instances, very few ASNs</td>
</tr>
<tr>
<td>Kolhatkar and Hirst, 2012</td>
<td>188 <em>this issue</em> instances from Medline abstracts</td>
</tr>
<tr>
<td>Botley, 2006</td>
<td>462 ASN instances <em>(not available)</em></td>
</tr>
</tbody>
</table>

ASNs are largely ignored in CL

Need a large-scale ASN antecedent corpus
Annotation challenges
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...

- Antecedents: complex and abstract entities
- Heterogeneous set of markables (e.g., NPs, VPs, sentences, clauses, ...)

What to Annotate?
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...

• Antecedents: complex and abstract entities

• Heterogeneous set of markables (e.g., NPs, VPs, sentences, clauses, ...)

Leads to large search space
What’s the “right” answer?

New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
Annotation data
# Shell Nouns: Categorization

*(Schmid, 2000)*

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual</td>
<td><em>fact, problem, reason</em></td>
</tr>
<tr>
<td>Linguistic</td>
<td><em>news, proposal, question</em></td>
</tr>
<tr>
<td>Mental</td>
<td><em>idea, belief, decision, issue</em></td>
</tr>
<tr>
<td>Modal</td>
<td><em>possibility, need, trend</em></td>
</tr>
<tr>
<td>Eventive</td>
<td><em>act, reaction, attempt</em></td>
</tr>
<tr>
<td>Circumstantial</td>
<td><em>situation, approach</em></td>
</tr>
</tbody>
</table>
### Shell Nouns: Selection

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual</td>
<td><em>fact, problem, reason</em></td>
</tr>
<tr>
<td>Linguistic</td>
<td><em>news, proposal, question</em></td>
</tr>
<tr>
<td>Mental</td>
<td><em>idea, belief, decision, issue</em></td>
</tr>
<tr>
<td>Modal</td>
<td><em>possibility, need, trend</em></td>
</tr>
<tr>
<td>Eventive</td>
<td><em>act, reaction, attempt</em></td>
</tr>
<tr>
<td>Circumstantial</td>
<td><em>situation, approach</em></td>
</tr>
</tbody>
</table>
# Shell Nouns: Selection

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual</td>
<td>fact, problem, reason</td>
</tr>
<tr>
<td>Linguistic</td>
<td>news, proposal, question</td>
</tr>
<tr>
<td>Mental</td>
<td>idea, belief, decision, issue</td>
</tr>
<tr>
<td>Modal</td>
<td>possibility, need, trend</td>
</tr>
<tr>
<td>Eventive</td>
<td>act, reaction, attempt</td>
</tr>
<tr>
<td>Circumstantial</td>
<td>situation, approach</td>
</tr>
</tbody>
</table>

High-frequency nouns from each category
The ASN Corpus

- ~475 instances per 6 selected shell nouns
  fact, reason, issue, decision, question, possibility
- Total: 2,822 ASN instances
Annotation methodology
Annotation Platform

- Crowdsourcing: CrowdFlower
- Quality control
- Gold questions
- Training phase
- Detailed results
  - Aggregated and full results with annotators’ demographic information
## CrowdFlower Confidence

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Trust</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.75</td>
<td>“a₁”</td>
</tr>
<tr>
<td>B</td>
<td>0.75</td>
<td>“a₂”</td>
</tr>
<tr>
<td>C</td>
<td>1.0</td>
<td>“a₂”</td>
</tr>
</tbody>
</table>
### CrowdFlower Confidence

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Trust</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.75</td>
<td>“a₁”</td>
</tr>
<tr>
<td>B</td>
<td>0.75</td>
<td>“a₂”</td>
</tr>
<tr>
<td>C</td>
<td>1.0</td>
<td>“a₂”</td>
</tr>
</tbody>
</table>

Score for “a₂” = 1.75
# CrowdFlower Confidence

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Trust</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.75</td>
<td>“a₁”</td>
</tr>
<tr>
<td>B</td>
<td>0.75</td>
<td>“a₂”</td>
</tr>
<tr>
<td>C</td>
<td>1.0</td>
<td>“a₂”</td>
</tr>
</tbody>
</table>

Crowd’s answer:
“a₂” with confidence $0.7 = \frac{1.75}{1.75 + 0.75}$
Annotation Tasks

ASN instances from the NYT

CrowdFlower Expt. 1

Identify the sentence containing antecedent

CrowdFlower Expt. 2

Identify the precise antecedent

Annotated ASN Corpus

Simple tasks do best with crowdsourcing
(Madnani et al. 2010, Wang et al. 2012)
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
Settings

- 2,822 instances
- 8 judgements per instance
- 8 cents per annotation
- Completion time: 3 days
Inter-annotator agreement for expt. 1
Confidence Distribution

- fact
- reason
- question
- issue
- decision
- possibility

Number of instances

Confidence
Confidence Distribution

Mean = 0.83

Mean = 0.82

Mean = 0.80

Mean = 0.83
Confidence Distribution

**fact**

- Mean = 0.83

**reason**

- Mean = 0.82

**question**

- Mean = 0.80

**issue**

- Mean = 0.61

**decision**

- Mean = 0.72

**possibility**

- Mean = 0.83
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings. Some lawmakers worry that cameras might compromise the rights of the litigants. But a 10-year experiment with courtroom cameras showed that televised access enhanced public understanding of the judicial system without harming the legal process. New York's backwardness on this issue hurts public confidence in the judiciary...
New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings.

Select one of the options (required)

- None
- one of only three states
- some form of audio-visual coverage
- some form of audio-visual coverage of court proceedings
- that do not allow some form of audio-visual coverage of court proceedings
- audio-visual coverage of court proceedings
- some form
- New York is one of only three states that do not allow some form of audio-visual coverage of court proceedings.
- only three states that do not allow some form of audio-visual coverage of court proceedings
- one of only three states that do not allow some form of audio-visual coverage of court proceedings
- allow some form of audio-visual coverage of court proceedings
Settings

- 2,323 high-confidence instances from CrowdFlower expt. 1
- 8 judgements per instance
- 6 cents per annotation
- Completion time: 7 days
Inter-annotator agreement for expt. 2
Challenge

It is believed that between 20 percent and 30 percent of Italy's economic output is submerged. The decision to finally take this fact into account...

**Answer 1:** that between 20 percent and 30 percent of Italy's economic output is submerged

**Answer 2:** between 20 percent and 30 percent of Italy's economic output is submerged

**Answer 3:** between 20 percent and 30 percent

Need coefficients that incorporate the notion of the distance between strings
Challenge

It is believed that between 20 percent and 30 percent of Italy's economic output is submerged. The decision to finally take this fact into account...

**Answer 1:** that between 20 percent and 30 percent of Italy's economic output is submerged

**Answer 2:** between 20 percent and 30 percent of Italy's economic output is submerged

**Answer 3:** between 20 percent and 30 percent

Need coefficients that incorporate the notion of the distance between strings
Challenge

It is believed that between 20 percent and 30 percent of Italy's economic output is submerged. The decision to finally take this fact into account...

**Answer 1:** that between 20 percent and 30 percent of Italy's economic output is submerged

**Answer 2:** between 20 percent and 30 percent of Italy's economic output is submerged

**Answer 3:** between 20 percent and 30 percent

Need coefficients that incorporate the notion of the distance between strings
Krippendorff’s $\alpha$

$$\alpha = 1 - \frac{D_o}{D_e}$$

- $\alpha = 1$, perfect reliability
- $\alpha = 0$, absence of reliability
- $\alpha < 0$, systematic disagreement or small sample size
α with Distance Metrics

<table>
<thead>
<tr>
<th></th>
<th>Jaccard</th>
<th></th>
<th>Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_o$</td>
<td>$D_e$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>A&amp;P, 2006</td>
<td>.53</td>
<td>.95</td>
<td>.45</td>
</tr>
<tr>
<td>Our results</td>
<td>.47</td>
<td>.96</td>
<td>.51</td>
</tr>
</tbody>
</table>

Jaccard$(A, B) = \frac{|A \cap B|}{|A \cup B|}$  
Dice$(A, B) = \frac{2|A \cap B|}{|A| + |B|}$

Artstein and Poesio, 2006: 20 annotators, 16 instances
Our work: 8 annotators, 2,323 instances
**α with Distance Metrics**

<table>
<thead>
<tr>
<th></th>
<th>Jaccard</th>
<th></th>
<th>Dice</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_o$</td>
<td>$D_e$</td>
<td>$\alpha$</td>
<td>$D_o$</td>
</tr>
<tr>
<td>A&amp;P, 2006</td>
<td>.53</td>
<td>.95</td>
<td>.45</td>
<td>.43</td>
</tr>
<tr>
<td>Our results</td>
<td>.47</td>
<td>.96</td>
<td>.51</td>
<td>.36</td>
</tr>
</tbody>
</table>

Jaccard($A, B$) = $\frac{|A \cap B|}{|A \cup B|}$  

Dice($A, B$) = $\frac{2|A \cap B|}{|A| + |B|}$

**Artstein and Poesio, 2006:** 20 annotators, 16 instances  
**Our work:** 8 annotators, 2,323 instances
Our work is not restricted to a particular syntactic type of the antecedent—rather we provide the flexibility to adapt the method if one wants to apply them to arbitrary text segments in specific domains and need serious adaptation if one wants to apply them to arbitrary English text corpora.

To create an initial annotated dataset, we collected a small but representative corpus of text and used it to train and test an anaphora resolution system. We focused on events and actions and used verbs as a proxy for the antecedent for the given anaphors within the given text.

### 3.1 Inter-annotator Agreement

Annotators had to mark the antecedents corresponding to each anaphor instance in their respective copies of the text, as shown in Figure 5. The marked antecedents are mutually exclusive for the non-overlap criterion by creating different copies of the same text corresponding to each anaphor instance.

#### Distance function:

Square-distance between non-overlapping tokens

- The general form of coefficient
- Unitizing coefficient
- Assesses agreement between annotators to given units of analysis
- When perfect reliability and perfect agreement are assumed, the disagreements are observed and expected distribution is computed
- Both disagreement quantities express the average squared differences between the observed and expected distributions
- Agreement and disagreement are measured as follows:

\[ D = \sum \frac{(o - e)^2}{e} \]

where

- \( D \) is the distance function
- \( o \) is the observed value
- \( e \) is the expected value

1. The argument \( \alpha \) is the form of coefficient
2. We denote the marked antecedents for the corresponding anaphors in CL
3. In our context, this means identifying and marking some work on effective annotation of abstract text.

### Data and Annotation

There is currently no English corpus annotated for anaphora antecedents. However, to the best of our knowledge, there are also some works on effective annotation of abstract text.

To make the task tractable, we assumed that annotator's task was to mark arbitrary text segments unseen by the developer for testing. The annotators were domain experts, and the remaining 0.5 instances were annotated by two annotators, a domain expert and a non-expert, and the remaining 0.5 instances were annotated by two annotators.

Instances were discarded as they had an unrelated surrounding context from Medline abstracts.

To create an initial annotated dataset, we collected 5,8 instances. 5,86 instances were independently annotated by two annotators, a domain expert, and a non-domain expert. The remaining 0.5 instances were annotated by two annotators, a domain expert, and a non-domain expert.

5,81 instances were independently annotated by two annotators, a domain expert, and a non-domain expert. The remaining 0.5 instances were annotated by only one annotator in Botley. 3 instances along with the test data were gathered from Poesio and Artstein.

#### Annotate

Annotators had to mark the antecedents corresponding to each anaphor instance in their respective copies of the text, as shown in Figure 5. The marked antecedents are mutually exclusive for the non-overlap criterion by creating different copies of the same text corresponding to each anaphor instance

#### Annotate

Annotators had to mark the antecedents corresponding to each anaphor instance in their respective copies of the text, as shown in Figure 5. The marked antecedents are mutually exclusive for the non-overlap criterion by creating different copies of the same text corresponding to each anaphor instance.

#### Annotate

Annotators had to mark the antecedents corresponding to each anaphor instance in their respective copies of the text, as shown in Figure 5. The marked antecedents are mutually exclusive for the non-overlap criterion by creating different copies of the same text corresponding to each anaphor instance.

#### Annotate

Annotators had to mark the antecedents corresponding to each anaphor instance in their respective copies of the text, as shown in Figure 5. The marked antecedents are mutually exclusive for the non-overlap criterion by creating different copies of the same text corresponding to each anaphor instance.

#### Annotate

Annotators had to mark the antecedents corresponding to each anaphor instance in their respective copies of the text, as shown in Figure 5. The marked antecedents are mutually exclusive for the non-overlap criterion by creating different copies of the same text corresponding to each anaphor instance.
## Unitizing $\alpha$

(Krippendorff, 2013)

<table>
<thead>
<tr>
<th>Intersection</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotator 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annotator 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Distance function:
square-distance between non-overlapping tokens

**Unitizing $\alpha = 0.54$**
CrowdFlower Confidence

**fact**

**reason**

**question**

**issue**

**decision**

**possibility**
CrowdFlower Confidence

- **Fact**: Mean = 0.74
- **Reason**: Mean = 0.71
- **Question**: Mean = 0.77
CrowdFlower Confidence

**fact**

Mean = 0.74

**reason**

Mean = 0.71

**question**

Mean = 0.77

**issue**

Mean = 0.60

**decision**

Mean = 0.59

**possibility**

Mean = 0.62
ASN Antecedent Corpus

- 1,810 high confidence (confidence ≥ 0.5) instances from the CrowdFlower expt. 2
Evaluation by Experts

- Goal
  Examine acceptability of crowd's answers
- Judges
  - Two highly-qualified academic editors
- Evaluation options
  - Perfectly, Reasonably, Implicitly, Not at all
Evaluation by Experts

Perfectly (P), Reasonably (R), Implicitly (I), Not at all (N)

<table>
<thead>
<tr>
<th></th>
<th>Judge A</th>
<th>Judge B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Judge A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>171</td>
<td>44</td>
</tr>
<tr>
<td>R</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>186</td>
<td>77</td>
</tr>
</tbody>
</table>
Evaluation by Experts

Perfectly (P), Reasonably (R), Implicitly (I), Not at all (N)

<table>
<thead>
<tr>
<th>Judge B</th>
<th>P</th>
<th>R</th>
<th>I</th>
<th>N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge A</td>
<td>P</td>
<td>171</td>
<td>44</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>12</td>
<td>27</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>186</td>
<td>77</td>
<td>24</td>
<td>13</td>
<td>300</td>
</tr>
</tbody>
</table>

According to the experts, ~95% instances had acceptable annotations.
Conclusion

• Examined feasibility of annotating antecedents of ASNs (e.g., *this issue*, *this fact*) using crowdsourcing

• Most of the time, at least half of the annotators converged on one answer

• Evaluated crowd-annotations using experts

• Resulted in an ASN antecedent corpus containing 1,810 instances with high-confidence annotation
Error Analysis

• Problem with agreeing on *None*
• Multiple possible answers
• Hard instances
• Different strings representing similar concepts
Several Vatican officials said, however, that any such talk has little meaning because the church does not take sides in elections. But the statements by several American bishops that Catholics who vote for Mr. Kerry would have to go to confession have raised the question in many corners about whether this is an official church position.

The church has not addressed this question publicly and, in fact, seems reluctant to be dragged into the fight..."
Any biography of Thomas More has to answer one fundamental question. Why? Why, out of all the many ambitious politicians of early Tudor England, did only one refuse to acquiesce to a simple piece of religious and political opportunism? What was it about More that set him apart and doomed him to a spectacularly avoidable execution?

The innovation of Peter Ackroyd’s new biography of More is that he places the answer to this question outside of More himself.