Annotating Anaphoric Shell Nouns with their Antecedents

Varada Kolhatkar and Graeme Hirst
University of Toronto

Heike Zinsmeister
Universität Stuttgart
Anaphoric Shell Nouns (ASNs) (Schmid, 2000)

Our paper: Annotating antecedents of ASNs such as *this issue*, *this fact*, *this possibility*

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

| Corpus                          | ASN instances                                                                 |
|---------------------------------|-------------------------------------------------------------------------------|
| Poesio and Artstein, 2008 ARRAU | 455 abstract anaphor instances, very few ASNs                                |
| Kolhatkar and Hirst, 2012       | 188 *this issue* instances from Medline abstracts                             |
| Botley, 2006                    | 462 ASN instances *(not available)*                                           |
# Current Research: Gap

| Corpus                        | ASN instances                                                                 |
|-------------------------------|-------------------------------------------------------------------------------|
| Poesio and Artstein, 2008     | 455 abstract anaphor instances, very few ASNs                                |
| ARRAU                         |                                                                               |
| Kolhatkar and Hirst, 2012     | 188 *this issue* instances from Medline abstracts                             |
| Botley, 2006                  | 462 ASN instances *(not available)*                                           |

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, ...)
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)

| Category          | Examples               |
|-------------------|------------------------|
| Factual           | fact, problem, reason  |
| Linguistic        | news, proposal, question |
| Mental            | idea, belief, decision, issue |
| Modal             | possibility, need, trend |
| Eventive          | act, reaction, attempt |
| Circumstantial    | situation, approach    |
## Shell Nouns: Selection

| Category     | Examples                      |
|--------------|-------------------------------|
| Factual      | fact, problem, reason         |
| Linguistic   | news, proposal, question      |
| Mental       | idea, belief, decision, issue |
| Modal        | possibility, need, trend      |
| Eventive     | act, reaction, attempt        |
| Circumstantial | situation, approach         |
# Shell Nouns: Selection

| Category       | Examples                      |
|----------------|-------------------------------|
| Factual        | *fact*, *problem*, *reason*   |
| Linguistic     | *news*, *proposal*, *question*|
| Mental         | *idea*, *belief*, *decision*, *issue* |
| Modal          | *possibility*, *need*, *trend*|
| Eventive       | *(Act, reaction, attempt)*    |
| Circumstantial | *(Situation, approach)*       |

High-frequency nouns from each category
The ASN Corpus

- Base corpus: The New York Times corpus (Sandhaus, 2008)
- ~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

| Annotator | Trust | Answer |
|-----------|-------|--------|
| A         | 0.75  | “a₁”   |
| B         | 0.75  | “a₂”   |
| C         | 1.0   | “a₂”   |
## CrowdFlower Confidence

| Annotator | Trust | Answer |
|-----------|-------|--------|
| A         | 0.75  | “a₁”   |
| B         | 0.75  | “a₂”   |
| C         | 1.0   | “a₂”   |

Score for “a₂” = 1.75
## CrowdFlower Confidence

| Annotator | Trust | Answer |
|-----------|-------|--------|
| A         | 0.75  | “a₁”   |
| B         | 0.75  | “a₂”   |
| C         | 1.0   | “a₂”   |

**score for “a₁” = 0.75**

**score for “a₂” = 1.75**

Crowd’s answer:
“a₂” with confidence 0.7 = 1.75/(1.75 + 0.75)
Annotation Tasks

ASN instances from the NYT

Identify the sentence containing antecedent

Identify the precise antecedent

Annotated ASN Corpus

CrowdFlower Expt. 1

CrowdFlower Expt. 2

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. I
Confidence Distribution

- **Fact**
  - Confidence distribution for fact
  - Number of instances

- **Reason**
  - Confidence distribution for reason
  - Number of instances

- **Question**
  - Confidence distribution for question
  - Number of instances

- **Issue**
  - Confidence distribution for issue
  - Number of instances

- **Decision**
  - Confidence distribution for decision
  - Number of instances

- **Possibility**
  - Confidence distribution for possibility
  - Number of instances

29
Confidence Distribution

- **fact**: Mean = 0.83
- **reason**: Mean = 0.82
- **question**: Mean = 0.80
- **issue**: Mean = 0.83
- **decision**: Mean = 0.83
- **possibility**: 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

|                | Jaccard |        |        | Dice |        |        |
|----------------|---------|--------|--------|------|--------|--------|
|                | $D_o$   | $D_e$  | $\alpha$ | $D_o$ | $D_e$  | $\alpha$ |
| A&P, 2006      | .53     | .95    | .45    | .43  | .94    | .55    |
| Our results    | .47     | .96    | .51    | .36  | .92    | .61    |

\[
\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}
\]

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

|       | Jaccard |       |      |       |      |
|-------|---------|-------|------|-------|------|
|       | $D_o$   | $D_e$ | $\alpha$ | $D_o$ | $D_e$ | $\alpha$ |
| A&P, 2006 | .53     | .95   | .45  | .43   | .94  | .55  |
| Our results | .47     | .96   | .51  | .36   | .92  | .61  |

\[
\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

\[
\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}
\]

Artstein and Poesio, 2006: 20 annotators, 16 instances
Our work: 8 annotators, 2,323 instances
Unitizing $\alpha$

(Krippendorff, 2013)

Distance function:
square-distance between non-overlapping tokens
Unitizing $\alpha$

(Krippendorff, 2013)

| Annotator 1 | $r_{11}$ | $r_{12}$ | $r_{13}$ | $r_{14}$ | $r_{15}$ | $r_{16}$ | $r_{17}$ | $r_{18}$ | $r_{19}$ |
|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
|             |          |          |          |          |          |          |          |          |          |
|             |          |          |          |          |          |          |          |          |          |
| $id_2$      |          |          |          |          |          |          |          |          |          |
| $id_3$      |          |          |          |          |          |          |          |          |          |
| $id_4$      |          |          |          |          |          |          |          |          |          |
| $id_5$      |          |          |          |          |          |          |          |          |          |

| Annotator 2 | $r_{21}$ | $r_{22}$ | $r_{23}$ | $r_{24}$ | $r_{25}$ | $r_{26}$ | $r_{27}$ | $r_{28}$ | $r_{29}$ | $r_{2,10}$ |
|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|------------|
|             |          |          |          |          |          |          |          |          |          |            |
|             |          |          |          |          |          |          |          |          |          |            |
| $id_1$      |          |          |          |          |          |          |          |          |          |            |
| $id_2$      |          |          |          |          |          |          |          |          |          |            |
| $id_3$      |          |          |          |          |          |          |          |          |          |            |
| $id_4$      |          |          |          |          |          |          |          |          |          |            |
| $id_5$      |          |          |          |          |          |          |          |          |          |            |

| Intersections | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 |
|---------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|

Distance function:
square-distance between non-overlapping tokens

Unitizing $\alpha = 0.54$
CrowdFlower Confidence

fact

reason

question

issue

decision

possibility
CrowdFlower Confidence

**Confidence**
- **fact** Mean = 0.74
- **reason** Mean = 0.71
- **question** Mean = 0.77

**Number of instances**
- **issue**
- **decision**
- **possibility**
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 $\geq 0.5$) 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)**

|                | Judge B |       |       |       |     |
|----------------|---------|-------|-------|-------|-----|
| Judge A        |         |       |       |       |     |
| P              | 171     | 44    | 11    | 7     | 233 |
| R              | 12      | 27    | 7     | 4     | 50  |
| I              | 2       | 4     | 6     | 1     | 13  |
| N              | 1       | 2     | 0     | 1     | 4   |
| **Total**      | 186     | 77    | 24    | 13    | 300 |

We asked two highly qualified judges to independently examine the quality of a sample of such antecedents. Usually, at least 5 annotators worked independently with a minimal set of annotation instruction, which is a necessary step before developing computational approaches to resolve SN antecedents. We also plan to distribute the annotated data as training data for SN resolution. We carried out crowdsourcing experiments to get a large sample of annotated SN antecedents. According to her the correct antecedent is rather a perfect one, while the other rated this answer as perfectly. Moreover, if they did not converge on a single SN antecedent, we plan to use this crowd-annotated SN antecedents. According to an SN resolver, we also plan to distribute the annotated data as training data for SN resolution. The result is quite encouraging considering the nature of the task. In this paper, we addressed the fundamental question: whether American bishops that Catholics who vote for Mr. Kerry would have to go to confession have raised the question in many corners about Catholics who vote for Mr. Kerry, who is the crowd's answer. One of our judges marked this sentence as perfectly, the other rated it as reasonably. This was a positive result; the crowd's answer only implicitly contains the actual antecedent, while the judge rated this answer as perfectly. These are borderline cases, and we used them in our evaluation. The church has not addressed the question in many corners about the official church position. However, the statements by several Vatican officials said, however, that any such talk has little meaning because the church does not publicly address the question in many corners about the official church position. Several Vatican officials said, however, that any such talk has little meaning because the church does not publicly address the question in many corners about the official church position. The confidence level, many instances with borderline confidence were marked by both judges, 83% by the crowd. The table shows the confusion matrix of the raters, which is a necessary step before developing computational approaches to resolve SN antecedents.

We thank the crowdflower team for their responses. This material is based upon work supported by the United States Air Force and the National Science Foundation's Faculty Research Exchange and the University of Toronto. Ontario Advanced Research Projects Agency under contract N0161-10-1280.
Evaluation by Experts

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

|          | Judge B |
|----------|---------|
|          | P       | R | I | N | Total |
| Judge A  | P       | 171 | 44 | 11 | 7   | 233  |
|          | R       | 12  | 27 | 7  | 4   | 50   |
|          | I       | 2   | 4  | 6  | 1   | 13   |
|          | N       | 1   | 2  | 0  | 1   | 4    |
| Total    | 186     | 77  | 24 | 13 | 300 |

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
Syntactic Type Distribution

- Sentences
- Verb Phrases
- Clauses
- Adjective Phrases
- Noun Phrases
- Prepositional Phrases

Sentences

Clauses

Noun Phrases

Verb Phrases

Prepositional Phrases

Adjective Phrases

Fact

Reason

Issue

Decision

Question

Possibility
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...”
Hard Examples

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.