

Identifying referents through annotation of dialogue transcripts

Ron Artstein

Department of Computer Science
University of Essex

Workshop on incrementality and clarification, 15 February 2007

Joint work with Massimo Poesio
Thanks to EPSRC grant GR/S76434/01, ARRAU
(Anaphora Resolution and Underspecification)

The ARRAU project

- The Arrau project set out to investigate:
 - ambiguous anaphoric expressions
 - anaphoric reference to abstract objects
- Create a corpus annotated for the above
- Annotation experiments:
 - test the proposed annotation scheme
 - allow us to study how people interpret text

Annotating anaphoric relations

- A corpus annotated with anaphoric relations is needed for development and evaluation of anaphoric resolvers (cf. Eckert and Strube, 2000)
- Schemes like MUC do not capture linguistic notions of coreference or anaphora (van Deemter and Kibble, 2000)
- Need to determine:
 - What should be annotated?
 - What **can** be annotated reliably?
- Answers to be determined empirically, through experiments.

Annotators and dialogue

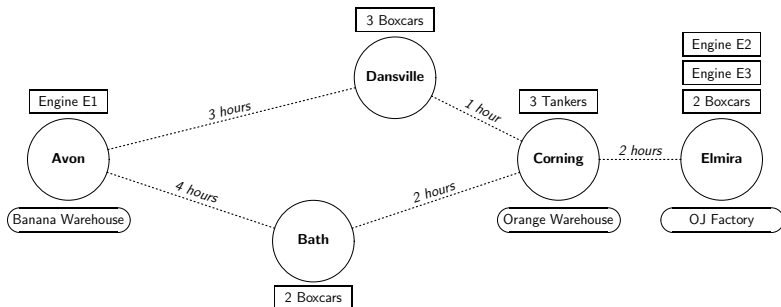
- Annotators of a dialogue transcript play the role of **overhearers** in a conversation (Schober and Clark, 1989)
- Overhearers may be part of a dialogue setting (broadly interpreted)
- Linguists who analyze data are in some sense overhearers

Outline

- 1 Motivation
- 2 The annotation experiment
 - Setup
 - Annotation scheme
 - Results
- 3 Reliability
 - Agreement measures
 - Agreement results

Materials

- Dialogue 14.2 from the TRAINS 93 corpus
 - Task-oriented.
 - Limited domain makes references tractable (somewhat).
- Map of the “TRAINS world”



Procedure

- 11 naïve participants, working simultaneously in one lab.
- MMAX2 tool for annotating transcripts.
- Short manual (8+6+2 pages).
- One hour of training: annotate a short dialogue together.
- Three passes over the text:
 - ① Identify objects only
 - ② Identify anaphoric relations and ambiguity
 - ③ Identify discourse antecedents
- Annotation completed in four two-hour sessions for dialogue transcript, Wall Street Journal text, and personal email.

First pass: marking objects

- All NPs are markables on the **phrase** level.
- Participants select the appropriate object for each markable.

One-click annotation

phrase utterance

Comment

Gender unmarked male female neuter unspecified

< > Reference unmarked referring non_referring

Category unmarked person animate concrete space time plan abstract unknown

< > Object unmarked engine railcar facility commodity place other

Engine unmarked engine engine_e1 engine_e2 engine_e3

Confidence unmarked 1 2 3 4 5

Suppress check Warn on extra attributes

Apply Undo changes

to front Auto-apply

Auto-apply is ON



Second pass: marking antecedents

- **Status** indicates whether a markable is new or old.

```
< > Status       unmarked  new_object  prev_phrase  prev_segment
< > Phrase_Antecedent  single_phrase  multiple_phrases
Single_phrase_antecedent110
< > Related_phrase  unmarked  no  part_of  member_of  converse
< > Ambiguity      unmarked  unambiguous  ambiguous  ambiguous_antecedent
```

- Old markables are linked to the previous mention.

U: <click> okay um so if we leave [Elmira] at [six AM] with [one tank of [orange juice]]
 S: uh-huh
 U: okay and then we need to take [that] to [Avon]

- Two interpretations may be marked for ambiguous items.



Inconsistency in labels

In principle we expect the objects at the two ends of an anaphoric link to be identical.

U: <click> okay um so if we leave [Elmira] at [six AM] with [one tank of [orange juice]]

S: uh-huh

U: okay and then we need to take [that] to [Avon]

- 250 markables \times 11 annotators = 2750 judgments
- 1195 instances of unambiguous markables with a single unambiguous antecedent
- 133 of the above (11%) show a mismatch between the objects

Many of the mismatches are errors – but some are more interesting.

Inconsistency = incrementality

- 27 u: okay um and then take um one one tank of oranges from Corning back to Elmira
- 28 s: um as I said you can't carry oranges in a tanker until they have been turned into orange juice you have to carry oranges in **a boxcar** until they are actually turned into juice
- 29 u: okay okay so we want **the boxcar** to leave Elmira for Corning at midnight

2 **the boxcar** = boxcar from Elmira, new object

4 **the boxcar** = boxcar from Elmira → **a boxcar** = boxcar

4 **the boxcar** = boxcar → **a boxcar** = boxcar

1 **the boxcar** = boxcar, discourse antecedent

Incrementality

Semantic analysis:

- **the boxcar** is a new discourse referent.
 - Not a canonical novel definite (Poesio and Vieira, 1998)
- **the boxcar** is anaphoric (e.g. “the boxcar which is needed for carrying the oranges”)
 - Accessible antecedent in a generic/intensional context
 - Incremental interpretation of discourse referent

Inconsistency = ambiguity

- 39 u: <click> okay um so if we leave Elmira at six a.m.
with one tank of orange juice
- 40 s: uh-huh
- 41 u: okay and then we need to take **that** to Avon

- 1 **that** = orange juice → one tank of orange juice = boxcar from Elmira
- 2 **that** = orange juice → one tank of orange juice = tanker car
- 1 **that** = tanker car → orange juice = orange juice
- 1 **that** = ambiguous
- 3 **that** = tanker car → one tank of orange juice = tanker car
- 2 **that** = tanker car, discourse antecedent
- 1 **that** = "1 tank oj", discourse antecedent



Ambiguity

Semantic analysis:

- **that** is ambiguous between the the tanker car and the juice that it contains
- Distinction irrelevant for the purpose of the plan
 - “Justified sloppiness” (Poesio, Reyle, and Stevenson, 2001/2007)

Inconsistency = unclear reference

- 67 u: okay and then also at midnight we need to send
 another boxcar to leave for Avon
- 68 s: from where
- 69 u: from Elmira
- 70 s: okay um you need to run that about an hour behind
- 71 u: okay
- 72 s: the other engine
- 73 u: <click> that's fine
- 74 s: so it would leave at one

	it → another boxcar	that	the other engine	Non-anaphoric
boxcar	1	1	2	1
engine	2		2	1 ¹

Unclear reference

Semantic analysis:

- Irrelevant whether **it** refers to the boxcar or engine
 - “Justified sloppiness”
- But the only source for the interpretation of “engine” is mention of **the other engine!**
- Also, **the other engine** refers to the only engine mentioned in the discourse so far. . .



Reliability measures

Qualitative analysis can only give an impression of whether the annotators agree.

What we learn from formal agreement measures:

- Quantify the amount of agreement (sort-of)
- Identify easy and difficult things to annotate
- Assess suitability of different agreement measures

Kappa (a.k.a. pi): chance-corrected agreement

observed agreement: mean pairwise agreement per item

Item: 7 boxcar, 4 engine

$$\text{agreement} = \frac{\binom{7}{2} + \binom{4}{2}}{\binom{11}{2}} = \frac{7 \times 6 + 4 \times 3}{11 \times 10} \approx 0.491$$

expected agreement: pairwise agreement expected by chance

$$K = \frac{\text{observed agreement} - \text{expected agreement}}{1 - \text{expected agreement}}$$

Krippendorff's alpha

Similar to K , but also applies to values on a scale.

- For nominal labels, $\alpha \approx K$
- For numerical labels, α related to F (ANOVA)

$$F = \frac{S_{\text{between}}^2}{S_{\text{within}}^2} \quad \alpha = 1 - \frac{S_{\text{within}}^2}{S_{\text{total}}^2}$$

Interpreting kappa and alpha

$K, \alpha = 1$: perfect agreement

$K, \alpha = 0$: chance agreement

$K, \alpha < 0$: systematic disagreement



Agreement on labels

	pairwise agreement		chance-corrected (K)	
	stage 1	stage 2	stage 1	stage 2
Ref/Nonref	0.87	0.88	0.45	0.47
Coarse object	0.79	0.80	0.74	0.75
Fine object	0.75	0.75	0.72	0.72
Category	0.81	0.82	0.75	0.76

The slight improvement in agreement from the first pass to the second is probably due to correction of errors:

- Total 108/11000 changes
- 72 increase agreement, 24 decrease



Agreement on detailed labels

	pairwise agreement		chance-corrected (K)	
	stage 1	stage 2	stage 1	stage 2
Ref/Nonref	0.87	0.88	0.45	0.47
Coarse object	0.79	0.80	0.74	0.75

Increase in chance-corrected agreement:

- random decision whether phrase is referential
- systematic agreement on what it refers to

Hypothesis: phrases that are genuinely ambiguous **FALSE**

Reality: random decision, then labeling the object as “other”

Confidence

Annotators indicated confidence in their annotation of each item.

- Confidence scale 1–5, mean 3.9
- Participants differ in confidence (means 2.5–4.8)
- Items differ in confidence: $F(249, 2500) = 2.85, p < .0001$
- Agreement on confidence of items is quite low: $\alpha = 0.14$

Confidence as a predictor

Can an individual confidence score serve as a predictor of the difficulty of an item?

- Agreement on an item is indicative of its difficulty
- High correlation between mean confidence per item and amount of agreement per item: $r = 0.74$
- Lower correlation between individual confidence per item and amount of agreement per item: $0.22 \leq r \leq 0.66$
- High correlation between an annotator's overall confidence and their predictive power: $r = 0.67, p < .05$

Cannot interpret the final correlation.

Conclusions

- Need to test annotators' interpretation of text
 - Not just devise a scheme with high agreement
- Devise multiple ways of marking reference
 - Inconsistencies indicative of annotation difficulty
 - Or interpretation difficulty
 - Or ambiguity
- Run annotation experiments with many coders
- Use formal reliability measures
 - Reliability is a research tool, not a hurdle for publication
- Confidence is indicative
 - but we're not sure what it indicates

References

- van Deemter, Kees and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.
- Eckert, Miriam and Michael Strube. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, 17(1):51–89.
- Poesio, Massimo, Uwe Reyle, and Rosemary Stevenson. 2001/2007. Justified sloppiness in anaphoric reference. In Harry Bunt and Reinhard Muskens, editors, *Computing Meaning, Volume 3*, volume 83 of *Studies in Linguistics and Philosophy*. Springer. To appear September 2007.
- Poesio, Massimo and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.
- Schober, Michael F. and Herbert H. Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.