The Arrau corpus of anaphoric relations

Ron Artstein[1] Massimo Poesio[2] | AACL, March 15, 2008

[1] Institute for Creative Technologies, University of Southern California

[2] Università di Trento and University of Essex

Corpus creation funded by EPSRC grant GR/S76434/01, ARRAU (Anaphora Resolution and Underspecification)

Travel support from projects funded by the U.S. Army Research, Development and Engineering Command (RDECOM)

Anaphora/coreference resolution

$$\begin{aligned}\text{Anaphora resolution} &\approx \text{Coreference resolution}\\ &\approx \text{Entity disambiguation}\end{aligned}$$

Identify mentions (expressions) that refer to the same entity.

- MUC and ACE initiatives

Annotating coreference relations

Implicit assumption: each mention co-refers with a unique previous expression.

5.5 : I have to get a boxcar
5.6 : to Corning
5.7 : and then I have to load **it** with oranges . . .

Trains 91, dialogue 1.1

Ambiguity

Sometimes there is no clear unique antecedent.

18.6 : it turns out that the boxcar at Elmira
18.7 : has a bad wheel
18.8 : and they're .. gonna start fixing **that** at midnight
18.9 : but it won't be ready until 8

Trains 91, dialogue 3.2

Abstract entities

Antecedent which is only implicitly evoked by preceding text.

7.3 : so we ship one
7.4 : boxcar
7.5 : of oranges to Elmira
7.6 : and **that** takes another 2 hours

Trains 91, dialogue 2.2

that ≈ “the shipping of one boxcar of oranges to Elmira”

The Arrau project

- EPSRC funding to the University of Essex, 2004–2007
- Explore “difficult” cases of anaphora
  - Ambiguous anaphoric relations
  - Reference to abstract objects (events, plans, actions. . .)
- Annotation experiments with multiple participants
- Annotated corpus consisting of multiple genres
  - Dialogue
  - Narrative
  - Newspaper (WSJ)

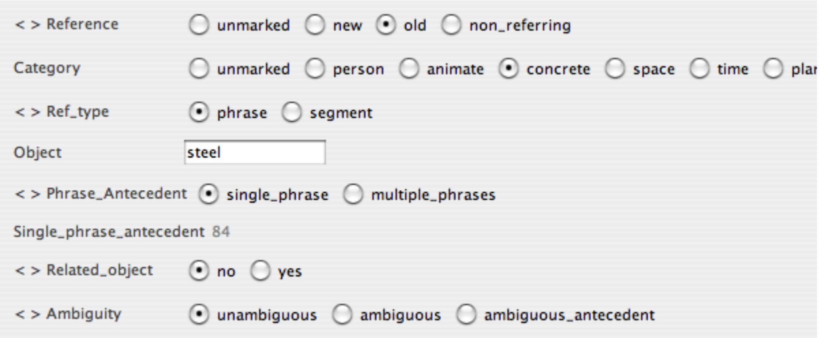
Annotation experiments

- Test annotation schemes for reliability
- Up to 20 participants annotating same text independently
- Short manual to tap into intuitions
- Lead to improved annotation scheme
- Main findings:
  - Reasonable agreement on coreference chains ($\alpha \approx 0.6$–0.7)
  - Spotting ambiguity is difficult
  - Ambiguity can be detected implicitly through disagreement
  - Annotators agree on general referent-evoking textual regions, but disagree on precise boundaries ($\alpha \approx 0.55$)

Annotation format and scheme

- MMAX 2 annotation tool (Müller and Strube, 2003)
  - Multi-level XML format
  - Visual tool
- All noun phrases marked for referential status (new/old/non-referring)
- Coreference links are **pointers** (not equivalence sets)
- Referent-evoking regions are clause-like units
- Limited bridging references
- Each item allows two distinct meanings

Noun phrases marked for referential status



< > Reference ○ unmarked ○ new ● old ○ non_referring

Category ○ unmarked ○ person ○ animate ● concrete ○ space ○ time ○ pla

< > Ref_type ● phrase ○ segment

Object steel

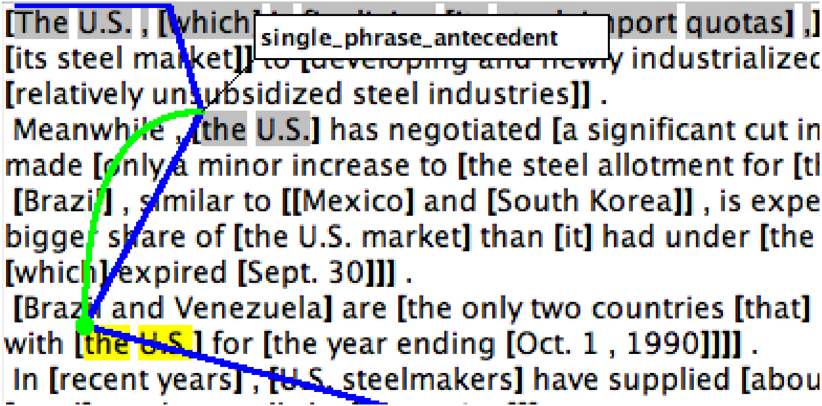
< > Phrase_Antecedent ● single_phrase ○ multiple_phrases

Single_phrase_antecedent 84

< > Related_object ● no ○ yes

< > Ambiguity ● unambiguous ○ ambiguous ○ ambiguous_antecedent

Coreference links are pointers

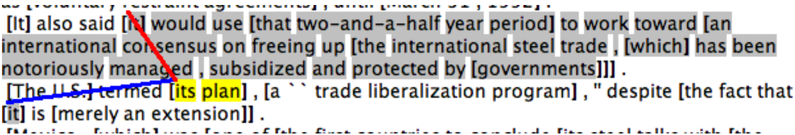


[The U.S. , [which]

single_phrase_antecedent

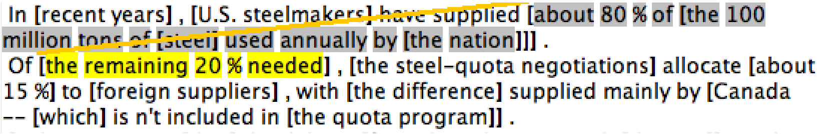
port quotas] ,]
[its steel market]] to [developing and newly industrializec
[relatively unsubsidized steel industries]] .
Meanwhile , [the U.S.] has negotiated [a significant cut in
made [only a minor increase to [the steel allotment for [tl
[Brazil] , similar to [[Mexico] and [South Korea]] , is expe
bigger share of [the U.S. market] than [it] had under [the
[which] expired [Sept. 30]]] .
[Brazil and Venezuela] are [the only two countries [that]
with [the U.S.] for [the year ending [Oct. 1 , 1990]]]] .
In [recent years] , [U.S. steelmakers] have supplied [abou

Referent-evoking regions



as [voluntary restraint agreements] , until [March 31 , 1992] .
[It] also said [it] would use [that two-and-a-half year period] to work toward [an international consensus on freeing up [the international steel trade , [which] has been notoriously managed , subsidized and protected by [governments]]] .
[The U.S.] termed [its plan] , [a `` trade liberalization program] , " despite [the fact that [it] is [merely an extension]] .
[Mexico , [which] was [one of [the first countries to conclude [its steel talks with [the

Limited bridging references



In [recent years] , [U.S. steelmakers] have supplied [about 80 % of [the 100 million tons of [steel] used annually by [the nation]]] .
Of [the remaining 20 % needed] , [the steel-quota negotiations] allocate [about 15 %] to [foreign suppliers] , with [the difference] supplied mainly by [Canada -- [which] is n't included in [the quota program]] .

Composition

Source	Texts	Markables				Words
		total	anaph[a]	seg	ambig	
Trains 91	16	2874	1679	143	19	14496
Trains 93	19	2342	1327	121	11	11287
Gnome	5	6045	2101	58	26	21599
Pear stories	20	3883	2194	50	10	14059
Wall St Jrnl	50	9177	2852	83	37	32771
Total	110	24321	10153	455	103	94212

[a] Those markables for which an explicit nominal antecedent was identified

Conversion, extension and use

- Coreference relations can be converted to equivalence sets of mentions by eliminating information on ambiguity.
- Wall Street Journal portion augmented by automatic conversion from Vieira-Poesio corpus and Moscow RST Discourse Treebank.
- Corpus used with the BART system, developed at the Johns Hopkins 2007 Summer Workshop on Natural Language Engineering.

Planned release

- Corpus currently undergoing verification and checking.
- Hope to release soon, via LDC.

References

Müller, Christoph and Michael Strube. 2003. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 198–207, Sapporo.