Anaphoric Annotation in the ARRAU Corpus

Massimo Poesio,[*] Ron Artstein[†]

[*]Università di Trento and University of Essex
Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
poesio [at] essex.ac.uk

[†]Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina Del Rey CA 90292, USA
artstein [at] ict.usc.edu

Abstract

Arrau is a new corpus annotated for anaphoric relations, with information about agreement and explicit representation of multiple antecedents for ambiguous anaphoric expressions and discourse antecedents for expressions which refer to abstract entities such as events, actions and plans. The corpus contains texts from different genres: task-oriented dialogues from the Trains-91 and Trains-93 corpus, narratives from the English Pear Stories corpus, newspaper articles from the Wall Street Journal portion of the Penn Treebank, and mixed text from the Gnome corpus.

1. Introduction

Although large-scale annotated corpora for the empirical study of anaphora such as Ontonotes (Hovy et al., 2006; Pradhan et al., 2007) are finally becoming available, such resources are still limited. Most annotation efforts concentrate on written text – indeed, on a very specific type of written text, newspaper articles. Also, because many theoretical questions about anaphora are still poorly understood (Zaenen, 2006), the range of anaphoric phenomena being annotated tends to be very narrow: often only relations between mentions realized with proper names, virtually never agreement information, and in very few cases bridging relations.

A specific problem with most existing schemes for anaphora annotation is the assumption that each anaphoric expression has a clearly identifiable antecedent which is realized through another noun phrase. However, anaphoric expressions can be ambiguous as to which object they denote, and in some cases this doesn't affect the overall interpretation, a phenomenon that Poesio et al. (2007) call "justified sloppiness". Furthermore, there have only been few attempts to mark discourse deixis (Webber, 1991), the anaphoric relation where an expression, typically a demonstrative, refers to something abstract like an event, action or plan which has been introduced in previous discourse but not with a referring noun phrase.

In this paper we introduce the Arrau corpus, created at the University of Essex between 2004 and 2007 as part of the Arrau project[1] to address these problems. The primary goal of our annotation effort was to develop methods for marking ambiguous anaphoric expressions, and expressions which refer to abstract entities such as events, actions and plans. We conducted a series of experiments to test the feasibility of annotating these phenomena, and then annotated texts from a variety of genres. In addition, we annotated information that we knew could be annotated reliably from previous efforts, including information about agreement features and bridging relations. A first release of the corpus has been completed and used in the ELERFED workshop;[2] we hope to make the corpus publicly available before LREC.

2. Previous annotation efforts

This work builds on several years of experience with anaphoric annotation summarized in Poesio (2004b). The annotation experiments discussed in Poesio and Vieira (1998) were primarily concerned with establishing whether annotation could reliably capture a distinction between coreference, bridging, and discourse novelty inspired by the proposals of Prince (1992), using newspaper articles from the Wall Street Journal portion of the Penn Treebank as data. These studies revealed the difficulty of annotating reliably the whole range of bridging relations, and distinguishing those cases from discourse-new. The result was the so-called Vieira-Poesio corpus of 34 documents, which was used in a number of subsequent studies.

Subsequent work on the GNOME corpus (Poesio, 2004a) focused on identifying a subset of bridging relations that could be reliably annotated, as well as annotating other types of information that could be useful in the study of anaphora in general and local salience in particular, such as agreement features, grammatical function, and animacy. The GNOME coding instructions allowed for ambiguity, but this feature was not systematically studied. Limited progress was made towards annotating discourse deixis, marking only the type of antecedent of abstract anaphora in an attempt to distinguish between references to events, temporal objects, propositions, and types (Poesio and Modjeska, 2005).

In subsequent work on the VENEX corpus (Poesio et al., 2004) we carried out preliminary studies of annotation of coreference in dialogue, including annotation of references to objects in the visual situation, and started experimenting with an early version of the MMAX tool (Müller and

[1]EPSRC grant number GR/S76434/01, http://cswww.essex.ac.uk/Research.nle/arrau/.

[2]http://www.clsp.jhu.edu/ws2007/groups/elerfed/

Strube, 2006).

3. Agreement on ambiguity and discourse deixis

In order to determine the best way to annotate anaphoric ambiguity and discourse deixis in Arrau we conducted a series of experiments, some with pen and paper, but mostly using an annotation tool, MMAX2 (Müller and Strube, 2006), which forced the coders to use a predefined scheme. In these annotation experiments, multiple annotators (as many as 20) worked independently on the same text, and formal reliability measures such as α (Krippendorff, 1980) were used to compare the annotations and identify easy and difficult parts of the task; agreement on anaphoric chains was in the range of $\alpha \approx 0.6$–0.7 (Poesio and Artstein, 2005b). We found that while annotators often missed the ambiguity of an item, it was possible to identify ambiguity implicitly when sets of annotators chose different antecedents for a single item (Poesio and Artstein, 2005a). For discourse deixis we found that annotators agreed on the general textual regions that evoke the referents, though they often disagreed on the exact boundaries, resulting in agreement of around $\alpha \approx 0.55$ (Artstein and Poesio, 2006).
The experiments led to progressive refinements of the annotation scheme to one which was more reliable yet expressed the distinctions of interest. We clarified the distinction between multiple antecedents for plural expressions and for ambiguous expressions, we added a way to mark ambiguity between discourse-old and discourse-new interpretations, and we constrained the marking of textual regions to predefined clause-level units. The resulting scheme is described in section 4.2.

4. The corpus

4.1. Composition

The ARRAU corpus contains texts from a mixture of genres, including dialogue, narrative, and a variety of genres of written text. Task-oriented dialogues include texts from the Trains-91 and Trains-93 corpora (Gross et al., 1993; Heeman and Allen, 1995). Spoken narratives include the full English Pear Stories corpus of Narratives (Chafe, 1980).[3] Examples of written text include five texts from the Gnome corpus not yet annotated for anaphoric relations and – the final and largest part of the corpus – newspaper text from the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993). Most of the WSJ texts are part of the RST Discourse Treebank (Carlson et al., 2003) and are being annotated as part of a joint effort with prof. Kibrik's group at the the Russian Academy of Sciences in Moscow.
The current composition of the corpus is summarized in Table 1. It is expected that the final version of the corpus will also include half of the texts in the RST discourse treebank, as well as the files from the Vieira-Poesio corpus and the previously annotated files in the GNOME corpus.

4.2. Annotation scheme

The corpus was created using the MMAX2 tool (Müller and Strube, 2006), which allows marking text units at different levels. All noun phrases are treated as markables which can be anaphoric or serve as antecedents (or both), and all clauses are treated as potential antecedents for discourse deixis. When NPs and clauses were not already marked we identified them using the Charniak parser (Charniak, 2000) and then corrected the output by hand.

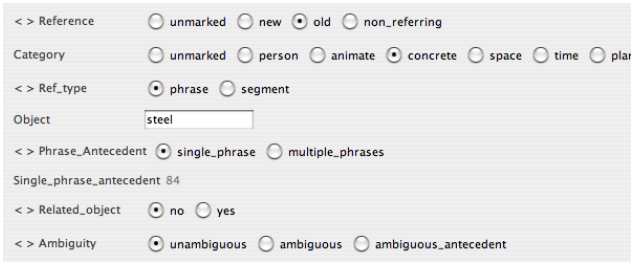

Figure 1: Attributes (partial screenshot)

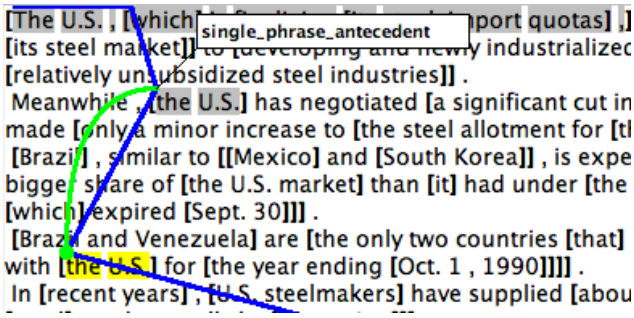

Figure 2: A pointer from a noun phrase to its antecedent (the green arc)

Each noun phrase is annotated with a set of attributes (Figure 1). These include gender, grammatical function, number, person, and a "category" attribute which combines animacy and a concrete/abstract distinction. The "reference" attribute indicates whether a noun phrase is anaphoric, discourse-new, or non-referential. If it is referential then the referent is identified – in a restricted domain like Trains the referent is selected from a list, otherwise it is entered as free text.

Anaphoric expressions are linked to previous discourse by pointers to their antecedents, which are entered through the MMAX2 graphical interface (Figure 2). Multiple pointers form a single anaphoric expression mark plural antecedents, and two distinct sets of pointers are available for each expression in order to indicate ambiguity (Poesio and Artstein, 2005a). Anaphora is therefore not an equivalence relation, and markables form more complex structures than equivalence sets indicating identity of reference. Reference to an event, action or plan is marked by a pointer from the referring NP to the clause that introduces the abstract entity (Artstein and Poesio, 2006) (Figure 3). The scheme also al-

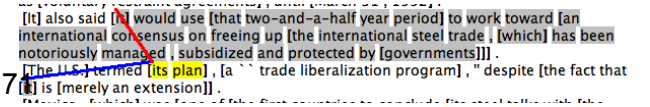

Figure 3: A pointer from a noun phrase to a textual region

[3] http://www.pearstories.org

Source	Texts	Markables				Words
		total	anaphoric[a]	discourse [b]	ambiguous[c]	
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]Markables with a nominal antecedent
[b]Pointing to a text region
[c]Identified explicitly as ambiguous

Table 1: Composition of the Arrau corpus

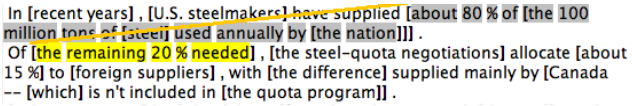

Figure 4: A pointer form a noun phrase to its converse

lows the marking of certain bridging relations, namely part-of, set membership, and a converse relation (Figure 4).

5. Preliminary experiments

Preliminary studies using the corpus to train and evaluate anaphora resolution systems were carried out at the 2007 Johns Hopkins summer workshop on natural language engineering. The system developed at the workshop treats anaphoric reference as an equivalence relation, so we created a new annotation level of markable sets, which included all the noun phrases which were either anaphors or antecedents. These markables were divided into equivalence sets derived from the original markable pointers; for ambiguous anaphors we just chose the first marked interpretation, assuming that this would be the most salient one. We also augmented the Wall Street Journal part of the corpus with additional texts from the Vieira and Poesio Corpus (Poesio and Vieira, 1998) and with texts from the RST discourse treebank annotated by prof. Kibrik's group. The composition of the extended corpus is shown in Table 2. We plan to test the corpus with the systems developed at the workshop.

6. Comparison with other corpora

The standard resources for evaluating anaphora resolution systems are the two corpora created as part of the Message Understanding Conference (MUC) and known as 'MUC-6' and 'MUC-7' corpora, and the series of corpora created as part of the Automatic Content Extraction (ACE) initiative. These corpora are all based on the MUCCS coding scheme (Hirschman and Chinchor, 1997), which was designed to study the use of coreference for information extraction and therefore incorporates a number of design decisions that are problematic from other perspectives, such as the decision to treat apposition and other forms of predication as well as identity as the same type of semantic relation. Also, these

Source	Texts	Markables		Words
		total[a]	coreferent[b]	
Arrau-WSJ	50	9177	3837	32771
VPC	35	8095	2844	31118
Moscow	40		2114	22529
Total[c]	105		7469	71612

[a]The anaphora resolution system only uses coreferent markables, so we did not extract all the markables from the Moscow corpus.
[b]Coreferent markables are those which participate in an anaphoric chain as either anaphor or antecedent.
[c]The total data is less than the sum of the individual components because two texts are annotated in both Arrau and VPC.

Table 2: Extended Arrau-WSJ corpus

corpora only mark coreference relations between mentions of a few specific types, such as people or organizations. The OntoNotes coreference corpus, a first part of which was released in May 2007 (Pradhan et al., 2007), overcomes many of the limitations of earlier efforts. All mentions are annotated, identity is distinguished from apposition, and references to some types of abstract objects – events – are annotated, although not the full form of discourse deixis. We are not aware of other efforts to create a corpus of anaphoric information in which ambiguity is allowed and discourse deixis is annotated. Also, to our knowledge, ARRAU is the only corpus in which additional information about grammatical function, agreement, and reference is marked.

7. Conclusions

The linguistic aspects of anaphora are not yet completely understood and as a consequence annotating anaphoric information is still an open problem; in addition, it is not completely clear the extent to which humans agree on the interpretation of such expressions. We believe these shortcomings of current work can only be addressed by carrying out further annotation experiments; this is the main justification of this annotation effort. Nevertheless, the resulting corpus is arguably the most complete resource currently available for the study of anaphora from a linguistic

perspective, and we believe that it may also be a valuable resource for evaluating application-oriented anaphoric resolvers.

8. Acknowledgments

This work was supported in part by EPSRC project GR/S76434/01, ARRAU, in part by a grant to the Center for Language and Speech Processing (CLSP) at Johns Hopkins University. The corpus was annotated by Laura Bostock, Tara Sutton, Josef Steinberger, Janet Hitzeman, Mijail Kabadjov, and Rhiannon Jones, together with the authors. We also wish to thank Prof. Andrej Kibrik and Olga Krasavina from the Russian Academy of Sciences in Moscow, and Christian Chiarcos from the University of Potsdam, with whom we are collaborating on annotating the RST discourse treebank; and Renata Vieira, for her part in the creation of the Vieira-Poesio corpus. Travel support for the second author comes from projects sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

9. References

Ron Artstein and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In David Schlangen and Raquel Fernández, editors, *brandial 2006: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 56–63, Potsdam, Germany, September.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech, and Language Technology*, chapter 5, pages 85–112. Kluwer, Dordrecht.

Wallace L. Chafe, editor. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 132–139, Seattle, April–May.

Derek Gross, James F. Allen, and David R. Traum. 1993. The Trains 91 dialogues. TRAINS Technical Note 92-1, University of Rochester Computer Science Department, July.

Peter A. Heeman and James Allen. 1995. The Trains 93 dialogues. TRAINS Technical Note 94-2, University of Rochester Computer Science Department, March.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 coreference task definition, version 3.0. In *MUC-7 Proceedings*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York, June. Association for Computational Linguistics.

Klaus Krippendorff, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 129–154. Sage, Beverly Hills, CA.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

C. Müller and M. Strube. 2006. Multi-level annotation of linguistic data with mmax2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, volume 3 of *English Corpus Linguistics*, pages 197–214. Peter Lang.

Massimo Poesio and Ron Artstein. 2005a. Annotating (anaphoric) ambiguity. In *Proceedings from the Corpus Linguistics Conference Series*, volume 1, Birmingham, England, July.

Massimo Poesio and Ron Artstein. 2005b. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Massimo Poesio and Natlia N. Modjeska. 2005. Focus, activation, and *this*-noun phrases: An empirical study. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, volume 263 of *Current Issues in Linguistic Theory*, pages 429–442. John Benjamins.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, June.

Massimo Poesio, Rodolfo Delmonte, Antonella Bristot, Luminita Chiran, and Sara Tonelli. 2004. The VENEX corpus of anaphoric information in spoken and written Italian. Manuscript. Available online at `http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf`.

Massimo Poesio, Uwe Reyle, and Rosemary Stevenson. 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*, pages 11–31. Springer.

Massimo Poesio. 2004a. Discourse annotation and semantic annotation in the GNOME corpus. In Bonnie Webber and Donna K. Byron, editors, *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79, Barcelona, July. Association for Computational Linguistics.

Massimo Poesio. 2004b. The MATE/GNOME proposals for anaphoric annotation, revisited. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge, Massachusetts, April-May. Association for Computational Linguistics.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Un-

restricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of ICSC 2007, International Conference on Semantic Computing*, pages 446–453, Irvine, CA.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In Sandra A. Thompson and William C. Mann, editors, *Discourse Description: Diverse Linguistic Analyses of a Fund-raising Text*, volume New Series 16 of *Pragmatics and Beyond*, pages 295–325. John Benjamins, Amterdam.

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Annie Zaenen. 2006. Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.