Issues in Anaphora Annotation: The Case of Encyclopaedic Animal Descriptions

Josef Meyer and Robert Dale
Language Technology Group, Macquarie University
{jmeyer|rdale}@ics.mq.edu.au

April 5, 2001

Abstract

Anaphora resolution has been a widely studied problem in natural language processing almost since work in the field began (see [Hirst 1981, Mitkov 1999] for summaries). However, most work focuses on the resolution of pronouns, or, at best, on a wider range of anaphoric expressions which co-refer with their antecedents. Real natural language texts exhibit a much richer range of anaphoric phenomena. In this paper, we describe an experiment in annotating a corpus for the variety of anaphors found, and indicate some of the problematic cases that robust natural language processing techniques have to deal with. As a result of this analysis we have identified several subtypes of associative anaphoric relationships that occur frequently in the corpus, and determined that it may be possible to identify the most common relationships using some fairly simple heuristics.

1 Introduction

This work is concerned with the development of strategies for anaphor resolution in real natural language texts. In particular, we are interested in the problems that arise in re-using small parts of a text outside of their original context, as is often required in systems which perform text summarisation [Paice 1990], answer extraction [Schwitter et al 1999], or in approaches to text generation which reuse text fragments mined from existing sources [Meyer and Dale 2000]. Consider the following example:

> *Caterpillar Hunter*, common name for a large, brightly colored, nocturnal, and predatory beetle (see GROUND BEETLE). Several species prey on caterpillars and earthworms. *One species* was imported into the United States in large numbers to combat the brown-tail moth.
>
> [Microsoft 1995]

Suppose that we had a system that extracted short segments from a text, and we wanted information on the introduction of predators to control pests. Only the last sentence of this text is relevant to this query; however, without the preceding reference to *Caterpillar Hunter*, a reader would be unable to work out that *one species* actually refers to one species of caterpillar hunter.

Existing work in anaphora resolution tends to focus only on pronominal reference, or on other forms of anaphora where the relationship between the referent of the anaphor and the referent of the antecedent is one of identity.[1] This paper reports on the annotation of a corpus of texts to make explicit a range of implicit anaphoric relationships between noun phrases. Developing the annotated corpus has two main purposes:

- to provide information on how frequently different types of anaphoric relationship occur in the corpus; and

[1] A notable exception is Vieira [1998], although even this work covers only a very small proportion of the types of anaphora we have observed in real texts.

⟨**S** id="s2"⟩
By ⟨**NP** id="2.1"⟩ 1958 ⟨**/NP**⟩
⟨**NP** id="2.2"⟩ the Havana brown ⟨**/NP**⟩
was firmly established
as ⟨**NP** id="2.3"⟩ a breed ⟨**/NP**⟩ . ⟨**/S**⟩
⟨**S** id="s3"⟩
⟨**NP** id="3.1"⟩ It ⟨**/NP**⟩
is short-haired and dark chocolate
in ⟨**NP** id="3.2"⟩ color ⟨**/NP**⟩ ,
with ⟨**NP** id="3.3"⟩ green, oval eyes ⟨**/NP**⟩ .
⟨**/S**⟩
⟨**S** id="s4"⟩
⟨**NP** id="4.1"⟩ The profile
of ⟨**NP** id="4.2"⟩ the head ⟨**/NP**⟩
⟨**/NP**⟩
is sloping and
⟨**NP** id="4.3"⟩ the tail ⟨**/NP**⟩
is long. ⟨**/S**⟩
⟨**S** id="s5"⟩
⟨**NP** id="5.1"⟩ The body ⟨**/NP**⟩
is muscular. ⟨**/S**⟩

Figure 1: An example of NP markup

⟨**BINDINGS**⟩
...

⟨**RE** np="2.2" de="topic020"⟩
the Havana brown ⟨**/RE**⟩
...
⟨**RE** np="3.1" de="topic020"⟩ It ⟨**/RE**⟩
...
⟨**RE** np="4.2" de="head017"⟩ the head ⟨**/RE**⟩
...
⟨**RE** np="4.3" de="tail018"⟩ the tail ⟨**/RE**⟩
⟨**RE** np="5.1" de="body019"⟩ The body ⟨**/RE**⟩
...
⟨**/BINDINGS**⟩
⟨**DE-RELS**⟩
⟨**REL** type="part/bpart" refs="topic020 head017"/⟩
⟨**REL** type="part/bpart" refs="topic020 tail018"/⟩
⟨**REL** type="part/bpart" refs="topic020 body019"/⟩
...
⟨**/DE-RELS**⟩

Figure 2: A fragment of reference structure

- to provide a resource which can be used in evaluating different anaphora resolution algorithms (or training, for statistical algorithms).

As a result of this analysis we have identified several subtypes of associative anaphoric relationships that occur frequently in the corpus, and determined that it may be possible to identify the most common relationships using some fairly simple heuristics.

Section 2 describes the approach we took to annotating a randomly selected 50 entries from Encarta [Microsoft 1995] and Grolier's encyclopedia [Groliers 1992]. Section 3 presents some preliminary findings, including frequencies for the commonest and most readily identified types of anaphoric relationships. Our experiment has uncovered some practical and theoretical issues that will be addressed in future work; these are discussed in Section 4.

2 Method

For this study we selected 50 entries from a collection of encyclopedia entries dealing with animals, 25 from Grolier's Encyclopedia [Groliers 1992], and 25 from Encarta [Microsoft 1995]. We then manually marked noun phrases within the corpus, assigning each an identifier; a fragment of the marked-up corpus is shown in Figure 1. Automatic identification of noun phrases is feasible, but early results were not sufficiently reliable for it to be of benefit in this study, and there are the obvious problems with ambiguity.

In the model of reference represented by this annotation scheme, noun phrases may be either referring or predicative. For example, the subject of the following sentence (*The African hunting dog, Lycaon pictus*) is referring and the object (*a wild CARNIVORE in the DOG family, Canidae, order Carnivora*) predicative:

> The African hunting dog, Lycaon pictus, is a wild CARNIVORE in the DOG family, Canidae, order Carnivora.

Each referring noun phrase specifies a DISCOURSE ENTITY; discourse entities can be connected via a number of different types of relationship, some of which form the basis for indirect anaphora. Each of

Type	Maximal	Total	% Total
Full NPs	940	1516	66.2
Full NPs followed by *'* or *'s*	0	11	0.0
Full NPs preceded by *of*	0	110	0.0
Full NPs (Possessive)	0	121	0.0
Pronouns (Neuter)	48	73	65.8
Pronouns (Feminine)	1	1	100.0
Pronouns (Possessive)	0	21	0.0
Pronouns (Non-possessive)	49	53	92.4
Pronouns (Total)	49	74	66.2

Table 1: Breakdown of NPs according to syntactic criteria.

the texts in our corpus has a fairly obvious TOPIC, which is generally the referent of the headword of the encyclopedia entry.

To encode these relationships, a separate file was used for representing the REFERENCE STRUCTURE of the texts. Each referring expression was associated with a key identifying that discourse entity which is its referent. All noun phrases were processed in a single pass. After all of the noun phrases had been associated with discourse entities, a second pass was made to identify and classify the relationships between discourse entities. The fragment in Figure 2 demonstrates a simplified version of this representation, where the ⟨BINDINGS¿⟩ structure indicates the mapping between referring expressions and discourse entities, and the ⟨DE-RELATIONS⟩ structure indicates the relationships between discourse entities. Annotating all of the relationships between referring expressions in a text would amount to something close to a full semantic analysis. This was not the objective of this work. Generally, only those relationships that were established as part of resolving an anaphoric noun phrase appear in the marked up text; so, for example, no relationship would be marked between subject and object in the sentence *The Havana brown has a long tail*. In addition to this, only relationships between the discourse referents of maximal noun phrases were considered; that is, those noun phrases embedded within other noun phrases were not considered. So, for example, in the noun phrase *the tip of its tail*, we have one discourse entity corresponding to the entire NP, but we do not include a distinct discourse entity for either *the tip* or *its tail*. Identifying relationships between DEs was the least reliable part of the process, for reasons outlined in Section 4.

3 Results

Table 1 gives a breakdown of the noun phrases in the text according to various syntactic criteria. As can be seen, maximal noun phrases account for only 66.2% of the noun phrases in the text; the other 33.8% are embedded within larger NPs. 92.4% of the non-possessive pronouns are maximal, with 7.6% appearing within larger NPs, as in the *the foot which it uses for cleaning*. Explicitly marked possession plays an important role in the corpus; 7.9% of the total number of noun phrases in the corpus are possessives, as in *its foot* and *the cat's tail*, and these could conceivably be handled by an extension to a straightforward coreference resolution mechanism, whereas forms such as *the foot* or *the tail* clearly require a more complex mechanism.

Table 2 gives information on how many times discourse entities are referred to. Out of a total of 655 discourse entities identified in the annotated corpus, 568 (86.7%) were only referred to once; this accounts for 65.2% of the 870 maximal noun phrases that were identified as referring rather than predicative. The topic is referred to an average of 4.33 times in each text, while other discourse referents (excluding the topic for each entry) occur an average of 1.09 times. Out of the 50 texts, there are only two in which the

Num. references:	1	2	3	4	5	6	7	8	9	10
Count of DEs	568	37	14	18	7	6	2	0	1	2
% of DEs:	86.7	5.6	2.1	2.7	1.0	0.9	0.3	0.0	0.1	0.3
% of maximal REs:	65.2	8.5	4.8	8.2	4.0	4.1	1.6	0.0	1.0	2.2
Topic DEs:	2	7	6	15	7	6	2	0	1	2
Other DEs:	566	30	8	3	0	0	0	0	0	0

Table 2: Distribution of the number of times a discourse entity is referred to using a maximal noun phrase.

topic is referred to only once.

Table 3 shows the frequency of the labels used in identifying types of anaphoric relationships between DEs that are the referents of maximal NPs in the corpus. As can be seen, of the marked anaphoric relationships, 96% have been classified. The relationships are arranged into a hierarchical structure. The hierarchy of tag types provides a natural way of clustering similar types of relationship to provide a more robust classification. The major types of relationships are as follows:

part: The most common type of relationship in the corpus is that between an animal and its parts. This accounts for 45.6% of the relationships. The following example shows the kinds of NPs that fall into this category; note that only the first NP uses an explicit anaphor, whereas the others are instances of what is generally referred to as associative anaphora.

> *Its legs* are white; *the hair* is short and sleek, *the limbs* slender, and *the muzzle* sharp; *the back* rises into a high arch immediately behind *the neck*.

subtype: The next most common type is the *subtype* relationship, which accounts for 24.9% of the relationships; this is used when referring to a particular subtype of a recently mentioned animal. The most common type of subtype relationship is based on common properties, like sex and age. The following example demonstrates again that the anaphoric relationship is implicit:

> *The male* is dark brown to black; *the female* is brown to reddish brown.

name: The relationship between an entity and the name for that entity is a slightly more complicated case; the most obvious example of this type of reference is *the name*, as in the following example:

> *The name* (meaning "double beam") alludes to the peculiar construction of certain tailbones, which had projections fore and aft and which protected blood vessels from friction with the ground.

These references combined make up 9.1% of the relationships.

attrib: The entity/attribute relationship accounts for 4.1% of the relationships in the corpus, as in the following example:

> They can live for more than 20 years in captivity; *life expectancy in the wild* is probably shorter.

Physical attributes (*attrib/phys*) consist of things like size and colour, developmental attributes (*attrib/devel*) consist of things like lifespan, cultural attributes (*attrib/cult*) things like beauty, and so on. The finer distinctions in this case are largely irrelevant, except in that not all attributes seem to appear without a possessive.

The noteworthy feature of all of these categories is that the type of relationship seems to be inferable from the ontological category of the anaphor.

Relationship name	Number	% Total
/attrib:	11	4.6
/attrib/behaviour:	2	0.8
/attrib/cult:	1	0.4
/attrib/devel:	3	1.2
/attrib/devel/base:	1	0.4
/attrib/devel/lifespan:	2	0.8
/attrib/devel/lifespan/wild:	1	0.4
/attrib/devel/lifespan/captive:	1	0.4
/attrib/phys:	8	3.3
/attrib/phys/base:	4	1.7
/attrib/phys/shape:	2	0.8
/attrib/phys/shape/base:	1	0.4
/attrib/phys/shape/build:	1	0.4
/attrib/phys/size/length:	1	0.4
/attrib/phys/colour:	1	0.4
/attrib/sense:	1	0.4
/disjoint:	2	0.8
/group:	2	0.8
/ident:	2	0.8
/ident/base:	1	0.4
/ident/age:	1	0.4
/instance:	3	1.2
/instance/base:	2	0.8
/instance/specific:	1	0.4
/member:	2	0.8
/member/base:	1	0.4
/member/group:	1	0.4
/name:	22	9.1
/name/base:	15	6.2
/name/common:	4	1.7
/name/science:	3	1.2

Relationship name	Number	% Total
/part:	110	45.6
/part/base:	1	0.4
/part/bpart:	108	44.8
/part/bpart/base:	107	44.4
/part/bpart/product:	1	0.4
/part/misc:	1	0.4
/partition:	4	1.7
/possible:	2	0.8
/possible/role/job:	1	0.4
/possible/association/misc:	1	0.4
/subset:	4	1.7
/subtype:	60	24.9
/subtype/base:	4	1.7
/subtype/misc:	10	4.1
/subtype/misc/base:	4	1.7
/subtype/misc/range:	1	0.4
/subtype/misc/taxon:	5	2.1
/subtype/property:	34	14.1
/subtype/property/[captivity]:	2	0.8
/subtype/property/[sex,age]:	1	0.4
/subtype/property/[sex,captivity]:	1	0.4
/subtype/property/age:	12	4.9
/subtype/property/breed:	2	0.8
/subtype/property/colour:	1	0.4
/subtype/property/misc:	1	0.4
/subtype/property/sex:	13	5.4
/subtype/property/sex+:	1	0.4
/subtype/taxon:	11	4.6
/subtype/temporal:	1	0.4
/other:	12	4.9
/other/ancestor:	1	0.4
/other/base:	7	2.9
/other/comparator:	1	0.4
/other/embed:	3	1.2
/other/embed/base:	2	0.8
/other/embed/subtype/taxon:	1	0.4
total:	241	100.0

Table 3: Raw frequencies of the different types of relationship that hold between top-level Research in the corpus.

4 Discussion

One of the aims of the work described here was to develop and refine an annotation process that can be applied to a wide range of texts. The work has highlighted a variety of methodological issues relating to reliability and standardisation, and theoretical issues resulting from the differences between the domain of the texts in our corpus and those of corpora used in previous studies. This section singles out some of the many problems that remain to be addressed, particularly with regard to processing texts from this genre.

4.1 Methodological issues

4.1.1 Reliability

The classification of anaphoric relationships used in this study needs to be related to developing standards for marking up discourse relations. The markup scheme currently in use has some similarities to the annotation standard developed by the MATE group Davies et al [1998], so this should be feasible.

Some measure of the reliability of the annotation is needed; for the present work, all annotation was carried out by one of the authors. We intend to carry out annotation experiments with a number of independent subjects, and we expect the results of this to be useful in refining the classification of different types of anaphoric relationships. Measuring the reliability of annotation, however, is not a trivial problem, as discussed in Vieira [1998].

4.1.2 Non-maximal noun phrases

There were a number of cases in which the relationship between maximal noun phrases was in fact determined by the relationship between embedded noun phrases. In the case of possessives indicating body part and attribute relationships, this actually simplified matters. However, in other cases it made it almost impossible to classify the type of relationship that existed between discourse entities:

⟨**S** id=”s5”⟩
⟨**NP** id=”5.1”⟩ Young monotremes ⟨**/NP**⟩
do not have
⟨**NP** id=”5.2”⟩ mouth parts suitable for suckling ⟨**/NP**⟩ ;
⟨**NP** id=”5.3”⟩
the liquid produced
by ⟨**NP** id=”5.4”⟩ the nippleless mammary organ ⟨**/NP**⟩
⟨**/NP**⟩
is licked from
⟨**NP** id=”5.5”⟩
the belly hair
of ⟨**NP** id=”5.6”⟩ the mother ⟨**/NP**⟩
⟨**/NP**⟩ .
⟨**/S**⟩

In this example, there is a relationship between the referent of NP 5.3 and that of NP 5.4 that would be most easily explained in terms of body part relationships between the referent of NP 5.6 and the referents of NPs 5.5 and 5.4.

4.2 Theoretical issues

4.2.1 Difficult cases

One example that was particularly difficult to deal with was the following:

⟨**S** id=”s4”⟩
After
⟨**NP** id=”4.1”⟩ monotreme eggs ⟨**/NP**⟩
are hatched,
⟨**NP** id=”4.2”⟩ the young ⟨**/NP**⟩
are helpless, and, in
⟨**NP** id=”4.3”⟩
the case of
⟨**NP** id=”4.4”⟩ the echidna ⟨**/NP**⟩
⟨**/NP**⟩ ,
are carried in
⟨**NP** id=”4.5”⟩ shallow abdominal pouches ⟨**/NP**⟩ .
⟨**/S**⟩

For a start, there is the issue of the null subject following NP 4.3. It is fairly clear that in this case the first and second verb phrases should not have the same entity in the role of patient. This either requires us to adopt a position in which a noun phrase can simultaneously have two referents, or equivalently, that this sort of conjunction involves a trace that does not necessarily have the same referent as the expression that introduces it. There is also the problem of determining a referent for NP 4.3 itself, if it is indeed a referring expression.

A second category of problem case occurs where relationships between subtypes and their supertypes are not always distinct. In the case of the English foxhound, for example, there are references throughout the text to not only *the English foxhound* and *the foxhound*, but also to *the American foxhound*. The relationship between the three is never explicitly stated, so a given instance of *the foxhound* could be ambiguous between the English foxhound, the American foxhound, and the supertype. A similar problem occurs with the entry on German shepherds, where a reference is made to both *today’s German shepherd* and its ancestors, making an exact determination of some referents, and the relationships between referents, difficult. In both cases the text can be understood without resolving the ambiguity. So we are left with the choice of either picking one option at random (the solution adopted in this study), or finding some way to represent the ambiguity within the markup.

4.2.2 General issues

When bridging references are taken into account, it is possible for the chain of inferences that link anaphor to antecedent to involve more than just the referents of these phrases. Take the following sentence from the entry on barbary apes:

⟨**S id=”s8”**⟩
⟨**NP id=”8.1”**⟩
The life span
in ⟨**NP id=”8.2”**⟩ captivity ⟨**/NP**⟩
⟨**/NP**⟩
is ⟨**NP id=”8.3”**⟩ more than 30 years ⟨**/NP**⟩ .
⟨**/S**⟩

One way to handle NP 8.1 is to posit the existence of an unexpressed discourse entity representing a (generic) captive barbary ape. As with all animals, this would have an associated life-span; by introducing this unexpressed DE, we would achieve some degree of consistency in representing the relationships between DEs. An alternative strategy is to increase the number of possible relationships between DEs to include something like *property-in-captivity*.

5 Conclusions and Further Work

This paper discusses the results of a first pass at classifying the anaphoric relationships occurring within a corpus of encyclopedia entries. The results show that certain kinds of associative anaphoric relationship occur particularly frequently in this corpus; more frequently than might be expected in other types of text. This suggests that for processing texts in this particular class, it should be possible to achieve better results by focusing on these types of relationship rather than using the fairly non-specific approaches that have previously been used in dealing with associative anaphora in unrestricted texts [Vieira 1998].

The process of annotating the anaphoric relationships within this corpus has raised a variety of interesting questions about the way that reference functions within texts that describe types of entity rather than focusing on specific individuals. These questions provide the basis for further work on the nature and function of reference, and more practically, on the representation of anaphoric relationships in annotated texts.

The next logical step to take, after some of the representational issues raised in Section 4 are clarified, is to annotate another corpus of texts using this scheme (and an overlapping set of categories). This is necessary not only to confirm that the annotation scheme is robust, but also to provide something against which the figures in this report can be compared.

References

Sarah Davies, Massimo Poesio, Florence Bruneseaux, and Laurent Romary. Annotating coreference in dialogues: proposal for a scheme for MATE. First draft, available as `http://www.cogsci.ed.ac.uk/~poesio/MATE/anno_manual.html`, July 1998.

Groliers (1992). *The New Grolier Multimedia Encyclopedia*. Copyright Grolier Incorporated, (c) 1987-1992 Online Computer Systems, Inc.

Graeme Hirst (1981). *Anaphora in natural language understanding: A survey*. Springer-Verlag, New York.

J. Meyer and R. Dale (2000). Building Hybrid Knowledge Representations from Text. Pages 158–165. In *Proceedings of the 23rd Australasian Computer Science Conference (ACSC2000)*. February. Canberra, Australia.

Microsoft (1995). *Microsoft (R) Encarta'95 Encyclopedia*. Copyright (c) 1994 Microsoft Corporation. Copyright (c) 1994 Funk and Wagnall's Corporation.

Maria Milosavljevic (1997). Augmenting the User's Knowledge via Comparison. Pages 119–130. In *Proceedings of the 6th International Conference on User Modelling*. 2–5 June 1997. Sardinia.

Ruslan Mitkov (1999). Anaphora resolution: the state of the art. Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution). University of Wolverhampton, Wolverhampton. Available as `http://www.wlv.ac.uk/~le1825/anaphora_resolution_papers/state.ps`

Chris D. Paice (1990). Constructing literature abstracts by computer: techniques and prospects. In *Information Processing and Management*, 26 (1), pp. 171–186.

Rolf Schwitter, Diego Mollá Aliod and Michael Hess (1999). ExtrAns — Answer Extraction from Technical Documents by Minimal Logical Forms and Selective Highlighting. Presented at *The Third International Tbilisi Symposium on Language, Logic and Computation*. September 12-16, 1999. Batumi, Georgia.

Renata Vieira (1998). *Definite description processing in unrestricted text*. PhD thesis, University of Edinburgh.