Towards Reliable Partial Anaphora Resolution

Sabine Bergler *

Computer Science Department, Concordia University

1455 de Maisonneuve Blvd. W., Montréal, Québec, H3G 1M8

bergler@cs.concordia.ca

Abstract

This paper assumes that currently, anaphora resolution at a desired level of reliability has to remain partial. It presents the thesis that multiple small ("expert") procedures of known reliability that are conceived for partial analysis have to be developed and combined in order to increase coverage. These resolution experts will be specific to style, domain, task, etc. The paper describes corpus analysis that suggests such experts and their potential ordering. A quick and partial implementation of the ideas is evaluated on Wall Street Journal articles.

Introduction

Totally correct anaphora resolution requires full natural language understanding, since anaphoric relations could be hidden in the context. At present, only partial natural language understanding is possible. This paper claims that one way to increase the reliability (or at least in assessing the reliability) of anaphora resolution lies in acknowledging and making use of this limitation.

Strategies of anaphora resolution depend on the genre and style of text under consideration, as the different style manuals for major newspapers show. Since many practical applications are limited to a certain genre, it is legitimate to optimize results by studying peculiarities of the genre.

We focus our attention on the Wall Street Journal corpus available on CD-ROM from the Association for Computational Linguistics. Our main interest at the outset was in assessing the lexical complexity of NP coreference to guide us in our development of a lexicon. We reported initial corpus analysis results that show the relative frequency of semantic relations that hold between elements in coreference chains [Bergler and Knoll, 1996]. Analyzing the reference chains[1] of 79 articles (28,798 words) from the Wall Street Journal we found that 35% of the subsequent references are actually equal to the first reference of that entity, 23% are close variations of the first reference (i.e. retain at least the same headword). Pronouns and appositions account for 22% and systematic lexical relation (synonymy and hyponymy) for 7%. We consider the remaining 13% to be tough cases that might require full syntactic, lexical, and semantic processing. The other 87% we expect can be addressed with a subset of these tools.

This paper presents further results of this corpus analysis which lead to some resolution strategies and presents an experiment in implementing some of these aspects in a knowledge poor system.

Corpus Analysis Results

The corpus study of 79 articles from the Wall Street Journal was performed manually by a single analyst, thus the results are as consistent as possible for a manual analysis. The analyst separated all NPs from the text, when appropriate separating referring sub-NPs from larger NPs, and for each NP placed it in a chain that contained a coreferent when possible or started a new chain. Chains were additionally annotated for a few features of interest, such as length of the article (*short* or *long*), the textual designator (see below), and whether the chain was part of the topic of the article (see next Section.) These annotations were ultimately left to the judgement of the analyst within a strict set of rules.

Particularly encouraging are the first two lines in Figure 1. It turns out that over a third of the coreferring NPs are identical and can therefore be recovered reliably and correctly without linguistic tools. Almost a quarter of the coreferring NPs are very close to the first NP in the reference chain, that is they share at least the headword, if not a larger substring with the first reference. Thus in theory almost 60% of coreferring NPs should be identifiable with very simple techniques once the NPs have been identified.

To put things into perspective, let us reconsider the numbers in Figure 1 with some additional information.

*This work is funded in part by the Natural Sciences and Engineering Research Council of Canada and Fonds pour la formation de chercheurs et l'aide à la recherche.

[1] Reference chains in this study contain all (partial) noun phrases that corefer in a text. They are thus different from [Morris and Hirst, 1991], who do not limit their reference chains to NPs.

Semantic relation	Total	Percent
Equal to first reference	1424	35%
Close to first reference	955	23%
Appositions	128	3%
Pronouns	756	19%
Acronyms	46	1%
Synonyms	73	2%
Hypernyms	174	4%
Neither of the above	520	13%
Total NPs	4076	100%

Figure 1: Semantic relations between elements in coreference chains

The 4076 NPs analyzed there constitute roughly half of the NPs counted in total in that collection of texts, namely 8,027. The 3,951 NPs that are not analyzed in Figure 1 are NPs that do not corefer with any other NP in the text. These singular occurrences account for 49% of all NPs. One obvious question is: can singular occurrences of entities be singled out? This is an open question. We address the easier problem of: how can we determine NPs which are likely to corefer and which are most important to the text overall?

Topicality

The same study showed that NP chains considered to be in the topic of an article usually require anaphora resolution and are lexically more complex than non-topical reference chains. For this study we defined a topic to be one of the NPs that occur in the headline or the first sentence[2] (see [Lundquist, 1989] for a motivation of this heuristic.) A text can have no more than 4 topics (this number was chosen intuitively.) The analyst decided how many topics there were in each article according to her understanding of it. The text in Figure 3 was assigned a single topical chain containing NPs 1, 2.1, 10, and 11. There are 185 topical chains in 79 articles, averaging 2.4 topics per article. 17% of the topical reference chains are singular occurrences, i.e. NPs that do not corefer.[3] That establishes that the topic of an article is usually referred to more than once. Intuitively we assume that the topic of an artice is more important to resolve than non-topical NPs. One partial strategy, consequently, is to establish potentially topical NPs in the first n sentences of a newspaper article and to resolve coreference only to these NPs. This strategy has the advantage of reducing the search space considerably and of focusing on important (topical) chains.

[2] The heuristic to find potential topics of an article is easy to implement: consider all NPs in the headline and the first sentence.

[3] There are two possible explanations why this number is relatively high: the Wall Street Journal contains several very short segments of very few sentences, thus a topic that is mentioned only once is a possibility. Also, headlines often use a summarizing term that never corefers with another NP, but rather corefers with the article as a whole.

Textual Designators

We feel that NP resolution in newspaper articles is straightforward (by design) for human readers because of recurrent terms that designate an entity. To assess this intuition quantitatively, [Bergler and Knoll, 1996] report the sum of distinct words within a chain over all non-singular chains. This count eliminates the recurrent words, whether empty (*the*) or descriptive (*company*).[4] The results show the significance of this phenomenon which we call *textual designator*. A textual designator in this study is the first non-pronomial reference to an entity. Consider the text in Figure 3. One chain consists of two identical NPs, *its Houston work force* (NPs 3 and 15.) Other textual designators of that text are NP1, NP6, and NP16. Counting the number of different words in each chain allows us to assess the lexical diversity within a chain and the contribution of the textual designator to that diversity. The results are summarized in Figure 2.

Chain type	Total chains	Non-singular chains	Excluding designator
Topic	1,632	1,544	981
Not topic	19,449	5,672	2,808

Figure 2: Number of different words per reference chain.

We find 1,632 different words in topical chains and 19,449 different words in non-topical chains. When we consider only the chains that actually involve coreference, the sums reduce to 1,544 different words for topical chains and 5,672 different words for non-topical chains. Removing the words that are part of the textual designator, we observe a drastic reduction to 981 different words for topical chains and 2,808 different words for non-topical chains.

The 185 topical chains average about 9 different words per chain while non-topical chains average 4. Removing the words of the textual designator reduces the different word count by 40% for topical and 86% for non-topical chains when counting all chains. If we count only chains that involve coreference the reductions are 36% for topical chains and 50% for non-topical chains. The number of different words excluding the words of the textual designator on average are 5 for topical chains and .6 for non-topical chains.[5]

These numbers suggest that the textual designator defined as the first reference to an entity leads to a strong resolution heuristic, one that is in fact stronger for non-topical chains. The numbers also show how surprisingly small the lexical diversity of words outside

[4] Note that the same word could potentially be counted in every single chain, as the recurrence is only eliminated within a chain.

[5] This average includes singular chains.

a textual designator are. The sum for both types of chains is 3,789. The textual designators are made up of 563 different words for topical non-singular chains and 2864 different words for non-topical, non-singular chains. The total number of different words for singular chains is 13,865. Thus the sum total of different words for first references is 17,292, a surprisingly high number considering that the overall corpus has only 28,798 words, which includes all duplicates.

Resolution Strategies

While we are still mining the results of our study for more data, some strategies for resolution in general are already emerging.

We believe that there is strong evidence for an approach to anaphora resolution in multiple passes, where earlier passes implement more reliable[6], less knowledge-intensive, and computationally less complex strategies with faster tools than later ones. For each pass, an expected reliability should ideally be known.

The expectation is that early, knowledge-poor, and highly reliable passes can be used for almost any task. Matching equal NPs, for instance, can be done with basic, fast tools independently of further linguistic processing. Anaphora resolution could then be tailored to particular needs by determining which levels of reliability are acceptable to the task and using the passes[7] up to that threshold. For the remaining resolution task, a domain- and genre-specific set of procedures has to be developed.

We argue that such a multi-pass approach has advantages to a monolithic approach, be it statistical or symbolic. While most symbolic anaphora resolution systems probably correctly identify identical NPs as coreferring, making this a first step and using very fast, low level tools can pre-process a text faster. The modular approach allows for use of the tools of choice at each level. Moreover, a text can be left partly resolved, potentially allowing anaphora resolution to be interleaved with other linguistic processing as required.

Another interesting result of our study is the fact that many NPs correctly resolve to more than one coreferring NP, and often resolve to the first reference. This provides support for the viability of partial parsing methods, because a missing link does not mean the rest of the chain is unresolvable. As mentioned above, this also allows for a focused partial resolution strategy that attempts to resolve subsequent NPs only to a set of NPs determined at the outset (e.g., topical NPs, predetermined subjects or persons, ...). This provides the basis for a series of principled heuristics. These partial resolution strategies are of great importance where the amount of text to be processed is large but the depth of processing is shallow, as for certain text annotation tasks.

The study presented above suggests the following strategies:

1. **Identical NPs**
   Prerequisites: NP boundaries
   Procedure: string matching
2. **Focused partial resolution**
   Prerequisites: NP boundaries, identification procedures of the NP chains of interest
   Procedure: according to desired resolution strategies
3. **Common head**
   Prerequisites: parsed NPs
   Procedure: matching head positions
4. **Appositions**
   Prerequisites: parsed NPs
   Procedure: matching syntactic pattern of apposition
5. **Extended head matching**
   Prerequisites: parsed NPs, lexicon or thesaurus
   Procedure: compare heads for synonymy, hypernomy
6. **Pronoun resolution**
   Prerequisites: parsed text
   Procedure: as described in the literature

These resolution strategies are not exhaustive, nor can an optimal ordering be assigned in general; the desired level of reliablity, the genre and style features, and any possible additional linguistic processing will determine different combinations, extensions, and orderings.

Knowledge Poor Resolution

We implemented an experimental resolution system (ERS) based on these ideas. The system uses as input the parse trees provided by the Penn Treebank on the ACL CD-ROM. These parse trees have been corrected manually for inconsistencies.

The system includes a (partial) implementation of all of the six strategies except for Focused Partial Resolution. The most carefully worked out strategy is pronoun resolution. Pronoun resolution makes use of the parse trees and follows the ideas in [Lappin and Leass, 1994, Hobbs, 1978]. The other strategies have only been partially or crudely implemented. The Common Head strategy, for instance, uses a crude heuristic to determine the head of a complex noun phrase that fails in certain cases. Extended Head Matching is limited to very few lexical items such as *company*, which receive special treatment. No lexicon is used, the required lexical knowledge has been provided in a list of gendered items. An additional limitation is the fact that the system considers coreference only within a sentence and between adjacent sentences.

Algorithm

For every NP in the text:

[6] Reliability here is with respect to errors of commission.

[7] An appropriate subset of the following heuristics can of course be combined into a single pass, which is the case in our experimental system presented below.

$_{(1}$ Telxon Corp. $_{1)}$ said $_{(2}$ $_{(2.1\ll\mathtt{ref=1}}$ its $_{2.1)}$ vice president for manufacturing $_{2)}$ resigned and $_{(3}$ $_{(3.1<\mathtt{ref=1}}$ its $_{3.1)}$ Houston work force $_{3)}$ has been trimmed by $_{(4}$ 40 people $_{4)}$, or about $_{(5<\mathtt{ref=4}}$ 15% $_{5)}$.
$_{(6<\mathtt{ref=1}}$ The maker of hand-held computers and computer systems $_{6)}$ said $_{(7}$ the personnel changes $_{7)}$ were needed to improve $_{(8}$ the efficiency $_{8)}$ of $_{(9}$ $_{(9.1\ll\mathtt{ref=6}}$ its $_{9.1)}$ manufacturing operation $_{9)}$.
$_{(10\ll\mathtt{ref=1}}$ The company $_{10)}$ said $_{(11\ll\mathtt{ref=10}}$ it $_{11)}$ hasn't named $_{(12}$ a successor $_{12)}$ to $_{(13>\mathtt{ref=14}}$ Ronald Bufton $_{13)}$, $_{(14<\mathtt{ref=2}}$ the vice president $_{14)}$ who resigned. $_{(15\ll\mathtt{ref=3}}$ $_{(15.1<\mathtt{ref=1}}$ Its $_{15.1)}$ Houston work force $_{15)}$ now totals $_{(16}$ 230 $_{16)}$.

Figure 3: Manually determined coreference in a short text from the Wall Street Journal

1. Determine candidate referents within the sentence. If none are found (i.e. lack of agreement), determine candidate referents in previous sentence.
2. Test each candidate referent for actual coreference using:
   (a) Common Head (with slight modifications)
   (b) Extended Head Matching (limited to few cases)
   (c) Appositions
   (d) Copula
3. If there is more than one possible coreference, select best.
4. Merge the new coreference pair with existing reference chains or start a new chain.

Sample Output

This algorithms is clearly too constrained to ever achieve full resolution, but except for the pronoun resolution, it was quickly implemented and performs surprisingly well.

Both strengths and limitations are best illustrated on a short example. Consider the text in Figure 3, which has been annotated with manually determined coreference links following the Lancaster notation [Fligelstone, 1992] with slight modifications

The annotation $_{(2.1<\mathtt{ref=1}}$ means that NP2.1 (a sub-NP of NP2) starts at this point and that it refers backwards to NP1. The $\ll$ sign indicates that this reference has also been detected by ERS.

ERS determined 4 reference chains in this article. The first chain consists of NPs 1, 2.1, 10, and 11. The second chain contains NPs 6 and 9.1, the third contains NPs 3 and 15, and chain number four contains NPs 16 and 15.1. The coreference link stipulated for chain four is wrong (an artifact of a strong bias towards intrasentential resolution.) All other stipulated coreference links are correct. There are six coreferring NPs whose coreference link has not been identified. Two of the stipulated chains could be merged (chains one and two.)

Preliminary Results

We evaluated ERS on a set of twelve short articles from the Wall Street Journal which had also been part of the corpus study described above. The chains of the study were compared to the chains stipulated by ERS and both were examined for correctness by a person not involved in the development of either.

Text	Sent.	NPs	Wrong		Omitted	
I	11	61	4	7%	13	21%
II	4	20	1	5%	6	30%
III	14	85	2	2%	20	24%
IV	8	53	1	2%	17	32%
V	4	14	0	0%	2	14%
VI	3	15	0	0%	1	7%
VII	3	17	1	6%	4	24%
VIII	2	13	0	0%	3	23%
IX	3	16	1	6%	3	19%
X	7	30	0	0%	5	17%
XI	7	38	16	42%	13	34%
XII	10	49	20	41%	18	37%
Σ	76	411	46	11%	105	26%

Figure 4: Evaluation of NP coreference

Figure 4 shows the length of the articles in number of sentences, the number of NPs per article, the number of NPs that have been placed in the wrong chain (**wrong**) and the number of coreference links that have not been identified (**omitted**.) ERS showed an error rate of 11% (wrongly dereferenced NPs) and an omission rate of 26%. Interestingly, the error rates for individual articles fluctuates between 0% and 7% for all but two articles. The two articles with over 40% error rate are both very typical Wall Street Journal articles that cover earnings of a particular company. The articles are very similar and the high error rate is due to the same shortcoming of ERS: it has no special treatment for amount terms and currency terms and thus creates chains that contain all NPs with headword *yen* or *million* (in these articles about Japanese companies *18.32 billion yen* are also expressed as *$128.9 million*.) Since none of these NPs should corefer, the error rate is extremely high but could be fixed with shallow lexical knowledge.

Before analyzing the performance of the heuristics in the evaluation, consider the results for reference chains. The twelve articles have 81 (non-singular) reference chains, ERS stipulates 50. 35 of the 81 chains have not been established by ERS (43%), while 17 (21%) contain NPs that do not corefer with any member of the chain. 14 chains (17%) were correct (agreement between study, ERS, and verification), another 19 chains were correct but incomplete (24%). Thus 33 chains were established correctly and contained no errors, this amounts to a 41% accuracy rate for chains. This figure is extremely low compared to the NP coreference resolution result because in the case of an incorrect coreference link in

Text	Real	ERS	Correct	Partly	Wrong	Omitted
I	10	7	3	1	3	3
II	4	4	2	1	1	2
III	18	10	4	4	2	8
IV	10	6	1	4	1	5
V	1	0	0	0	0	1
VI	1	1	0	1	0	0
VII	4	1	0	0	1	3
VIII	2	1	0	1	0	1
IX	4	3	1	1	1	1
X	4	4	1	3	0	1
XI	10	6	0	2	4	4
XII	13	7	2	1	4	6
Σ	81	50	14	19	17	35

Figure 5: Evaluation of reference chains

a chain the entire chain is discounted. Correct chains are those were ERS picked up all the NPs that the analysts determined as belonging to that chain. Partly correct chains are those were every coreference link is correct but some are missing. Wrong chains are those that include an incorrect coreference link and omitted chains are those that the analysts determined to exist but which have no counterpart in ERS's output.[8]

Causes of Errors

An important first observation is that the human analysts did not agree in all cases, which reminds us to take figures and percentages with a grain of salt. Disagreements between the analysts are of course few. The majority of ERS's errors stem from three sources: lack of a lexicon, lack of syntactic finesse, and simplistic head matching for the Common Head heuristic.

The lack of any lexicon tricks ERS into matching all amounts measured in *yen* as possessing the same head and thus coreferring. A lexicon could also avoid pronoun agreement mistakes to a larger extent.

ERS lacks syntactic finesse even though it uses the rather sophisticated Penn Treebank parse trees, because its extraction of the head is based on heuristics rather than syntactic knowledge.

Most mistakes, by far, are due to matching identical heads where the modifying information in the NP makes it clear that no coreference exists, such as *an issue price of $849* ... and *a conversion price of $25* .... These, as well as omissions, would be reduced by considering the compositional semantics of the NPs. Omissions would also be reduced by implementing the resolution strategies fully (acronyms, hyponymic relations of head nouns, etc.)

[8] Note that the columns for correct, partly correct, wrong, and omitted chains do not add up to the number of real chains. This is due to several factors. Take for instance Text II, given in Figure 3, where two partly correct chains should have been merged and ERS's incorrect chain does not correspond to any real chain, thus there is an overcount of two chains. For the entire test set there is an overcount of 4 chains from three articles.

Conclusion

This paper reviews results from previous corpus analysis and presents new data that show that simple resolution procedures based on lexical similarity can achieve partial anaphora resolution. We advocate a multipass approach, sequencing appropriate simple resolution strategies for any given application. These passes can be interleaved with other linguistic processing and thus afford more flexibility than a monolithic anaphora resolution strategy.

We have tested these ideas with a crude experimental system that had an error rate of 11% and an omission rate of 26%. This result confirms the estimate from our corpus analysis and error analysis indicates refinements required for a robust system.

Acknowledgements

The manual corpus analysis was done by Sonja Knoll. ERS has been implemented by Dr. Xiaobin Li. Jennifer Scott did the evaluation. Thanks also to the anonymous reviewers, whose comments helped to improve the paper.

References

[Bergler and Knoll, 1996] S. Bergler and S. Knoll. Coreference patterns in the wall street journal. In C. Percy, C.F. Meyer, and I. Lancashire, editors, *Synchronic corpus linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*. Rodopi, Amsterdam, 1996.

[Fligelstone, 1992] S. Fligelstone. Developing a scheme for annotating text to show anaphoric relations. In G. Leitner, editor, *New Directions in English Language Corpora*. Mouton deGruyter, Berlin, 1992.

[Hobbs, 1978] J.R. Hobbs. Resolving pronoun references. *Lingua*, 44:311–338, 1978.

[Lappin and Leass, 1994] S. Lappin and H.J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[Lundquist, 1989] L. Lundquist. Modality and text constitution. In M-E. Conte, J.S. Petöfi, and E. Sözer, editors, *Text and Discourse Connectedness. Proceedings of the Conference on Connexity and Coherence, Urbino, July 1984*. John Benjamins Publishing Co., Amsterdam, 1989.

[Morris and Hirst, 1991] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.