

Resolving Pronouns Robustly: Plumbing the Depths of Shallowness

Advait Siddharthan

Natural Language and Information Processing Group
Computer Laboratory, University of Cambridge
as372@cl.cam.ac.uk

Abstract

In this paper, we explore various shallow techniques for salience-based pronoun resolution. We use simple noun chunking at the syntactic analysis stage and extract grammatical function information by pattern matching. Identifying subjects and objects is critical to salience calculations. We report that this important subject-object distinction can be made reliably with our shallow approach. We then explore a range of shallow semantic inference procedures (making use of only the text under consideration and lexical knowledge bases like WordNet) that significantly improve third-person pronoun resolution across a range of genre. We show how the same framework can also resolve relative pronouns with high accuracy. This is a result that might be relevant to other NLP applications like parsing.

1 Introduction

Pronoun resolution systems need to take a range of factors, both syntactic and semantic, into account. Most algorithms do this in stages, by first identifying possible antecedents, then applying a set of filters to rule out some of them and finally applying a decision procedure to select one of the remaining candidates. For example, salience based algorithms (Lappin and Leass, 1994; Kennedy and Boguraev, 1996) first calculate salience scores for potential antecedents based on their syntactic roles and recency, then apply a set of semantic and syntactic filters to rule out potential antecedents and

finally attach the pronoun to the most salient remaining potential antecedent.

However, the performance of such systems appears to plateau at around 60-65% on unrestricted text (Barbu and Mitkov, 2001; Preiss, 2002). It appears that weights for various salience features, trained to give high performance on particular genre, need to be retrained to work on other genre. However, there remains a strong preference for antecedents that are subjects, and to a lesser extent direct objects, across genre.

In section 2, we show how this crucial subject-object distinction can be made reliably using pattern matching on chunked text. This is a level of processing that is even shallower than that used by Kennedy and Boguraev (1996), and guarantees an analysis for every sentence, with a computational complexity that is linear in sentence length.

Anaphora resolution algorithms need to fall back on more elaborate inference mechanisms when salience alone does not return a reliable answer. Unfortunately, knowledge-intensive approaches do not scale up well when attempts are made to apply them to unrestricted domains; hence the importance of shallow inference procedures, which we describe in section 3.2.

We then present a treatment of relative pronouns in section 4, describe our corpus in section 5 and evaluate our algorithm in section 6.

2 Extracting GRs by Pattern Matching

Grammatical function is an important determinant of salience. As anaphora resolution algorithms have a strong subject preference, it is important that we are able to reliably differentiate between subjects and objects.

While most implementations use some form of

		B&C	Charniak	Collins 1	Collins 2	<i>Us</i>
subj	precision(%)	84	91	89	90	88
	recall(%)	88	85	80	83	88
	F-measure	0.86	0.88	0.85	0.84	0.88
dobj	precision(%)	86	82	83	83	89
	recall(%)	84	67	62	55	76
	F-measure	0.85	0.74	0.71	0.66	0.82
iobj	precision(%)	39	60	50	50	26
	recall(%)	84	32	32	32	89
	F-measure	0.53	0.41	0.39	0.39	0.40

Table 1: Evaluation of Grammatical Relation Extraction

parser or verb frame information to decide grammatical function, we do this using only pattern matching on noun-chunked text. We use the Edinburgh LT TTT toolkit (Grover et al., 2000) for POS tagging and noun chunking. We then use an ordered sequence of simple pattern matching rules to decide the grammatical function of noun groups. In the following patterns, the superscript of NP_i gives its grammatical function:

1. Prep NP_i^{obliq}
2. NP_i^{subj} [“, [^Verb]+,” | “Prep NP”]* Verb
3. Verb NP_i^{dobj}
4. Verb [NP]+ NP_i^{iobj}

The first pattern (gfun=*oblique*) looks back for a preposition. The second (gfun=*subject*) looks ahead for a verb, jumping over appositives and PPs. The third (gfun=*direct obj*) and fourth (gfun=*indirect obj*) patterns look back for a verb.

Preiss (2002) evaluated the performance of four parsers (Briscoe and Carroll (2002), Charniak (2000) and two versions of Collins (1997)) using the Carroll et al. (1999) evaluation corpus¹. We compare the performance of our approach with the results reported by Preiss (2002) in table 1.

Our approach identifies the object of any preposition as *oblique*, which results in very low recall for *iobj*. The results for our algorithm in

¹The evaluation corpus for GRs is available at <http://www.cogs.susx.ac.uk/lab/nlp/carroll/greval.html>

the *iobj* row in table 1 are actually for the conflated *iobj/oblique* class; i.e. NPs that match pattern 1 are also labelled as *iobj*. This results in high recall and low precision. The inability to differentiate *iobjs* from *oblique* references is not a problem for pronoun resolution as the Lappin and Leass (1994) paper uses the same weights for oblique and indirect object emphasis.

On the other hand, an important class of errors our algorithm makes is that of labelling temporal adjuncts as objects; for example, in *The judge said Friday that...* We overcome this in the agreement filter, where we say that hypernyms of the WordNet classes *time period* and *time unit* can only be antecedents if they appear in the subject position. This filter (like many of the filters we use) can be too restrictive; for example, consider:

- i John prefers Friday.
- ii It is a convenient day.

Our algorithm therefore has a mechanism to relax filters if no antecedent is found.

Our results indicate that subjects and direct objects can be determined reliably without resorting to parsing. This is significant, because our approach guarantees an analysis for every sentence, with a complexity that is linear in sentence length.

3 Resolving Third-Person Pronouns

3.1 Agreement Features

We use the four standard agreement features, for *number*, *person*, *gender* and *animacy*. We implement the features as lists of allowed values:

1. *number* = (s)ingular, (p)lural
2. *person* = (f)irst, (s)econd, (t)hird
3. *gender* = (m)ale, (f)emale, (n)euter
4. *animacy* = (a)nimate, (i)nanimate

This allows us to underspecify features when we have inadequate information. Having separate *animacy* and *gender* features allows us to handle companies and animals in an elegant way. For a company, we set:

- *gender* = {n}
- *animacy* = {a}

For an animal, we set:

- *gender* = {m/f,n}
- *animacy* = {a}

Then, for example, the pronoun *it* can refer to something with *gender*={n} and *animacy*={a} (like a company or animal) or something with *animacy*={i}. However, *he* can only refer to something with *gender*={m} and *animacy*={a} (an animal but not a company).

We also implement an additional *speaker-quote* agreement feature. This enforces two restrictions: firstly, third person pronouns within quotes cannot co-refer with the speaker of the quote and secondly, pronouns that are speakers of quotes cannot co-refer with noun phrases (apart from first person pronouns) within the quote. And, as described in section 2, we implement a filter for temporal adjuncts.

3.2 Inferring Agreement Values

Of the four standard agreement features — *number*, *person*, *gender* and *animacy*, values for the first two are available from the POS tagger; however, the tagger does not provide gender and animacy information. To get the most out of our agreement filters, we need to infer as much agreement information as possible. Ge et al. (1998) present an unsupervised approach to learning gender information from a corpus. We take an alternative approach to the problem. In edited text, animacy and gender information for a potential antecedent is usually available in some form elsewhere in the text, usually in other references to the

same referent. We try and retrieve this information using shallow inference mechanisms. We run through the set of noun phrases in iterations that:

1. Look for keywords in the NP
2. Try to co-refer the NP with another NP
3. Collect information about the head noun in WordNet
4. Infer from appositives and existential constructs
5. Make use of any reliable verb frames

In each iteration, we only consider noun phrases that have some agreement information (*animacy* or *gender*) missing.

In the first iteration, we look for keywords in an NP; for example, key words like *Inc.*, *Lmt.*, *PLC.* and *Corp.* suggest that the noun phrase is a company (*gender*={n} and *animacy*={a}) and titles like *Mrs.* and *Ms.* suggest that the noun phrase is a female person (*gender*={f} and *animacy*={a}). We use a list of 19 keywords.

In the second iteration, we try and co-refer an NP with an NP for which we have the required information. For example, in:

Pierre Vinken^x, 61 years old, will join the board as a nonexecutive director Nov. 29. **Mr. Vinken**^y is chairman of Elsevier N.V., the Dutch publishing group.

we can find agreement values for *x* by co-referring it to *y*, which the first iteration has dealt with.

The first two iterations largely deal with proper nouns, a particularly troublesome class. The third iteration deals with common nouns and involves a look-up of the head noun in WordNet. If the head noun is a hypernym of *human*, *animal* or *organisation* we set *animacy*={a}, otherwise we set *animacy*={i}. Gender information is sometimes available for humans in WordNet; for example if the head noun is *son*, *woman*, *widow* or *spinster*. WordNet also recognises some place names, particularly countries and cities.

The fourth iteration makes use of information contained in appositives and existential constructs; for example, in:

J.P. Bolduc^x, vice chairman^y of W.R. Grace Co., was elected a director.

and:

Finmeccanica^x is an Italian state-owned holding company^y with interests in the mechanical engineering industry.

we assign $animacy=\{a\}$ to x using the WordNet class of the head noun of y (*chairman* and *company*). We also set $gender=\{n\}$ for *Finmeccanica* and rule out $gender=\{n\}$ for *J.P. Bolduc*.

The fifth iteration makes use of the reliable verb frames. For example, the subject of verbs like *said*, *reported*, *stated* are assigned $animacy=\{a\}$.

3.3 Saliency and Syntax Filters

We use the following Lappin and Leass (1994) saliency features:

Saliency Factor	L&L Weight
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect Object / Oblique	40
Head Noun emphasis	80

We also consider *possessives*, giving them a weight equal to the weight of their enclosing NP minus ten. The additional features we consider are the membership of a co-reference class, the WordNet category of the co-reference class and noun phrase recency (distance of the potential antecedent measured in noun phrases). We discuss the weighting of the features further in section 6.1 on methodology.

Our syntax filter is an implementation of Kennedy and Boguraev (1996).

4 Resolving Relative Pronouns

Current parsers like the RASP (Briscoe and Carroll, 2002) take the view that determining what a relative pronoun refers to is not a problem that can always be solved in a syntactic framework; hence non-restrictive relative clauses are treated as text adjuncts, leaving the attachment decisions to anaphora resolution algorithms. However, relative pronouns have been largely ignored in the anaphora resolution literature.

In a previous paper (Siddharthan, 2002), we treated relative pronoun resolution as a clause attachment issue and approaches it in a machine learning framework, using WordNet classes and acquired prepositional preferences as features. In this section we treat it as an anaphora resolution problem, and provide a resolution mechanism based on saliency, agreement and syntactic filters.

4.1 Syntactic Filter

The antecedent of a relative pronoun is usually only separated from it by prepositional phrases or appositives; for example, in:

One man who is likely to reap the benefits is Vaino Heikkinen¹, aged 67, a farmer in Lieksa, 10km from the Soviet border, who¹ claims a Finnish record for shooting 36 bears since 1948.

and:

‘The pace of life was slower in those days,’ says 51-year-old Cathy Tinsall² from South London, who² had five children, three of them boys.

Our syntactic filter rules out any potential antecedent that is separated from the relative pronoun by any other category. This filter can be too restrictive. If no antecedent is found, we do away with the syntactic filter completely and try again.

4.2 Agreement Values

The relative pronoun *who* has $num=\{s,p\}$, $gender=\{m,f,n\}$ and $anim=\{a\}$. This allows *who* to refer to people, companies and animals, but nothing inanimate.

The relative pronoun *which* has $num=\{s,p\}$, $gender=\{n\}$ and $anim=\{a,i\}$. This allows *which* to refer to companies, animals and inanimate objects, but not people.

The relative pronoun *that* has $num=\{s,p\}$, $gender=\{m,f,n\}$ and $anim=\{a,i\}$. This allows *that* to refer to any noun phrase.

4.3 Saliency

We use the same saliency function as for third person pronouns; however, we weight it according to the relative pronoun and the animacy of the

Genre / Corpus	Training Set		Test Set	
	3 rd	Rel.	3 rd	Rel.
Guardian News	93	33	81	24
Guardian Sports	99	25	105	22
Guardian Opinion	93	24	88	20
NY Times News	117	35	122	41
NYT Sports	94	15	93	28
NYT Opinion	92	25	111	35
Literature	231	33	216	11
Comp. Manuals	89	42	-	-
Travelogues	-	-	93	27
Medical Articles	-	-	70	23
Total	908	230	979	231

Table 2: Number of 3rd Person and Relative Pronouns in our Corpus

co-reference class under consideration. For *who*, we increase the salience of potential antecedents that are people ($anim=\{a\}$ and $gend=\{m,f\}$). For *which*, we increase the salience of potential antecedents that are organisations or animals ($anim=\{a\}$ and $gen=\{n\}$).

5 The Corpus

Due to the lack of a standardised evaluation corpus for pronoun resolution, we have constructed an annotated corpus, the contents of which are described in table 2. The training and test corpora contain some genre in common (articles from the news, sports and guest column sections of one British and one American daily). The literature component of the training corpus consists of Beatrix Potter, H.H. Munro, Rudyard Kipling and Anna Sewell. The literature component of the test corpus consists of Aesop, Lewis Carroll and Agatha Christie. In addition, we have included some genre in the test set that we have not trained on, specifically travelogues (from the Lonely Planet guide) and medical articles.

We expect that this corpus will not overlap with corpora traditionally used in NLP that algorithms might have been trained on, and hence can be useful to other researchers as an independent evaluation corpus. Our annotation marks sentences and noun phrases and assigns each NP an index and an

optional co-reference index; for example, in:

(S1 (NP Mr Gilchrist 93) denied (NP-PRP he 94#93) was scare-mongering.)

the pronoun *he* has index 94 and co-refers with the noun phrase with index 93.

Pronouns in the corpus are co-referenced with the most recent antecedent. However, earlier antecedents can be recovered for evaluation purposes by following the co-reference chains backward. Pronouns with no antecedent in the discourse are given the co-reference index #-1. Plural pronouns that have more than one NP as antecedents are, for the moment, given the co-reference index #-2. In future, they could be dealt with using multiple #s.

6 Evaluation

6.1 Methodology

We now consider the question of what evaluation criterion should be used in the training stage. Our *gold standard* is marked up with chains of co-references and we have two options. Suppose our algorithm has resolved the pronouns as below:

Although **Hindley**¹'s own plans are still in place, police sources say they may have to be revised. "There will be no big send-off," said one **officer**². Feelings about **her**^{3#2} still run very high so all arrangements have to be carefully worked out. Just 12 people had been invited to attend the service including **her**^{4#3} mother.

We could treat the pronoun *her*^{4#3} as correctly resolved as it co-refers correctly with *her*³. As salience decreases very fast with distance, the salience of a class tends to be dictated by its most recent member. By verifying only the most recent antecedent, we are evaluating how well salience is working. In future, we refer to the evaluation on the most recent antecedent as *Eval-Salience*.

However, if (as above) the most recent antecedent is a pronoun (*her*^{3#2}), we should chain back all the way to decide if the pronoun has been resolved correctly. In this example our algorithm has resolved *her*^{4#3} incorrectly to *officer*². Ultimately, this is what we are interested in, and from now on, we refer to this "absolute" evaluation as *Eval-Absolute*. *Eval-Salience* is an indica-

Genre / Corpus	Training Corpus			Test Corpus		
	Base Algo	+ WordNet	+ Inference	Base Algo	+ WordNet	+ Inference
Guardian Opinion	.60 / .65	.79 / .81	.80 / .84	.61 / .73	.69 / .77	.83 / .85
Guardian News	.58 / .64	.77 / .79	.80 / .81	.61 / .69	.56 / .77	.60 / .78
Guardian Sports	.57 / .68	.53 / .74	.80 / .85	.60 / .71	.71 / .79	.84 / .87
NY Times Opinion	.60 / .76	.65 / .77	.81 / .88	.56 / .65	.72 / .79	.85 / .88
NY Times News	.53 / .64	.68 / .77	.82 / .86	.68 / .75	.75 / .79	.84 / .85
NY Times Sports	.70 / .76	.77 / .83	.69 / .75	.62 / .70	.71 / .82	.80 / .84
Literature	.61 / .75	.67 / .80	.73 / .84	.55 / .62	.68 / .71	.74 / .84
Computer Manuals	.66 / .72	.72 / .76	.74 / .78	-	-	-
Travelogues	-	-	-	.66 / .73	.77 / .79	.84 / .87
Medical Articles	-	-	-	.54 / .72	.65 / .83	.89 / .90
Average	.61 / .71	.69 / .79	.76 / .82	.60 / .70	.69 / .79	.79 / .85

Table 3: Results for Third Person Pronouns — Accuracy is reported as *Eval-Absolute* / *Eval-Salience*

tor of how well our algorithm can perform. *Eval-Absolute* measures how well it does. The difference is a measure of how far errors propagate.

We use *Eval-Salience* in the training phase as training on *Eval-Absolute* would result in preferentially fixing the errors that (purely by luck) happen to propagate a long way.

We use the training stage to determine the weights for the salience features, as well as to decide the number of WordNet senses to consider and the order in which to use our inference rules. As our aim is to build a genre-independent system, we need to make sure we do not over-train on our data. We do this by trying to ensure that the training improves results on all the training genre individually, not just the whole corpus collectively. We found that altering the original Lappin and Leass (1994) weights in different ways gave improved performance on some genre, but also resulted in worse performance on other genre. For genre-independent performance, the exact salience weights were not significant, as long as there was a strong subject preference.

6.2 Results

We present our results for third person pronouns in table 3. The results for the basic algorithm on our corpus are comparable to those reported by Preiss (2002) and Barbu and Mitkov (2001) for completely different corpora. There is a big improvement when we use WordNet to obtain agree-

ment values (section 3.1). There is a further improvement when we infer agreement values for agreement features (3.2) and enforce *speaker-quote* agreement (section 3.1). The fact that we report better results on the test corpus suggests that we have not over-trained our system. It is interesting to note that the *Eval-Salience* measure appears to stay reasonably constant across data sets. However, the *Eval-Absolute* measure can vary wildly, from *Eval-Salience* in the best case when errors do not propagate at all, to 20% below *Eval-Salience* when they propagate far. This suggests that traditional evaluations of pronoun resolution algorithms on small corpora can involve a fair bit of luck. However, *Eval-Absolute* becomes more reliable as the evaluation corpus gets larger.

We present our results for relative pronouns in table 4. We report a 10% improvement over the local attachment baseline. These results are comparable to those previously reported by us (Siddharthan, 2002), where we used machine learning techniques on WordNet Classes and prepositional preferences and evaluated on the Penn WSJ Treebank (Marcus et al., 1993).

7 Conclusions and Future Work

We have described various shallow techniques for salience-based anaphora resolution. We have described how important grammatical function information can be obtained reliably by pattern matching on chunked text and shown that acquir-

	Training Corpus	Test Corpus
who	.98 (.82)	.98 (.87)
which	.97 (.85)	.86 (.78)
that	.85 (.81)	.96 (.93)
Average	.94 (.82)	.94 (.86)

Table 4: Precision results for Relative Pronouns — The local attachment baseline is shown in brackets

ing agreement information for a NP from the surrounding text and WordNet can significantly boost results *across genre*. In contrast, adjusting the weights of salience features results in only genre-specific improvements. We have also shown that the same framework can be used to resolve relative pronouns with high accuracy. This is a result that might be significant to other fields like parsing.

Despite our attempts at inferring agreement information, many of the mistakes our algorithm makes remain due to insufficient knowledge of animacy and gender. This suggests that it might be worthwhile considering other techniques for obtaining information about them; for example, gender information could perhaps be obtained from census data and Orăsan and Evans (2001) provide machine learning techniques for inferring animacy. Most of the remaining errors were due to sentences containing parenthetical information that led to an incorrect focus shift; like the example in section 6.1. Accurately identifying such sentences would significantly aid our anaphora resolution algorithm.

Acknowledgements

We are grateful to Ted Briscoe for valuable discussions based on a first draft of this paper, and to three anonymous referees for providing useful feedback that has helped improve this paper further.

References

- Catalina Barbu and Ruslan Mitkov. 2001. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Toulouse, France, pages 34–41.
- Ted Briscoe and John Carroll. 2002. Robust accurate

statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria*, pages 1499–1504.

- John Carroll, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL99 workshop on Linguistically Interpreted Corpora (LINC)*, Bergen, Norway, June 12.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of ANLP-NAACL 2000, Seattle, Washington*, pages 132–139.

Michael Collins. 1997. Three generative, lexicalized models for statistical parsing. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics*, pages 16–23.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–171.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT - A flexible tokenisation tool. In *Proceedings of Second International Conference on Language Resources and Evaluation*.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora in a wider context: Tracking discourse referents. In *European Conference on Artificial Intelligence*, pages 582–586. John Wiley and Sons, Ltd, London/New York.

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

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

Constantin Orăsan and Richard Evans. 2001. Learning to identify animate references. In Walter Daelemans and Rémi Zajac, editors, *Proceedings of CoNLL-2001*, pages 129–136, Toulouse, France, July, 6 – 7.

Judita Preiss. 2002. Choosing a parser for anaphora resolution. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, Lisbon, Portugal, pages 175–180.

Advaith Siddharthan. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, USA, pages 60–65.